# Statistical Models for the Analysis of Short User-Generated Documents

## Author Identification for Conversational Documents

Doctoral Dissertation submitted to the
Faculty of Informatics of the *Università della Svizzera Italiana*
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

presented by
**Giacomo Inches**

under the supervision of
Prof. Fabio Crestani

April 2015

## Dissertation Committee

| | |
|---|---|
| **Prof. Michael Bronstein** | Università della Svizzera Italiana, Switzerland |
| **Prof. Mehdi Jazayeri** | Università della Svizzera Italiana, Switzerland |
| **Prof. Fazli Can** | Bilkent University, Ankara, Turkey |
| **Prof. Douglas W. Oard** | University of Maryland, College Park, U.S.A. |

Dissertation accepted on 30 April 2015

**Prof. Fabio Crestani**
Research Advisor
Università della Svizzera Italiana, Switzerland

**Prof. Igor Pivkin**
PhD Program Director

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Giacomo Inches
Lugano, 30 April 2015

*To my ♡ family*

*Il tempo è un concetto utile, ma non fondamentale.*

Carlo Rovelli, TEDxLakeComo 2012

# Abstract

In recent years short user-generated documents have been gaining popularity on the Internet and attention in the research communities. This kind of documents are generated by users of the various online services: platforms for instant messaging communication, for real-time status posting, for discussing and for writing reviews. Each of these services allows users to generate written texts with particular properties and which might require specific algorithms for being analysed.

In this dissertation we are presenting our work which aims at analysing this kind of documents. We conducted qualitative and quantitative studies to identify the properties that might allow for characterising them. We compared the properties of these documents with the properties of standard documents employed in the literature, such as newspaper articles, and defined a set of characteristics that are distinctive of the documents generated online. We also observed two classes within the online user-generated documents: the conversational documents and those involving group discussions.

We later focused on the class of conversational documents, that are short and spontaneous. We created a novel collection of real conversational documents retrieved online (e.g. Internet Relay Chat) and distributed it as part of an international competition (PAN @ CLEF'12). The competition was about author characterisation, which is one of the possible studies of authorship attribution documented in the literature. Another field of study is authorship identification, that became our main topic of research. We approached the authorship identification problem in its closed-class variant. For each problem we employed documents from the collection we released and from a collection of Twitter messages, as representative of conversational or short user-generated documents. We proved the unsuitability of standard authorship identification techniques for conversational documents and proposed novel methods capable of reaching better accuracy rates. As opposed to standard methods that worked well only for few authors, the proposed technique allowed for reaching significant results even for hundreds of users.

# Acknowledgements

My first thank you goes to Fabio Crestani, who gave me the opportunity of joining the Ph.D. program at USI. I am grateful to him for setting up such an outstanding research group in Information Retrieval. During my 5+ years at USI I had the pleasure to collaborate with loyal friends and great colleagues (in pseudo-random order): Marcello Paolo Scipioni, Shima Gerani and Mehdi Mirzaaghaei, Ilya Markov and Maia Khassina, Mostafa Keikha, Mark Carman, Parvaz Mahdabi and Nicolas Schiper, Monica Landoni, Morgan Harvey and Ana Javornik, Marco Pasch, Luca Colombo, Andrei Rikitianskii, Alessandro Margara (*parallelisation is an art*), Konstantin Rubinov, Az Azrinudin Alidin, Davide Eynard, Giordano Tamburrelli, Robert Gwadera, Eduardo Bezerra, Slobodan Lukovic, Igor Kaitović, Liudmila Kazak, Roberto Stefanini, Leandro Fiorin, Mattia Vivanti, Amir Malekpour, Philippe Moret, Antonio Taddeo, Tim Winkler, Marco D'Ambros, Cyrus Hall, Jochen Wuttke, Adina Mosincat (*rest in peace*), Sasa Nesic, Nemanja Memarovic, Nicolò Perino, Aibek Sarimbekov, Maksym Zavershynskyi, Alberto Bacchelli, Alessio Gambi, Alessandro Antonucci, Marcin Nowak, Panos Hilaris, Sebastian Daum, Alessandro Inversini, Saman Kamran, Domenico Bianculli, Lile Hattori, Alexander Tomic, Spicuglia Sebastiano, Simone Fulvio Rollini, Alessandro Rigazzi, Ioannis Athanasiadis, Marco Primi, Tim Winkler, Giovanni Ansaloni, Paolo Bonzini, Alessandra Gorla, Giacomo Toffetti Carughi, Elisa Rubegni, Cheng Zhang, Anna Förster, Francesco Alberti, Dmitry Anisimov, Samuel Benz, Sandeep Kumar Dey, Parisa Jalili Marandi, Ricardo Padilha, Matteo Agosti, Stefano Vaghi, Alberto Ferrante, Ivan Elhart, Giovanni Toffetti Carughi, Andrea Mattavelli, Andrea Gallidabino, Marcello Romanelli, Koorosh Khazaei, Elena Khramtcova, Andrea Rosà, Randolf Schaerfig, Daniele Sciascia, Evangelos Niforatos and many others. Each of you supported me in different ways, in particular during our meetings, lunch discussions, conferences or summer school trips.

Thanks to the TILO's friends: Nicola, Federico, Riccardo, Daniele, Antonio, Veronica, Silvia, Cecilia, Livia, Debora, Roberta, Chiara, Stefania.

I am very indebted to the "Decanato": Elisa "the Queen" Larghi, Janine Caggiano, Cristina Spinedi, Diana Corica, Danijela Milicevic and Marisa Clemenz. Thanks to Silvia Invrea, Annelore Denti, Anna Giovanetti, Nicole Bandion, Gilda Schertenleib, Ladina Caprez and Rosario Maccarrone for the enriching and friendly collaboration. I am grateful to Mauro Prevostini for giving me the possibility to interact with high school students and their teachers following the different programs at the Faculty.

Thanks to Gabriella Pasi and Andrea Emilio Rizzoli I was helping with teaching and to other faculty members who supported me at different levels: Mark Langheinrich, Michele Lanza, Antonio Carzaniga, Kai Hormann, Mauro Pezzè, Matthias Hauswirth, Fernando Pedone, Laura Pozzi, Nathaniel Nystrom, Luca Maria Gambardella, Igor Pivkin.

I am obliged to the many invited speakers and visiting students I had the honour to meet at USI and to the many researchers I met at different conferences who dedicated some of their time to provide me with feedback, in particular: Javier Parapar, Elena Sivogolovko, Daria Bogdanova, Emanuele Panzeri, Giorgio Ghisalberti, Omar Alonso, Gloria Bordogna, Alfredo Cuzzocrea, Hussein Suleman, Gianni Amati, Evangelos Kanoulas, Peter Ingwersen, Thomas Gottron, Jacques Savoy, Mounia Lalmas, Djoerd Heimstra, Massimiliano Ciaramita, Nicola Ferro, Paul Thomas, Richard Bache, Peter Mika, Stefano Mizzaro, Kalervo Järvelin, Joachim Pfister, Leif Azzopardi, Fabrizio Silvestri, Mark Sanderson, Ian Ruthven, David Elsweiler.

A big and special "thank you!" to Andrea Basso and Paolo Rosso. Andrea allowed me to spend the 6 most amazing weeks of my life in the U.S.A. working in the AT&T Research Lab with extraordinary colleagues: Behzad Shahraray, Bernard S. Renger, David C. Gibbon, Eric Zavesky, Zhu Liu, Patrick Haffner, Junlan Feng and Ovidiu Dan. A big thank you to Jane Bazzoni, Daniela and Luca for hosting me. Paolo introduced me to the exceptional PAN team. Thanks to Martin Potthast, Benno Stein, Tim Gollub, Matthias Hagen, Alberto Barrón-Cedeño, Parth Gupta, Patrick Juola and Efstathios Stamatatos for their feedback and support in organising a subtask of PAN 2012.

I should not forget to mention Alessandra Franzini, who proofreaded the text of the dissertation, and my colleagues at Newscron, who supported me in the last sprint: Peter Hogenkamp, Elia Palme, Patrick Lardi, Mirko Pinna, Roger Lüchinger, Ahmed Fouad, Martin Weigert, Domagoj Kulundžić, Filip Matijević and Stephanie Grubenmann.

I am particularly grateful to the dissertation committee members Doug Oard, Fazli Can, Mehdi Jazayeri and Michael Bronstein for their feedback and the time they spent examining this thesis. Their valuable comments, suggestions and critics made this work more complete and outstanding.

This list will never be complete without the family who raised me when I was a child and the family that will be with me for the rest of my life. Thank you dad Sabatino and mum Serenella: you made me the man I am today and gave me the instruments and the knowledge I need to sail through the ocean of life. I will *always* be grateful to you. Thank you to my brothers Matteo and Filippo, together with their partners Nastasja and Sara. A brother is forever. Thanks grandma Maria, uncles and aunts for your silent and unconditional support. Thanks to my in-laws Aurelio and Valeria: I am very indebted to you. Thanks to Paolo & Francesca and the Rossini clan.

This work was only possible thanks to the unconditional support, patience and love of my wife Elena. You only know what this thesis means to me, the time, energy and emotions that are all in here. This work is all to you Elena and to our sweet and lovely Camilla and Carlotta. Thank you. G.

# Contents

# Figures

# Tables

# Chapter 1

# Introduction

*I'm a genetic optimist. I've been told, "Jeff, you're fooling yourself; the problem is unsolvable." But I don't think so. It just takes a lot of time, patience and experimentation.*

Jeffrey Bezos

## 1.1 Description of the problem

One of the most important needs of the human being is communication. Since ancient times he has expressed this necessity with verbal and written symbols, that later turned into letters, alphabets and then structured texts. Written texts, in fact, represented for hundreds of years the main way of communication for people, especially those separated by long distances. Letters were the main instrument for actively communicating and establishing a "dialog" between individuals that could not meet in person. Other forms of written communication included newspaper articles, to report recent events, and poems, books, or magazines, to spread knowledge of in-depth events or thoughts. For this reason, scientists started to analyze this kind of texts when they first wanted to investigate written documents [84, 27]. They were, in fact, the only available texts.

With the advent of computers first and then of the Internet, researchers were able to investigate, year after year, larger and larger collections of written documents, until faced with the recent challenges posed by all the novel kinds of written data (e.g. websites, emails, blogs, chat transcripts, etc). Over the past few years, the creation of such novel texts was made easier by the spread of small and smart devices with internet access, such as smartphones and tablets. As a matter of fact, technological innovations have always influenced text production: Gutenberg and his printing machine completely changed the way people approached written texts and the same can be said for personal computers and the Internet. However, the possibility of accessing

the Internet and its related services "anytime" and "anywhere" with a portable device represents another step forward in the way people are communicating and producing texts. Most (if not all) of the available services for communicating online make use of a form of textual interaction, thus generating a lot of interesting data for researchers. Among the various novel online services are:

- Blogs, in which people write short articles on topics of personal interest and engaged readers can leave their comments;
- Chat messages, in particular those originating from the Internet Relay Chat (IRC) providers but also within other softwares, like Skype or Facebook;
- Twitter messages, launched as a SMS based service and now a platform to share links, emotions, feelings or statuses more in general;
- Online fora, where communities of people discuss particular topics in an extensive way;
- Review websites, where users can share their experience of products or services;
- Email, used to communicate as in traditional letters but with the possibility of reaching multiple interlocutors at the same time.

This Ph.D. dissertation is placed within this context, in the general domain of the textual documents produced on the Internet. Moreover, it has a special focus on conversational documents generated online. We will define conversational documents later in Chapter 2. As to the first scenario, the general domain of textual documents produced on the Internet, there were not many previous studies investigating the features of the different types of documents emerging online. On the other hand, the research already conducted into conversational documents was limited and many problems remained without answers, in particular the ones related to authorship attribution and identification. We will analyze these two scenarios more in detail in the following paragraphs.

**Textual documents on the Internet** While the literature is abundant with work analysing in detail some specific collections of documents (e.g. blogs or reviews), there is a gap in the analysis of documents originated by services developed in the latest years (e.g. twitter, chats, fora, newsgroups). What is missing, in fact, is a comparative analysis among all these documents themselves and also between these documents and the standard documents traditionally employed in the literature (i.e. newspaper articles). These latter, in fact, were extensively employed in the literature due to their longer availability, while the former were just emerging at the time of starting our Ph.D. studies.

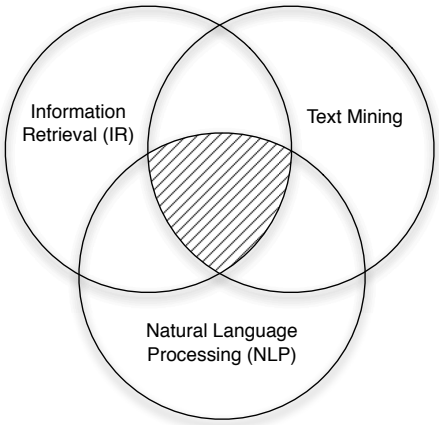



*Figure 1.1.* Positioning of the thesis work with respect to the different research areas.

**Conversational documents** Among the emerging set of documents, conversational documents are the ones for which less research had already been conducted, at least within our domain of reference. For this reason we decided to explore this kind of documents in our work. Moreover, the problem of authorship identification was never studied for conversational documents and we also filled this gap with our work.

In the next section (Section 1.2) we will further specify the research areas touched by this work, while in Section 1.3 we will highlight the research questions that drove our investigations. In Section 1.4 we will present the contributions that this work provided to advance the state of the art. The contributions are reflected also in the structure of the thesis presented in Section 1.5 and in the related publications, listed in Section 1.6.

## 1.2 Areas of research

This work is placed mainly in the area of Information Retrieval (IR), that is the reference area of the IR research group at the Università della Svizzera italiana (USI) where we operated during our Ph.D. studies. Moreover, many of the techniques employed in our work (e.g. Zipf's and Heap's Law, cosine similarity, Kullback-Leibler Divergence) have their main application in IR. However, our work covers a broader area than the interests of the IR community: we are not focused in particular on search, personalisation or interfaces, but on the more general context of user-generated documents on the Internet and Author Attribution. Other communities rather than the IR one, in fact, have already produced some work in these contexts and are more receptive on these topics. We are referring to the Text Mining community and the Natural Language Processing (NLP) one.

The goal of Text Mining is the extraction of novel knowledge from written texts, while NLP aims at understanding the language used to produce the written texts. To give an example, IR deals with "return me the documents that are most related to a given text", Text Mining aims at "finding the relations between a given text and a set of documents" while NLP wants to "find the structure of the given text and the set of documents". This is clearly a simplification but it serves to give a qualitative idea of the different goals of each research area. As should be clear from the example, despite the different purposes of each research community, the boundaries among the three are not really sharp. To the contrary, it is not unusual for researchers in one community to make use of techniques of the others to improve their systems or algorithms. As depicted in Figure 1.1 this is also the case of our work. Although our main research area is IR, we made use of concepts and techniques from Text Mining and NLP, in particular in our studies of the general properties of documents on the Internet (Chapter 4) and authorship identification (Chapter 5). Moreover, in Chapter 2 we will present the key concepts of IR and their relationship with our work. In the same chapter we will also give a quick introduction to Text Mining and explain its relationship with IR. Despite being influential at different level, we do not provide a dedicated introduction to NLP because its end effect on our work is limited and mostly applicative.

## 1.3 Research questions

Our work is guided by several Research Questions (RQ), each of which is addressed in a particular chapter of our work.

**RQ1** What are the differences between traditional collections of documents employed in Information Retrieval (e.g. newspaper articles) and recent collections of documents generated online (e.g. Internet Relay Chat -IRC- logs, Twitter)?

- Is there any difference within these collections of online user-generated documents: i.e. are Twitter documents different from IRC logs or from blogs?
- What are the specific characteristics of conversational documents (e.g. IRC logs)?

**RQ2** Is there a good representative collection for conversational documents, with a large number of documents and a variety of topics?

**RQ3** Are traditional methods of authorship identification suitable also for conversational documents?

- Are traditional methods of authorship identification suitable also for hundreds or thousands of authors, as in the case with conversational documents?

- How does the number of interlocutors of an author affect the performance of the author identification?

## 1.4 Thesis contributions

The contributions of this work are many and at different levels, as can be seen by the previous research questions. Like in a funnel (Figure 1.2), we first started our analysis by looking at the general features of some relevant collections of short-user documents generated online, for answering the first research question and its sub-questions. After having characterised the documents generated online and presented their differences with respect to the traditional documents employed in IR, we decided to restrict the documents under investigation to a single type, that is conversational documents. This decision was motivated by the challenging structure of these documents and the limited research already conducted on them. The first contribution in this field was the creation of a collection that could serve as reference for all interested researchers on this topic. In fact, a broad and diverse collection of conversational documents was missing in the literature and our contribution filled this gap. This answers the second research question. However, for readability and presentation purposes, we first introduce this contribution in the dissertation (in Chapter 3) and later the research work that answered the first research question (in Chapter 4). Finally, we decided to focus on a specific problem for the conversational documents: the authorship identification problem. In fact, while authorship identification has already been investigated for certain kinds of documents, from blogs to letters and diaries, this is the first time it has been addressed for this kind of documents. We think that this problem might also have a lot of implications not only in the academia but also in the industry, e.g. in the field of surveillance or security. This part of our work addressed the third research question and is presented in Chapter 5.

Although the main contributions are spread within the main chapters of the Ph.D. dissertation, in the two appendixes attached hereto we present two practical applications of the studies conducted on conversational documents. In particular, we present an analysis of Twitter for Social Tv (Appendix A) and the results of the competition of sexual predator identification in the conversational documents we organised (Appendix B). The study on Social TV employing Twitter documents is important because it inspired our later work on authorship identification. Indeed, the work on Social TV focused on identifying classes of users, allowing us to find methods (e.g. entropy measurement) that represent the foundation of more complex algorithms (e.g. distances between user profiles) later employed in the problem of user identification. However, due to the specificity of the problem dealt with ih this study, we decided not to incorporate it into the main story presented in the dissertation but to leave it as a separate specific report in the first appendix. Moreover, in the second appendix we report the specific details of the sexual predator identification competition we organised. These

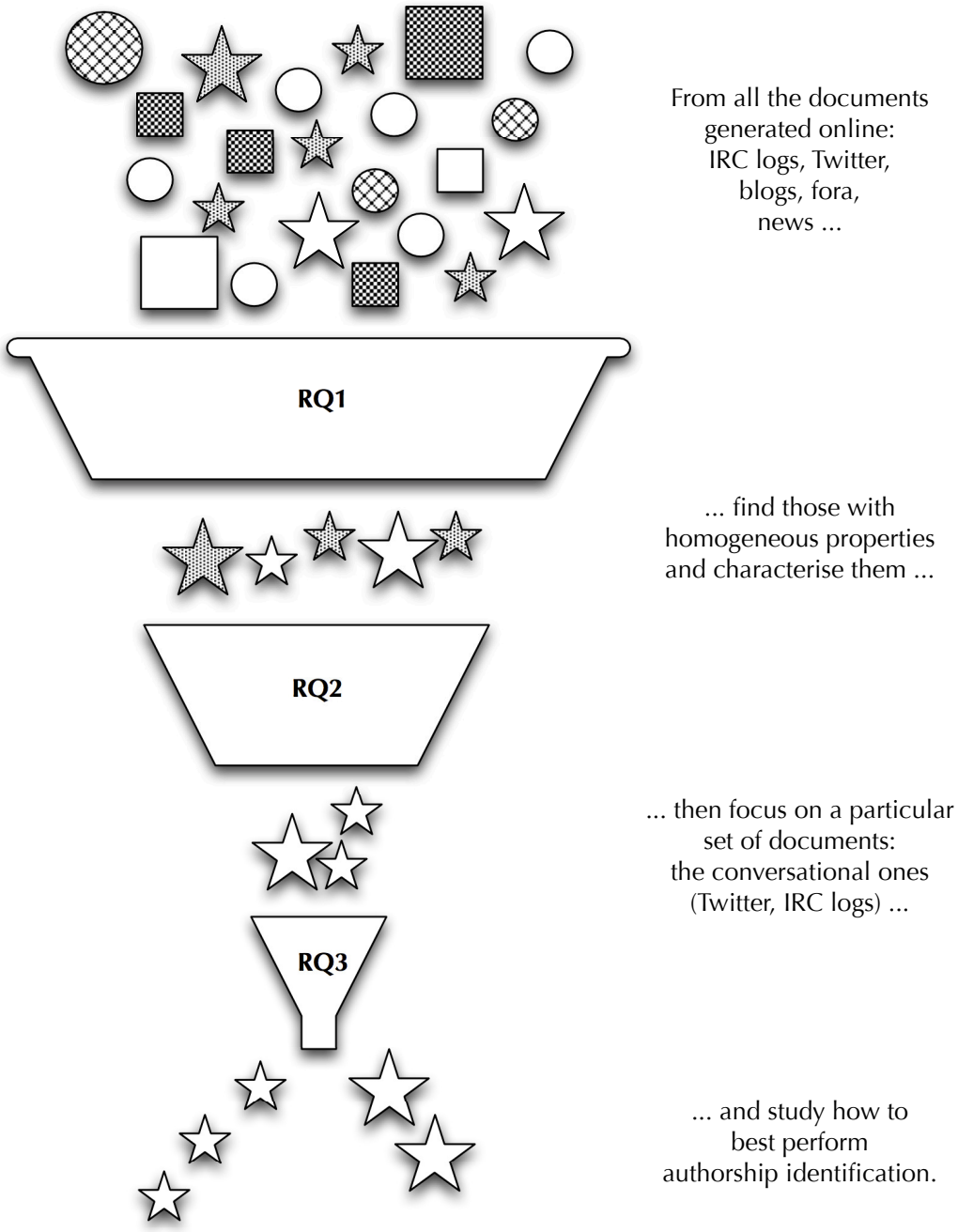



*Figure 1.2.* Contributions of the thesis.

details are less relevant than the collection realized to obtain them. For this reason, we decided to present the collection as part of our main contribution and leave the details of the organised competition in the appendix. Despite being in the appendix, these studies were instrumental in getting a broader knowledge of the literature of authorship attribution, identification and profiling. They were also instrumental in developing some of the approaches for authorship identification described in Chapter 5.

The exploratory nature of our work clearly emerges from the presentation of our contributions. This represented a challenge for our work, in particular for placing it within a specific research area or a particular research topic. In fact, unlike other dissertations where the focus is only on improving the state of the art of a particular area (e.g. improve algorithms for retrieving suggestions based on our current location), we first needed to set up our goals at different levels of detail and within different research areas (not a single one as already stated). Especially in the first part of our work, where we conducted a comparative analysis, we operated on a higher level that touched all the different areas presented in Figure 1.1 and we had to find the right community to present our work to. Despite being less general than the first one, also in the second part of the work we faced the problem of not having a specific reference community. In fact for conversational documents there was no previous work to compare with and in particular no standard collections and associated ground truth. For this reason we had to set up our own. Finally, in the last part of the work, which is even more specific than the second one, we could rely on the collections we created previously and, to some extent, also on some previous work to compare our results with. In fact, we demonstrated how traditional algorithms for authorship identification were failing in the case of conversational documents and presented our own solution.

To conclude, our work was part of (and partially funded by) the Chat Miner[1] project of the Swiss National Science Foundation. This project also helped us to focus on the conversational documents and author identification task in our work.

### Summary of contributions

1. We characterised the online user-generated documents and
    - identified features that distinguished them against traditional documents employed in the literature, in particular in IR;
    - identified features that separate them into "conversational" documents (Twitter, IRC logs) and "discussion" documents.
2. We created an evaluation framework for conversational documents (online-chats) that was used in several international competitions (CLEF-PAN 2012, CLEF-PAN 2013) and in our own experiments.

[1]Mining Conversational Content for Topic Modelling and Author Identification, `http://p3.snf.ch/project-130208`

3. We developed novel algorithms for authorship identification in conversational documents that improved the accuracy of standard algorithms or allowed for good accuracy but with a smaller number of features.

## 1.5 Thesis Outline

**Chapter 1** In this chapter we introduced the topic of our dissertation. We highlighted the problems to be solved, the research areas touched during our work as well as the research questions that guided us during the Ph.D. studies. We described the contribution of our work and its related publications.

**Chapter 2** This chapter gives an overview of the different research areas touched in our work while presenting the background of the main ones. We will focus primarily on IR, which is our reference research area, and then move on to Text Mining. Later we will introduce Conversational Text as related to Text Mining and present its features. We will finally give a summary of the authorship attribution and authorship identification problems.

**Chapter 3** In this chapter we will present a set of relevant collections already present in the literature. This set represents a sample of online user-generated documents plus two representative sets of documents traditionally employed in IR. We will illustrate the positive and negative aspects of each collection of documents and motivate the need for creating a new collection of conversational documents. The properties and use-case scenario of this new collection are also presented. This represents our first contribution answering **RQ2**.

**Chapter 4** This chapter contains a comparative study of the different collections of online user-generated documents and traditional documents. We will employ different criteria (qualitative and quantitative) to analyse the different collections. We will identify differences between the collections as well as within the online user-generated documents themselves. This is our second contribution that answers **RQ1**.

**Chapter 5** In this chapter we will focus on a particular problem: authorship identification. We approached this problem by verifying the limited impact of standard techniques of authorship identification for collections of conversational documents. For this reason we proposed a novel technique that performed better for conversational documents. This is our third contribution that answers **RQ3**.

**Chapter 6** In this chapter we will draw the final conclusions and summarise the importance of our work. We will also sketch some future work that might supplement that already conducted within this dissertation.

**Appendix A** In the first appendix we will present a work on Social TV that represents an application of the studies conducted in Chapter 4 and inspired the studies conducted in Chapter 5. Furthermore, this work also helped us to understand the importance of the short user-generated documents not only for academia but also for applied research to focus on the problem of user characterisation.

**Appendix B** In this appendix we will present an application of the collection we created and introduced in Chapter 3. This collection of conversational documents served as a testbed for participants in an international contest. The appendix reports all the technical details of the competition: purpose of the contest, participants approaches, evaluation framework and final remarks.

**Appendix C** In this appendix we will provide the tabular representation for the values of accuracy computed during the experimental evaluation. In fact, Chapter 5 only presents the values in graphs.

**Appendix D** In this appendix we will display some samples from the log files of our code for the experimental evaluation of Chapter 5. This serves to understand the running times of the algorithms on the different collections under the chosen experimental settings.

## 1.6 Publications

These are the publications that resulted from the work reported in this dissertation:

P1 G. Inches, M. J. Carman, and F. Crestani, **Statistics of Online User-Generated Short Documents**, in ECIR 2010: Proceedings of the 32nd European Conference on IR Research on Advances in Information Retrieval, 2010, vol. 5993, pp. 649-652, Milton Keynes, UK.

P2 G. Inches., M. J. Carman, and F. Crestani, **Investigating the Statistical Properties of User-Generated Documents**, in FAQS 2011: Proceedings of the 9th International Conference on Flexible Query Answering Systems, 2011, vol. 7022, pp. 198-209, Ghent, Belgium.

P3 G. Inches and F. Crestani, **Online Conversation Mining for Author Characterization and Topic Identification**, in Proceedings of the 4th workshop on Workshop for Ph.D. students in information & knowledge management - PIKM '11, 2011, pp. 19-26, Glasgow, UK.

P4 G. Inches, A. Basso, and F. Crestani, **On The Generation of Rich Content Metadata from Social Media**, in Proceedings of the 3rd international workshop on Search and mining user-generated contents - SMUC '11, 2011, pp. 85-92, Glasgow, UK.

P5 G. Inches, F. Crestani, **Overview of the International Sexual Predator Identification Competition** at PAN-2012. CLEF (Online Working Notes/Labs/-Workshop).

P6 G. Inches, M.Harvey, F. Crestani, **Finding Participants in a Chat: Authorship Attribution for Conversational Documents** in ASE/IEEE International Conference on Social Computing, 2013, pp. 272-279, Washington D.C., USA.

P7 G. Inches, F. Crestani, **An Introduction to the Novel Challenges in Information Retrieval for Social Media** In Bridging Between Information Retrieval and Databases, volume 8173 of Lecture Notes in Computer Science, pp. 1–30. Springer Berlin Heidelberg, 2014.

# Chapter 2

# Background

*The older I get, the more I see how much motivations matter. [...] If you don't love something, you're not going to go the extra mile, work the extra weekend, challenge the status quo as much.*

Steve Jobs

In this chapter we will introduce the research areas connected to our work and present their related state of the art. We will start in Sections 2.1 and 2.2 with an overview of the main research areas dealt with in our work, namely *Information Retrieval* (IR) and *Text Mining*. We will then continue presenting the specific area of *conversational texts* (Section 2.3), a cross-boundary between IR and Text Mining, we are predominantly working in. In the following Section (2.4) we will discuss, instead, the literature review of an important aspect of conversational texts investigated, namely, *authorship identification* (Section 5.1). We will emphasise the importance of authorship identification in relation to other problems of authorship characterisation and illustrate the *Sexual Predator Identification* contest that we organised as part of our work on authorship attribution (Section 2.4.1). A discussion on the the available collections in literature and the motivations for a new and better corpus is postponed to Chapter 3, that also presents the new corpus realised and released as part of this doctoral work.

## 2.1 Information Retrieval

Information Retrieval (IR) is the research area "concerned with the structure, analysis, organization, storage, searching, and retrieval of information" (according to a classical definition [111], recently recalled in [27]). IR has been involved with the study of written documents since its early stages.

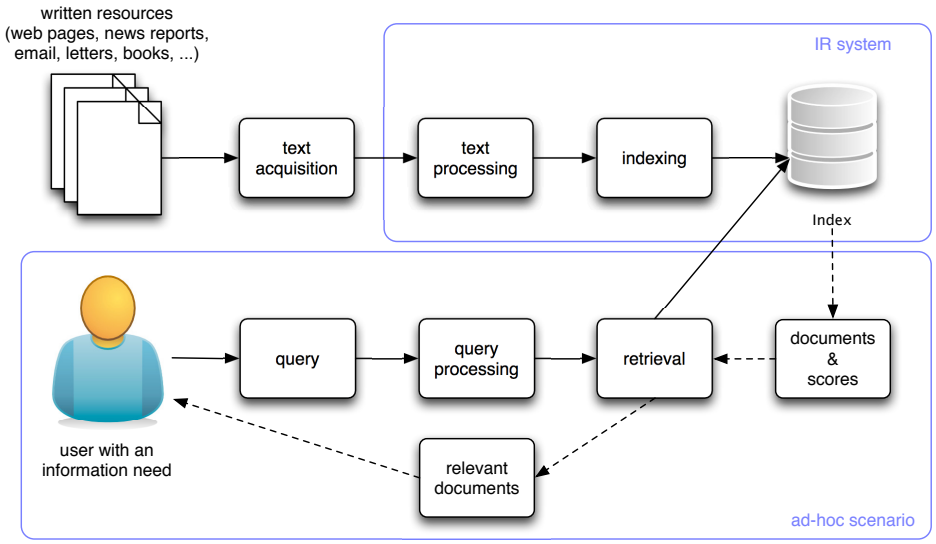



*Figure 2.1.* From written resources to indexing and a classical (ad-hoc) problem of Information Retrieval, from a user need to the returned list of relevant documents.

A milestone in Information Retrieval is the advent of the Internet in the late '90s. This coincided with a declining interest in traditional written documents (e.g. books, letters, newspaper articles) in favour of novel kinds of documents, mainly generated online. This reflected the changes in the way people started to communicate, for example substituting traditional letters on paper with electronic letters (email) or moving debates from written magazines and journals to online fora or blogs. Other services were created online, allowing for faster communication: platform for product reviews (like Tripadvisor), Internet Relay Chats (IRC) and microblog platforms (like Twitter). This novel set of documents called for an in-depth analysis to understand their properties with respect to traditional written documents and the applicability of traditional IR techniques. In Chapter 3 we will present a detailed overview of some relevant collections of online user-generated documents: IRC logs, Twitter, online blogs and discussion fora, SMS, email messages.

### 2.1.1 A Classical Information Retrieval System

In classical IR systems (Figure 2.1), textual documents are generally acquired using a system that automatically downloads them from the Internet (crawler) or reads them from a particular repository. These documents are then transformed into a simplified form thanks to a set of text processing techniques, some of which are presented later (Chapter 4). Documents in this simplified form are then stored in an index, where the information about terms and documents is made explicit. For example, given a term, one can find all the documents that contain that term (inverted index), or given a

document, the list of terms is made available (direct index). The index is consequently the main source of data for any application based on an IR system. Examples include the classical ad-hoc search (e.g. Google, Yahoo or Bing search engines), in which a user tries to retrieve documents containing some information and expresses this need by entering a representative set of terms (called query) into the system. The system, as a result, answers with a list of relevant documents (Figure 2.1).

In this dissertation we mainly focus on the text processing component (Figure 2.1) of the IR system, which we will present in the next section. In Chapter 4, in particular, we will focus on the major activities that can be found in the text processing component (i.e. parsing, stemming, stopwords removal) to study the features of traditional and online user-generated collections. Along with these activities we will perform other analyses on the same collections using other techniques, in particular the Part-Of-Speech (POS) processing. In Chapter 5, instead, we will focus on the problem of authorship attribution, making use of similarity measures. In particular we will employ measures of divergence, that are widely used both in the text processing component of the IR system and in the retrieval one.

### 2.1.2 Text Processing

Parsing is the first activity performed by an IR system. A program called "parser" reads the texts collected and divides them into terms based on characters' boundaries, such as whitespace or punctuation, preserving, or not, particular characters sequences (words but also links, markers, etc ), numbers, punctuation or symbols. After that, a list of characters representing words or terms is associated to each parsed text. All the terms that occur at least once in the text make up the vocabulary. Terms that appear just once in the text are called singleton, while the most frequent terms are inserted into the so-called stopwords list. Generally the most common terms in a text are functional words, e.g. propositions, connectors like "the" or "and" that do not contribute much to characterize the topicality of the text and therefore can be removed. This procedure is called "stopwords removal" and has the benefit of reducing the size of the index. Sometimes, this also provides a better retrieval of the documents, in particular for the ad-hoc scenario (e.g. searching a document online by keywords). Despite that, for other kinds of applications, such as authorship attribution, the functional words play an important role in charaterizing the style of the author of one text (Chapter 5) or influence the syntactical properties of the elements in the text and therefore the stopwords removal is neither desired nor applied (e.g. POS analysis in Chapter 4). The last important operation present in the text processing block is "stemming". Stemming is an operation carried out at term level that reduces all the words to their root: for example "win, winner, winning" are all reduced through stemming to a shorter form, in this case win. This also reduces the dimension of the index and in classical ad-hoc retrieval it may improve the performance of the system when the terms both in the query and in the vocabulary are reduced to the same basic form. This is not always the case for online

user-generated documents, as it will be explained in Chapter 4.

### 2.1.3 Similarity Measures

To measure the degree of similarity between a document and a query, or between two texts, different metrics can be applied, depending on the particular model assumed. One of the most common models used in literature is the vector space model. The vector space model assumes that each document is a vector and each term in the document is a component of this vector. With this assumption, the ordering of the words in each document is no longer important and each word represents a dimension in the vector space [81]. Models like this, that assume independence of the terms from their position in a document, are often referred as "bag-of-words" approaches. The measure of the distance in the case of the vector space model is the *cosine similarity*, that measures the angle between two vectors (documents) in the space, thus their degree of similarity. If the two vectors are overlapping, their similarity is 1, if they are unrelated, their similarity is consequently 0. The name cosine similarity derives from the behaviour of this similarity measure. We will present the cosine similarity in details, showing its formulation, in Chapter 4, where we used it to measure the mutual similarity of documents belonging to the same collection.

Another approach for measuring the distances between two texts is the probabilistic model, that assumes documents are generated according to a particular distribution of terms, modelled by a probability distribution. A particular technique based on a probabilist approach is Language Models. One of the most common models employed is the Unigram Language Model, for which each term is estimated independently, as apposed to models that consider estimations for groups of terms (e.g. two terms - bigram, three terms - trigrams, ...) [81]. As in the case of the vector space model, this is another example of bag-of-words approach and in Chapter 5 in our experiments on authorship attribution we will make use of this model for conversational documents. In the case of Language Models the similarity between documents is computed with metrics derived from information theory or statistics. Measures of similarity from information theory are the Kullback-Leibler Divergence (KLD, and its particular case the Janson-Shannon divergence) and the Mutual Information (MI), while the $\chi^2$ is an example of a metric derived from statistics. It has to be noted that these metrics (KLD, MI, $\chi^2$) are not similarity measures in a strict sense [26], rather they are used to measure the similarity between texts or documents. Among other applications where they can be employed are features selection, that is the process of deciding the most representative features (e.g. terms) to use to represent an item (e.g. a document or user profile). In Chapter 5 we make use of KLD and $\chi^2$ for both applications: first, for defining the most discriminative terms to be used for representing a user profile, then as a measure of distance between user profiles to identify users.

### 2.1.4 Topic Identification

Topic Modeling is an evolution of Language Models but assumes a more complicated statistical process for generating documents than those employed in the standard Language Models [132, 137, 138]. In Topic Models based approaches to IR and Text Mining, instead, the documents are no longer treated as distributions over terms, rather as distributions over a latent topic space, which has a much lower dimensionality than the original term space. In fact, document representation is done using a low dimensionality topic frequency vector rather than a high dimensional term frequency vector. Each document is then seen as a combination of topics, where each topic is itself defined by a distribution over words. These distributions can be learned directly from the corpus of documents using standard statistical techniques (e.g. Latent Dirichlet allocation - LDA) [45]. Topic Models in their classical definition were first described in [18] and have shown in the past only relatively limited success for ad-hoc (standard Web) IR search [134, 135]. Their widespread adoption has been obstructed by scaling issues for the estimation techniques used, although much recent work has been devoted to developing algorithms for scaling up Topic Model techniques to large document collections [134, 4, 123].

Topic Models based on LDA have being successfully applied to microblogs for new users selection and statuses ranking [102] and a recent work compared the differences between social media (Twitter) and traditional media (e.g. newspapers or online news) employing Topic Models as an instrument of evaluation [140, 52]. The application of LDA on conversational documents is not totally new and has been employed to detect topic change in IRC Logs for the purpose of segmenting the chats [124] or characterising the documents themselves [30]. Older work analysed the interactions between users with respect to topic change [90], or linked this to the users' social network [125].

A limitation of Topic Models in the classical formulations is the a-priori definition of the number of topics. In a dynamic context, such as the conversational one (described in Section 2.3), it might be of a great importance to be able to identify different topics at a different time, or the evolutions of these topics along the conversation, from general ones to specific ones, or vice-versa. A possible solution to this problem might be represented by the usage of Hierarchical Topical Models [17], in which the number of topics is still fixed but different levels of granularity are introduced. This allows to model the switch from general to specific topics (from the top to the bottom of the topic hierarchy tree) or from specific to general topics (from the bottom to the top of the hierarchy tree). For example, some authors studied the best way of finding hierarchies of topics within different sources (Blog, Q&A systems, Twitter) [143], while others decided to use LDA for finding topical and useful Twitter documents [103]. Other recent findings, however, indicate that ad-hoc techniques such as two-step labelling or threshold noise filtering [23] as well as spectral clustering [119], might perform better than hierarchical models. The problem of identifying hierarchies of topics, in fact, is not a trivial problem and would have required many additional studies. For these reasons

we left it to future work, as we left to future work its combination with the models we developed in Chapter 5 for author identification [110]. We will explain these future research directions in Chapter 6, while in the rest of the dissertation we focused on general properties of documents, looking at the topical properties of their words (called burstiness and defined in Section 4.2.4). In the authorship attribution problem we focused on predefined topics only, which are the themes or channels of the IRC logs constituting the collection employed in our study of Chapter 5.

## 2.2 Text Mining

Since some of the techniques described later in the dissertation do not fall into a strict definition of IR, they will be presented in the next section. They fall within the broader area of Text Mining, which somehow intersects the area of IR and creates a cross-boundary region where our work can be located (See Figure 1.1).

There is not a single definition of *Text Mining* but researchers [54] seem to agree with the one originally proposed by Hearst [49], where Text [Data] Mining was associated with the discovery of novel information from textual documents. In this work authors make also a sharp distinction between Text Mining and IR. Text Mining focuses on the extraction of new knowledge from textual data, while IR focuses on finding relevant documents according to a user's needs (as we explained in the previous section). We already presented the core components of IR and the broad definition of Text Mining allows for different techniques to be considered useful for the purpose of extracting new information from textual documents. These techniques range from generative to discriminative models, from supervised to unsupervised approaches. Generative methods are approaches that explicitly or implicitly model the distribution of inputs as well as the outputs, because by sampling from them we can generate synthetic data points in the input space. Approaches that model the posterior probabilities directly are called discriminative models instead. Similarly, in supervised learning the available data comprise both examples of the input and their corresponding target values, while for unsupervised methods the target values are not available [16]. These techniques belong also to the Machine Learning techniques, the aim of which is to develop systems capable of learning from data. In the case of Text Mining the data are texts and thanks to the learning systems, the information extracted from the texts can later be used in other different applications, including IR.

An overview of different Text Mining techniques and applications can be found in two surveys [13, 14] where the following areas of application of Text Mining are identified:

- Clustering and Classification;
- Document (Information) Retrieval and representation;

- Email surveillance and filtering;

- Anomaly and Trend detection.

The same areas are also highlighted in a more recent survey [15] where, instead, the attention is moved towards some interesting applications: keyword extraction from text [109], multilingual document clustering [9], text visualisation [100], spam email identification [61] and text mining and cybercrime [70]. These applications are interesting because they are all somehow related to IR, in particular the last work is interesting. Indeed, this is one of the first studies where Text Mining and conversational documents are directly linked. We are presenting the conversational documents in general in the next section and in different ways in all the other parts of the dissertation (Chapters from 3 to 5). Conversational documents, in fact, can be considered one of the main representatives of the novel class of online user-generated documents. These documents represent, as said in Section 2.1, the evolution of the traditional edited documents, such as newspapers or letters. We will investigate this relationship in detail in Chapter 4.

While recently a lot of researchers have started to investigate some classes of these online user-generated documents (e.g. blogs, discussion fora, Twitter as microblogging platformss), for other documents only little and fragmented research has been conducted. Among the documents needing in-depth study are conversational documents which we have made the focus of our dissertation. For this reason in the next section we will report the theoretical characteristics of conversational documents according to literature, while in Chapter 3 we will list the relevant collections currently available together with the one we developed. Furthermore, in Chapter 4 we will provide an experimental analysis on the characteristics of these documents, with a comparison between classes of online user-generated documents and those from more traditional collections.

To conclude this overview of Text Mining, it is worth mentioning a concept around which we were working and for which the role of Text Mining is central. It is the relationship between Text Mining and authorship, which has been treated in works on text mining and cybercrime [70] as well as in others about authorship identification [72, 77, 78]. The aim of these works is to extract novel knowledge about the people involved in the conversation, for the purpose of identifying their misbehaviour (in the case of cybercrime) or profiling them to provide better services (in the case of blogs or recommender systems). Given this strong relationship, we will later investigate the literature (Sections 2.4) for a better understanding of the area of authorship attribution, while in the immediate next section we will give an introduction to conversational documents.

## 2.3 Conversations and Conversational Texts

In this section we define the theoretical characteristics of conversational documents. While the observation of the actual features of these documents is part of our contribution and is extensively treated in Chapter 4, the overview given here is a general description and serves as an introduction.

In this dissertation we employ textual documents as main input data for our studies, in particular those textual documents produced by users online. Among these, however, we focus on a particular category of documents, those that present conversational properties. Example of conversational documents are those produced by users writing in IRC systems (e.g. freenode[1], quakenet[2], etc ) or instant messaging systems.

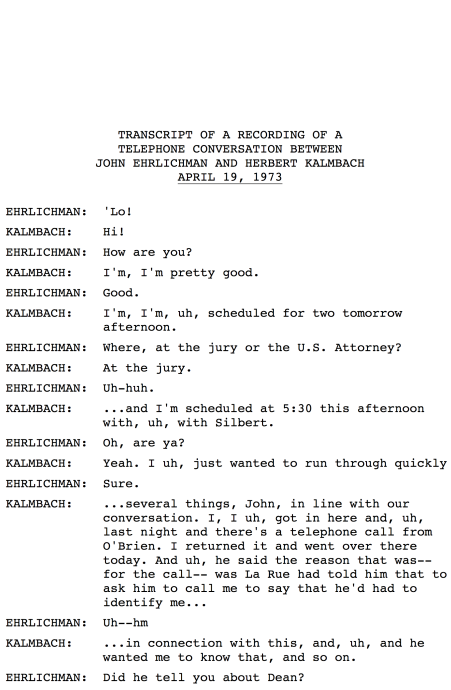
To understand the properties of conversational documents produced online, we first have to observe the properties of a "regular" conversation produced offline, i.e. a dialog between two (or more) people who physically meet. If we observe an in-person conversation, we see two or more people talking together, in the same place and at the same time, probably discussing some topics they share. More formally, a conversation is defined as a series of dialog acts, composed by a series of speech acts which share a common ground and which actively contribute to the discourse [67]. This contribution to the conversation is divided in two parts: presentation and acceptance. In the first part, the speech act (message) is produced by the first speaker, while in the second the other interlocutor (the hearer) has to "acknowledge" the reception of the message: by nodding, by showing interest or by citing (also verbatim) what was said. These actions repeated continuously during the conversation originate different dialog acts, thus a discourse. Moreover, the principle of grounding implicitly introduced above says that in a conversation, unlike in a monologue, the partners must share a common ground that evolves during the conversation.

Other important features of the in-person conversations is the turn and turn-taking rule. This rule establishes who is supposed to speak next, given the current speaker. Besides mentioning the actual rules, which can be found in [67], it is interesting to notice that some of these properties may also involve non-verbal signals, like gestures or facial expressions, that are not straightforward to emulate when the conversation moves online. One solution to this problem are the so-called *emoticons* (a particular sequence of punctuations associated with some emotions) that became the standard-de-facto for emphasising particular expressions or tried to emulate real facial expressions. We will study the use of emoticons later in Chapter 4, as an example of the different properties of online communications.

Finally, another important element of the conversation is *silence*, or absence of speech acts, that generally means disagreement or intent of closing the conversation. This is especially true if silence is followed or anticipated by the so called *dialog pairs*,

[1] `http://freenode.net`

[2] `http://www.quakenet.org`



TRANSCRIPT OF A RECORDING OF A
TELEPHONE CONVERSATION BETWEEN
JOHN EHRLICHMAN AND HERBERT KALMBACH
APRIL 19, 1973

EHRLICHMAN: 'Lo!

KALMBACH: Hi!

EHRLICHMAN: How are you?

KALMBACH: I'm, I'm pretty good.

EHRLICHMAN: Good.

KALMBACH: I'm, I'm, uh, scheduled for two tomorrow afternoon.

EHRLICHMAN: Where, at the jury or the U.S. Attorney?

KALMBACH: At the jury.

EHRLICHMAN: Uh-huh.

KALMBACH: ...and I'm scheduled at 5:30 this afternoon with, uh, with Silbert.

EHRLICHMAN: Oh, are ya?

KALMBACH: Yeah. I uh, just wanted to run through quickly

EHRLICHMAN: Sure.

KALMBACH: ...several things, John, in line with our conversation. I, I uh, got in here and, uh, last night and there's a telephone call from O'Brien. I returned it and went over there today. And uh, he said the reason that was-- for the call-- was La Rue had told him that to ask him to call me to say that he'd had to identify me...

EHRLICHMAN: Uh--hm

KALMBACH: ...in connection with this, and, uh, and he wanted me to know that, and so on.

EHRLICHMAN: Did he tell you about Dean?

*(a)* Transcript of a telephone conversation

[02:46] <sxr2rw> can anyone suggest and online spyware/addware/game/blocker?
[02:46] <sxr2rw> I know about adaware pe and spybot
[02:46] <sxr2rw> But I need something for my computers that stay resident and is free.
[02:51] <dzzwygyzg> try editing your hosts file
[03:48] <zuzu> when did this channel become moderated?
[03:49] <szgz2zyzdusya> hey
[03:49] <zuzu> hey
[03:49] <szgz2zyzdusya> whats blendering guru
[03:50] <zuzu> heh, nothing...no free time anymore
[03:50] <szgz2zyzdusya> wow
[03:50] <szgz2zyzdusya> i hear u
[03:50] <zuzu> i'm severly lacking in my contribution to open source projects lately
[03:50] <szgz2zyzdusya> wow
[03:50] <szgz2zyzdusya> what did u used to do
[03:51] <zuzu> many things
[03:51] <zuzu> don't remember them all
[03:51] <szgz2zyzdusya> programming?
[03:51] <szgz2zyzdusya> lol
[03:51] <zuzu> yes
[03:51] <szgz2zyzdusya> cool
[03:51] <szgz2zyzdusya> hey if ur making money thats what counts
[03:51] <zuzu> i think i'm still a dev for php, not sure
[03:52] <szgz2zyzdusya> hehe
[03:52] <szgz2zyzdusya> i make music. a little 3d and thats about it
[03:52] <zuzu> i sit in front of a computer and write code all day like a tool
[03:52] <szgz2zyzdusya> wow
[03:53] <szgz2zyzdusya> thats scary
[03:53] <szgz2zyzdusya> must be very stressful
[03:53] <zuzu> at times
[03:53] <zuzu> you come here much?
[03:53] <szgz2zyzdusya> when i have the time
[03:53] <zuzu> when did they add +m to the channel modes?
[03:53] <szgz2zyzdusya> i have no idea
[03:53] <sunlxrnn> guru: about 2 weeks ago
[03:54] <szgz2zyzdusya> oh
[03:54] <szgz2zyzdusya> hey sur
[03:54] <sunlxrnn> hi sanabriamusic
[03:54] <zuzu> Surphaze: wow, guess i haven't been active for a while so that explains it
[03:54] <zuzu> Surphaze: any idea what the reason was?
[03:54] <szgz2zyzdusya> i just updated my page a little
http://sanabriamusic.bravehost.com/home.html
[03:54] <szgz2zyzdusya> i am a new blender user
[03:55] <zuzu> speaking of blender, i installed the latest version of it a few days ago and still haven't even looked at it
[03:55] <szgz2zyzdusya> lol wow
[03:55] <szgz2zyzdusya> is better to do it with calmness

*(b)* Online conversation

*Figure 2.2.* Examples of conversations on different media: the telephone (a transcript from the Watergate case, from `http://www.nixonlibrary.gov`) and the Internet (an extract from `http://www.irclog.org`).

which are conventional sets of speech acts that generally come together, like greetings formula. This is even more evident during online conversations, where a long break between two messages might mean an interruption of the communication due to technical issues. This however might also mean the end of the current chat section and the end of the conversation, as in the majority of the cases.

The problem of finding when an online conversation is starting and ending is called *chat disentanglement*. Despite the fact that different techniques have already been proposed to solve it [33, 32, 133], it is still a challenging problem. While performing chat disentanglement seems obvious for person to person conversations or on the telephone, when people end the conversation by physically leaving it, the same does not happen when considering online chat sessions. In this case conversations might last for a longer and more dispersed time and might also involve more than just two users, who alternate in the conversations. In our work we will not address this problem, which is a difficult one [31], instead we will make some simplifications on the way the online conversations are disentangled and will focus on extracting novel information from the textual messages produced in the conversation.

We can, therefore, summarise all the properties of the "regular" chat presented above in the following big categories, which will later serve us as a reference and simplify the analysis of the online chat. A "regular" conversation involves [67]:

1. proximity in time, given the turn-taking rules, with little or no silence, otherwise the conversation is considered finished;

2. proximity in space, since interlocutors make use of non-verbal expressions, gestures, facial expressions and coreference to sustain the conversation and make it lively;

3. proximity in topics, given the grounding principle that makes the interlocutors to share information and the contribution principle, that enforce the participants in the chat to add new knowledge to the shared common ground.

These three principles are still valid when moving to online conversations. In fact, if we relax the proximity in space property (condition 2) and accept the introduction of novel ways of communicating (like emoticons), we will notice that all the properties are still valid. What changed is simply the medium that transports the message, the speech acts. The conversational pattern repeats, as it does for "regular" conversations, e.g. on the phone. The phone is just another medium that (should) facilitates conversations, relaxing the proximity in space condition and substituting the non-verbal and facial expressions with voice emphasis or novel interjections[3].

In Figure 2.2 we report two examples of conversations on a particular medium, where the proximity in space condition is relaxed. It is interesting to notice that despite being one conversation on the phone and the other online, some patterns are recurring, such as the use of interjections *uh-uhu*, *oh*, *yeah*, *wow*, *hey* or the cross mentioning of elements of the question in the answer, too (e.g. *Where, at the jury or the U.S. Attorney? ... At the jury.* or *I am a new blender user ... specking of blender, ...*). Furthermore, it is worth mentioning that algorithms that aim to disentangle online conversations [33, 32, 133] are generally based on the above mentioned three principles. In fact, to reconstruct online dialogs one needs proximity in time of the messages, their topical similarity and the coreference of elements in the text (e.g. names, pronouns, mentions).

To conclude this section on general properties of conversations, we should mention an interesting booklet on "text mining in conversations" [21]. This booklet is more focused on the analysis of chat transcript for summarisation and abstraction but contains a nice introduction on the general topic of conversations and on the relevance of topic identification (topic labelling and topic mining) for the conversations, too.

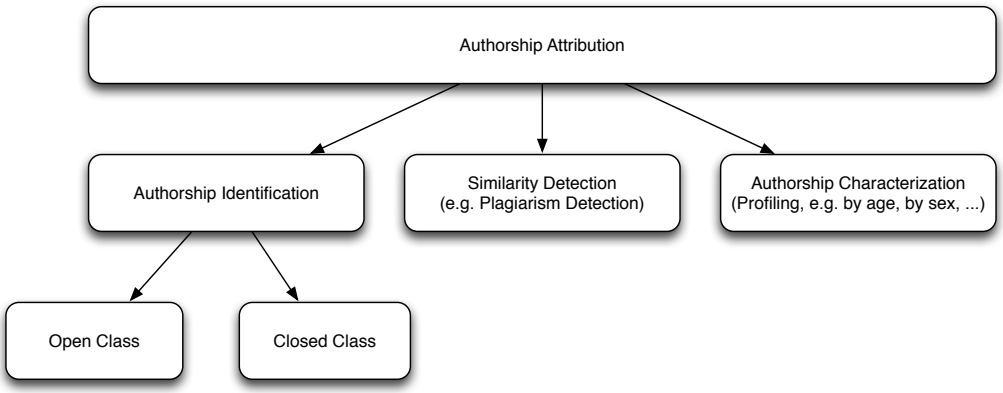



*Figure 2.3.* Different categories of the authorship attribution problem.

## 2.4 Authorship Attribution

The problem of investigating the authorship of a standard textual document is a multifaceted one, which has been widely treated in the literature. There is an agreement among researchers [64, 121, 73] to divide the problem into the following classes of problems, based on the number of authors involved and the purpose of authorship analysis. We summarize these different classes in Figure 2.3.

**Authorship Identification** whose focus is on finding or validating the author of a document; sometimes in the literature authorship attribution is also used for authorship identification.

**Similarity Detection** aims at finding the variations in the writing style of an author or the similarities among the writings of different authors, mostly for the purpose of detecting plagiarism. Plagiarism detection aims at detecting wether a text is original or has been copied ("plagiarised") from other sources without acknowledgement.

**Authorship Characterisation** is the task of assigning the writings of an author to a set of categories according to the author's sociolinguistic attributes. Examples of common attributes are gender, language background, and education level.

Authorship identification is traditionally divided into two classes of problems, based on the evidence available: *closed class* and *open class*. In the closed class problem, given a text one should attribute it to one author from a predefined group of authors, where the training and testing sets contain the same authors. In the open class problem, however, the set of possible authors may not be limited to a predefined subset but may

[3]Interjection: an exclamation, especially as a part of speech (e.g. ah!, dear me!) from *The Oxford English Dictionary*

involve other authors from outside the predefined set [64] and these should generally be identified as "unknown". The authorship characterisation problem is mostly known as Profiling (or Stylometry [51, 64]) and focuses on identifying properties of the authors of a given text, such as age, sex, dialect, etc. [3, 73]. In the Similarity Detection problem, instead, the focus is on discovering if a certain part of an author's text has been reused in another [98]. Reusing a text with explicit acknowledgement is considered citation or quotation, while failing to give proper credit to the original source when reusing a text is called plagiarism [99].

In our work we decided to focus on the authorship identification problem (Chapter 5), in particular in the subclass of closed problems, ignoring the Similarity Detection task. We were also involved in a small project where we worked on authorship characterisation. Since that was a self-contained work but gave us a lot of intuitions and ideas also for the authorship identification part, we will report it in Appendix A. We decided to ignore Similarity Detection because it is not yet a problem of interest in the case of short user-generated or conversational documents. Moreover, the nature of these documents does not make them suitable for that kind of analysis. Finally, for solving the similarity detection problem one should apply techniques that are very different from the ones used in the other two authorship scenarios.

There is an abundance of previous work on authorship identification, the first of which dates back to 1887 [84] and interested readers might have a look at the work of Holmes and Juola [64] for a more in-depth study of the history of stylometry. Despite this, the problem of authorship identification is still of a great interest due to the availability in the last year of novel datasets and new processing techniques. While in the past researchers concentrated their efforts on collections of formal edited documents, such as letters or newspaper articles [64], in recent years attention has moved to online user-generated documents, such as emails or online conversations. We consider these documents novel because they were not analysed in the past and they present characteristics that make them more difficult to be analysed. In the previous section, we already highlighted some of the features of conversational documents and we will analyse the features of these online user-generated documents in greater detail in Chapter 4. In Chapter 5, instead, we will give a detailed overview of previous work within the field of Authorship Identification, that was too specific to be mentioned in this chapter.

### 2.4.1 The PAN Evaluation Laboratory

The yearly PAN laboratory (lab) competition is an example of such ongoing research in the field of authorship identification [2, 65]. The acronym PAN stands for *evaluation lab on uncovering plagiarism, authorship, and social software misuse*[4] and it now reflects the nature of the laboratory. PAN originally took place in 2007 as a workshop of SIGIR, the ACM Special Interest Group in Information Retrieval, one of the main venues for

[4] `http://pan.webis.de`

researchers in IR, with the focus on Plagiarism Analysis, Authorship Identification and Near-Duplicate Detection [122]. Since then the PAN lab has taken place each year regularly and since 2010 it has been hosted by the Conference and Labs of the Evaluation Forum[5] (CLEF). PAN has become one of the biggest and most active (i.e. attended) laboratories of CLEF and has turned into a cutting-edge research venue. In 2013, as part of the evaluation process, it introduced a software submission procedure that allowed for measuring not only performances in terms of precision, recall or standard performance metrics, but also in terms of computational time[6] [42]. Moreover, under the hat of PAN different novel experimental campaigns have been conducted: besides the classical plagiarism identification and authorship identification, backbones of PAN, in 2010 and 2011 a task on Wikipedia vandalism detection was introduced. In 2012 we introduced an innovative task called *Sexual Predator Identification in Online Conversations* [56] and in 2013 we helped with the *Users Profiling in Social Media* task [104].

For simplicity, in the task we only concentrated on the identification of a "sexual predator" inside a chat and we did not deal with other kinds of misbehaviour or media. A "sexual predator" is defined in the New Oxford American Dictionary as *"a person or group that ruthlessly exploits others"*, while Wikipedia noticed how the definition *"is used pejoratively to describe a person seen as obtaining or trying to obtain sexual contact with another person in a metaphorically "predatory" manner"*. We refer to these interpretations of the term "sexual predator" for the competition.

In defining the Sexual Predator Identification in the online conversations task, we were inspired by some previous works [83, 70, 96] that addressed a similar problem, despite the fact that each author conducted his experiment independently and on a dedicated collection. For this reason we wanted to advance the state of the art and therefore organised a task that served as a unique venue with a uniform testbed for all interested researchers. We developed a challenging collection that all of the 16 participants in the task had to use as a reference for their experiments. Moreover, to the best of our knowledge we were the first to propose the following two kinds of problems to the research community. The formulation of the problem is the following:

*Given a collection containing chat logs involving two (or more) persons the participants had to:*

1. *identify the predators among all users in the different conversations (problem 1)*
2. *identify the part (the lines) of the conversations which are the most distinctive of the predator behaviour (problem 2)*.

We will present and discuss the features of the developed collection in Chapter 3, where we will also describe the rationale behind it and present other collections already

[5] `http://www.clef-initiative.eu`

[6] Measuring the computational time was possible thanks to the software submission system, that allowed to run each participants software on the same computing resources.

introduced in the literature. In the same chapter we will also motivate the unsuitability of these collections for our work. Finally we will report the participants' approaches and results to the task in Appendix B, as they are not strictly relevant to this dissertation.

To conclude, we should consider PAN interesting not only for researchers in the authorship attribution field, but also for those interested in text mining or text processing. This is due to the challenges involved in the nature of documents PAN tasks have to deal with. Challenges might arise when preprocessing them (how to expand short documents? How to deal with dozens of spelling mistakes?) as well as when analysing them (what are the best features to profile users?). They should also be interesting for companies, in particular those providing services for facilitating user conversations and collaboration. For example, providers of the IRC service or Social Networks might want to be able to identify particular users based on their behaviour (or misbehaviour) or be able to profile them to better target their advertisement.

## 2.5 Summary

In this section we presented the main research areas touched during our work, in particular IR and Text Mining. We highlighted the most important aspects of IR employed in the rest of the dissertation, from text processing to similarity evaluation. We then mentioned the Text Mining methods useful for conversational documents and gave a general introduction on the general features of short and conversational documents, that are the main object of our investigation. To conclude the chapter, we illustrated the state of the art of authorship attribution, identifying the three problems it might be divided into (Authorship Identification, Similarity Detection, Authorship Characterisation). We then focused on a particular problem, authorship identification, presented the competition we organised within PAN at CLEF and gave an extended overview of the literature for authorship identification. In the next chapters we will make use of all the concepts explained here and we are going in-depth on some arguments briefly introduced in this chapter (e.g. existing collections, observed properties of conversational documents, authorship identification for conversational documents).

# Chapter 3

# A Novel Corpus of Conversational Documents

*Don't do anything that someone else can do. Don't undertake a project unless it is manifestly important and nearly impossible.*

Edwin H. Land

This chapter is entirely dedicated to collections of textual documents directly related to our work. We will first introduce examples of documents for each type of collection, from chats to newspapers. We will then list all the collections and provide a description for each of them. In the first part of the chapter (Section 3.1) we will present the established collections of traditional documents widely used in IR (Section 3.1.3) and the collections of online user-generated documents already introduced and employed in the literature in the last few years (Section 3.1.2).

In the second part of the chapter (Section 3.2) we will present the collection of conversational documents that we created and that is part of the contribution of our work. We will illustrate the motivation and methodology behind this collection, created as part of the "Sexual Predator Identification" task in PAN 2012 and give a detailed overview of the different sources constituting the collection. We will indeed employ part of this collection for our studies in Chapter 5, while interested users might want to look at the result of the "Sexual Predator Identification" task available in Appedix B.

## 3.1 Collections in the Literature

In this first section we are introducing collections that are already employed in the literature. We will start by giving an example of the different types of documents in each collection, then we will list the important datasets relevant to the dissertation.

### 3.1.1 Example of Traditional and User-Generated Documents

In the previous chapters we already mentioned the differences between traditional IR collections and collections of online user-generated documents. In this section we want to give examples of these documents to facilitate the understanding of the different problems that might arise when dealing with each type of documents.

In Figure 3.1 we illustrate two examples of documents traditionally employed in IR: newspaper articles. The two articles were randomly chosen from two different collections that are also used in our experiments and that are presented in more detail in Section 3.1.3. The first text is from The Wall Street Journal, included in the TREC Ad-hoc collection, while the second one is from La Stampa, an Italian newspaper part of the 2004 CLEF collection (Section 3.1.3). From a first rough qualitative observation of the two documents, we can notice a careful writing style and a limited presence of author's sentiment in reporting the objective events. Things would be different if we dealt with other types of documents, for example letters or diaries. These are however less popular in the literature, therefore we did not take them into consideration in our studies and focused on newspaper articles instead, which are the most employed collections in literature. One last remark is about the documents' length, which is in the order of hundreds of words and that is a distinctive trait of these traditional collections.

Online user-generated documents, in fact, are different from the traditional ones, especially for their length, as we will state in Section 3.1.2. Moreover, since it is only in the last decade that these documents became so popular, we characterise them with the adjective "novel". In Figure 3.2 we present some examples of these novel collections. A first observation highlights the media these novel documents were generated from: the Internet. This influences the nature of these documents generated using different online services. Online chats, for example, are mostly used by people who want to have simultaneous conversations, emulating in person conversations. In microblogs the conversational aspect is less evident, since the interlocutors might not answer or do answer at a later stage. In blogs, fora and reviews, instead, people are more reflective since they are not communicating in a synchronous way and therefore they produce texts that are longer and more structured. These are just some of the first noticeable differences between online user-generated documents and traditional ones. Other differences, such as the structure of the documents and their topical content, are highlighted in detail in the next chapter (Chapter 4), where we will study both the qualitative and quantitative differences of the documents of traditional and novel collections. In the following sections, instead, we illustrate in detail the different collections, which are listed in Table 3.1. The same table shows which collection was employed in the dissertation as well as which collection we decided not to use (and why).

Chairman Lloyd Bentsen of the Senate Finance Committee is considering imposing a tax on the short-term trades of securities by currently tax-free pension funds. In an interview, the Texas Democrat said he was considering the new levy "among other options" as part of his drive to raise enough taxes to meet deficit-reduction goals this year. He added that he hasn't made any decision about the pension option. The notion stems from Sen. Bentsen's interest in encouraging long-range thinking by American business interests and in discouraging short-term trading for quick profits. "I'm deeply concerned about the churning of stocks and short-term horizons," he said. "That's been particularly true of the pension managers and some of the tax-free funds" Sen. Bentsen and his committee soon will begin drafting legislation designed to raise some $ 5.3 billion in taxes to reduce the budget deficit. A package of proposals meeting a similar obligation is pending in the House, and is scheduled to be voted on as early as next week. The House version of the revenue-raising plan, passed last week by the House Ways and Means Committee, doesn't include a similar tax on pension-fund trades. A proposal to tax pension-fund trading would cause tremors on Wall Street, where such a levy would cut broker commissions by reducing trading. The tax would also be a clear violation of President Bush's "no tax" pledge. But the idea would be a shrewd political response by Sen. Bentsen to efforts by Treasury Secretary Nicholas Brady to devise measures to discourage short-term thinking by U.S. businesses. A senior Treasury official said the department hasn't considered any proposals to tax pension funds, and noted that that would clearly violate the president's tax stance. But some administration officials are sympathetic to the idea. Taxing pension funds also has bipartisan support on Capitol Hill. Republican Sen. Nancy Kassebaum of Kansas wrote a column in the New York Times Aug. 30 proposing a 20% capital-gains tax on pension investments held for less than three months. Under Sen. Kassebaum's proposal, the tax would be reduced each quarter and phased out for assets held more than a year. Sen. Kassebaum said she sent a copy of the column to Sen. Bentsen last week and received a note from the Finance Committee chairman saying he'd like to discuss the idea with her. Staff aides to Sen. Kassebaum and Senate Minority Leader Robert Dole (R., Kan.) also met with Treasury officials last week to promote the idea of taxing pension funds.

*(a)* The Wall Street Journal

NEW YORK UNA corsa in taxi a West Palm Beach, in Florida, sabato sera. Ci sono 6 dollari da pagare. I clienti non li hanno, o comunque non hanno voglia di sborsarli. L'autista sta per arrabbiarsi ma non ne ha il tempo: dal sedile posteriore spunta la canna di una pistola che si va a posare proprio sulla sua nuca. Parte un colpo e il conto e' saldato. Uno di quegli episodi di ordinaria violenza cosi' frequenti nelle citta' americane? Questo e' speciale: la mano che ha premuto il grilletto appartiene a una bambina di tredici anni che frequenta la prima media e conduce una vita - forse non felice ma certamente non da delinquenza di strada - con la madre. Il tassista,Yves Quettant di 39 anni, muore sul colpo. La bambina e i suoi due amici assieme ai quali era salita sul taxi si allontanano senza troppa fretta. Qualcuno li nota e poi li descrive alla polizia.Vengono arrestati e lei racconta tutto. Ma non ci sono crolli nella sua confessione. Non era brilla, forse era in preda all'eccitazione del sabato sera ma non al punto da non sapere cosa facesse. Anzi lo sa talmente bene che spiega tutto per filo e per segno, senza mai alterare il tono della sua voce. Non un pianto, non un rimorso,dice ancora sconvolto il sergente John English, che ha raccolto la deposizione della bambina. In lei c'era solo una freddezza assoluta.Ha compiuto un omicidio a sangue freddo e la cosa non l'ha minimamente scalfita. Poco prima, quando era tornata a casa, era stato lo stesso con la madre. Non mi ha detto nulla, non ha fatto il minimo accenno a qualcosa di insolito accadutole nella serata; per lei era stata una serata di divertimento come altre, dice la donna.Quando sono venuti ad arrestare la figlia, lei l'ha accompagnata al posto di polizia convinta che di li' a poco l'equivoco sarebbe stato chiarito. Come poteva la sua bambina avere ucciso una persona e starsene cosi' calma, sorridente e tranquilla? E invece ecco che al commissariato il sergente English non deve neanche essere particolarmente stringente. La giovane assassina ammette tutto come se raccontasse un film: Non me ne importa nulla. Solito discorso sulla facilita' con cui chiunque in questo Paese puo' procurarsi un'arma? Solite considerazioni sulla familiarita' con la violenza,per cui la soppressione di una vita cessa di costituire un tabù? Il sergente English non sa di queste cose. Fa il suo mestiere da dieci anni, ne ha viste tante, ma di fronte a questa bambina ha un solo commento: Agghiacciante.

*(b)* La Stampa

*Figure 3.1.* Two examples of traditional documents: newspaper articles. The first article is from "The Wall Street Journal", in English language, while the second is from "La Stampa", an Italian newspaper.

i wanted u 2 know i wuld b gone for a coupla dayz cuz she in opelika n dad say he dont wanna go back n forth each day

*(a)* Online Chat: IRC logs.

Data at your fingertips: A new version of the Google #Analytics App for #Android `http://ow.ly/pnuIH` #SEO #search #datamining

*(b)* Microblog: Twitter.

I ran this game for a few years on a different forum site I used to be on.. Pic any type of movie,anything you can think of.. There are 10 posts per topic /type of movie. The person posting the # 10 reply picks a new topic & posts also #1 of that next topic.. All posts must be numbered so we know when & who is tenth .. Heres the 1st topic... Harrison Ford Movies: #1)- Indiana Jones & The Last Crusade

*(c)* Blog: MySpace.

It's actually pretty amazing just how vastly the difference in perception can be when doing things like flirting, as you say. When one party is thinking about the interaction in a different context than the other there is HUGE room for misunderstanding even when the signals are unambiguous. I'd also agree that I'm skeptical about this kind of system; my gut tells me it would need to have a huge amount of information on each person in the chat, across multiple chats with multiple parties, to even begin to build up a profile that could have a hope in hell of being accurate. However, it may be that they're looking to see if that isn't the case - we may need much less information than we think to come to these kinds of conclusions.Some people are very, very good at reading people - can take one look at someone, see a relatively small number of factors but put them together into a framework that suggests lots of other probabilities about the person that turns out to be startlingly accurate. I could see them trying this to see if it's possible for algorithms to pull off this same kind of feat. If they find there's something to it, it's cool and worth further exploration; if they don't find something to it they can at least start to figure out what the lower boundary might be for the amount of data needed to start getting there.

*(d)* Forum: Slashdot.

It seems that Dawson intentionally talks with the dying Jen Brooks such as video e-mail from Beyond the Grave in the series final,the second part in a second jump before a half a decade, with Dawson and co. in adulthood (unofficially along the Bund, Joey, Jen and Andie strike in 'Future Tense' meet again in five years). This reinforces the idea that the series works better than FINISHED coda in a season more than four to six season - I go so far that the insistence by the successor of 'Coda' , a successor superfluous, from the obligation to begin Time ( 'Winner' ), with 'all good things, At least, so that a merciful farewell gift Jen, over the next two years vainly trying to retrieve a niche in Dawsonland. I think Kevin Williamson has decided to sacrifice Jen in the finals finally discover because they are the martyrs of the 'Dawson's Creek'. its the best part of dawson"s creek i ever seen, belive me u must be enjoyed it

*(e)* Review: Ciao.

Please find attached the note I prepared in response to the request for information received last Monday (sorry, it is in Spanish, please let me know if you have someone who can quickly put it in English or if you would like me to translate and re-send it) . I have limited the information to what was actually requested. All of the data contained in the same is duly documented and can be checked in the stack of related documents I faxed you and Mark on Wednesday night. Please let me know if you have any questions or comments. Regards,

*(f)* Email: Enron.

*Figure 3.2.* Examples of different collections of online user-generated documents. Each text represents a single instance of the different documents.

| Dataset | Type | # Posts | Employed in |
|---|---|---|---|
| CAW 2.0 - Ciao | Review | 20K | Chapter 4 |
| CAW 2.0 - Kongregate | Chat | 145K | Chapter 4 |
| CAW 2.0 - Twitter | Microblog | 900K | Chapter 4 |
| CAW 2.0 - Myspace | Blog | 380K | Chapter 4 |
| CAW 2.0 - Slashdot | Forum | 140K | Chapter 4 |
| MS Twitter Conversations | Microblog | 1300K | Chapter 5<br>Chapter 3 |
| CLEF-PAN 2012 SPI | Chat | 345K | Chapter 5<br>Appendix B |

*(a)* Collections of online user-generated documents employed in our studies.

| Dataset | Type | # Posts | Not employed because |
|---|---|---|---|
| IRC Conversational | Chat | 2K | *collection too small and not topical* |
| NPS Chat Corpus | Chat | 10K | *collection too small and not topical* |
| NUS SMS Corpus | SMS | 10K | *collection too small and not topical* |
| TREC Microblog | Microblog | 16000K | *collection too big; no conversations* |

*(b)* Collections of online user-generated documents not employed in our studies.

| Dataset | Type | # Posts | Employed in |
|---|---|---|---|
| TREC Ad-hoc - WSJ | Newspaper | 210k | Chapter 4 |
| TREC Ad-hoc - FT | Newspaper | 170K | Chapter 4 |
| TREC Ad-hoc - AP | Newspaper | 240K | Chapter 4 & 5 |
| CLEF 2004 - Glasgow Herald | Newspaper | 26K | Chapter 5 |
| CLEF 2004 - La Stampa | Newspaper | 35K | Chapter 5 |

*(c)* Collections of traditional documents.

*Table 3.1.* List of collections and their type, with number of documents or posts per collection.

### 3.1.2 Collections of Online User-generated Documents

Different collections of online user-generated documents can be found in the literature. In this section we are presenting the most relevant and most widely employed collections in the literature. This list includes datasets that were used in this work and other datasets that were not used. Our aim is to give a complete picture of the existing collections in the literature, with a particular attention to conversational (e.g. chats) documents, highlighting their main characteristics. This also serves as justification for the creation of a novel collection, mentioned in the table (CLEF-PAN 2012 SPI) and described later in Section 3.2. We also report two datasets of traditional edited

documents (newspapers), that we employed in our comparative study with online user-generated documents (Chapter 4) and authorship attribution (Chapter 5). It is also to be noted that we did not include in this list of collections those that include transcripts of in-person chat or speech-to-text conversations [21], because these have different characteristics than documents generated online, as illustrated in the previous chapter (Chapter 2).

### The CAW 2.0 Datasets

The first collection we are presenting is the one developed for the Workshop for Content Analysis in Web 2.0 (CAW) introduced by J.Codina et al. [60]. We used it for the first part of our work (Chapter 4), in which we studied the properties of different set of online user-generated documents. The collection, in fact, consists of 5 distinct collections of documents crawled from 5 different online sources: *Ciao* (a movie rating service), *Kongregate* (Internet Relay Chat of online gamers), *Twitter* (short messages), *Myspace* (forum discussions) and *Slashdot* (comments on news-posts). The collection[1] is divided into training and testing sets and for our experiments we only used the training part of the dataset, which was enough for our purposes. We will present the actual statistics of the dataset in Chapter 4 where we will analyse it in detail. However one important aspect is worth mentioning now: its great novelty at the time of its creation. For example, in 2009 when the collection was released, Twitter was just emerging from the startups world but was already included in it. Moreover, in this collection we find both conversational documents and posts in blogs or fora, that were just starting to capture the attention of the research community. For example, the Blog Track in TREC released the first corpus in 2006, while the study on conversations started at the same time or later, between 2008-2009. We can then conclude that this collection was really state-of-the-art and was well suited for a comparative study, as we did. However, we could not use the set on chats for our experiments on authorship attribution (Chapter 5) because it contains 13 conversations only, which would not be enough to conduct any detailed (and not quantitative) analysis. This is the first argument towards the development of a novel collection.

### IRC Conversational Dataset

This dataset was created with the purpose of developing algorithms able to automatically segment online conversations [32]. It was generated by recording all the messages on IRC channel #LINUX at `www.freenode.net`. When we started our work, it seemed that this was the only and best dataset to be used for our analysis, as it was also quite popular among the NLP community [79]. However, for our goals we later realised that this dataset was too small (apparently a single or few sessions) and with some important information missing (e.g. timestamps). The timestamps are employed to divide a

[1] Dataset and details available at `http://caw2.barcelonamedia.org/`

single log file into a set of chat segments (later called threads or conversations). The segments allow for a coherent analysis of the conversations herein contained. Moreover, the dataset captured only a single channel in the big panorama of IRC channels and providers, while we ideally aimed at a broader dataset, for example containing a lot of channels with different topicalities (as we are doing in our collections presented in Section 3.2).

We had a similar problem with the dataset presented by Layton et al. in [78], which was crawled from a single IRC channel #Ubuntu and then reduced to a set of around 2500 messages only.

### The NPS Chat Corpus

This corpus consists of 10,567 posts out of approximately 500,000 posts gathered from various online chat services as explained by Forsythand and Martell [38]. It is distributed as part of the Natural Language Toolkit (NLTK)[2] or through the Linguistic Data Consortium (LDC)[3]. Like the IRC Conversational Dataset, it contains a set of POS annotated posts divided by author characteristics (i.e. age and sex) but it does not contain any information on the original source of such messages (i.e. from which online service they originated) and on the exact extension in time of each chat. In fact, messages do not contain any timestamp information or any information about the length of the thread they are in. If we assume that each file is a conversation or thread, then also this dataset contains too few threads to be employed in our later analysis (Chapters 4 and 5). Moreover, the collection does not contain a description of the topics associated with each conversation, that might help in processing the texts. Finally, it is interesting to notice that this dataset arose from one of the most active groups in the field of conversational documents analysis. However, being this group strictly linked with organisations devoted to the national (U.S.) security[4], we believe that only limited information can be provided to the public. For example, some more detailed information on the larger NPS Chat dataset is only made available in thesis works (e.g. [79] or [30]), while this larger dataset is not made publicly available.

### NUS SMS Corpus Base License

This collection is slightly different from the previous ones because it does not include logs of IRC chats but only logs of SMS exchanges. It was developed at the National University of Singapore (NUS) and contains hundreds of messages collected on a voluntary basis among computer science students. After its creation in 2005 described

[2] `http://nltk.org`

[3] `http://www.ldc.upenn.edu`, catalog number LDC2010T05

[4] NPS stands for Naval Postgraduate Schoool, whose mission is "to provide high-quality, relevant and unique advanced education and research programs that increase the combat effectiveness of the Naval Services, other Armed Forces of the U.S. and our partners, to enhance our national security" according to its website `http://nps.edu`.

by How and yen Kan [55], it was left without any update for quite some time, which is one of the reasons why we did not consider it in our work. In the last few years (2012-1013), however, it has been updated on a regular basis and it is now growing week after week, making it an online corpus, as announced by Chen and Kan [24]. Due to the low number of interlocutors and the outdatedness of the collection at the time of doing experiments, we did not consider it in the current work but left it for future studies.

### Microsoft Conversation in Twitter Collection

This collection contains a corpus of 1.3 million Twitter conversations, which the authors made available in 2010 [107]. However, it was suddenly removed from the Internet[5], possibly due to violations of Twitter's terms of service, which do not allow Twitter messages to be redistributed. In their work the authors [107] identified sets of users "talking" together in the Twitter collection and studied this behaviour. They realised that "the proportion of posts on Twitter that are conversational in nature are somewhere around 37%". These conclusions are interesting and together with our observations of Chapter 4 led us to choose this collection for our extended experiments on authorship identification in conversational documents in Chapter 5. This served also as inspiration for some complementary experiments on users characterisation conducted on a custom Twitter dataset and presented in Appendix A.

### The Microblog Track in TREC Corpus

The first Microblog collection[6] was released in 2011 as part of the first Microblog Track in TREC. It represented the "evolution" of the previous Blog Track, that ceased in 2010 [113]. The organiser of TREC 2011 Microblog track released a tool for obtaining identifiers for approximately 16 million Tweets and each participant in the track had to autonomously download every Twitter message with the provided tool, in order not to violate the Twitter service agreement. The corpus was designed to be a reusable and representative sample of the twittersphere, including both important and spam tweets, therefore a must-have for all researchers interested in analysing documents from social media. This collection served primarily as a testbed for the problem of ad-hoc retrieval, with attention to the temporal dimension, i.e. retrieval of important past tweets given a query and a certain date. In 2013 the track moved to an API based collection (collection-as-a-service), which allows users to query it and get tweets accordingly, instead of downloading a massive number of documents (around 240 million, according

[5] A trace of the original page promoting the collections can still be found here: `http://web.archive.org/web/20100606154107/http://research.microsoft.com/en-us/downloads/8f8d5323-0732-4ba0-8c6d-a5304967cc3f/default.aspx`

[6] `http://trec.nist.gov/data/tweets/`

to the organizers), as it was in the first edition. Unlike in the previous Twitter collection, these do not contain any grouping of the messages into conversations. Moreover they are really huge collections. For these reasons the identification of the conversations and the filtering of non relevant messages (e.g. in a language other than English or spam messages) is too complicated and computationally expensive. In addition to this, with an API based system we could not have retrieved all the necessary documents for our analysis. Given all these issues, in our experiments we employed the previous Twitter collections (the one in the CAW 2.0 datasets and the Microsoft Conversation one) as representative of Twitter messages. CAW and Microsoft are, in fact, enough for our general (Chapter 4) and detailed studies (Chapter 5). To conclude, the Twitter collections presented in this section are the most popular and best suit our experimental needs. In literature there are however several other collections of Twitter documents, e.g. the one employed in the RepLab[7] as part of CLEF, which interested users could investigate depending on their needs.

### 3.1.3 Collections of Traditional Documents

#### TREC Ad-hoc (Tipster)

The TREC Ad-hoc collection is the result of the TREC conference series, originated from the TIPSTER project[8]. The Ad-hoc collection contains different datasets, each with its own characteristics and its own set of topics (questions), and its own corresponding set of relevance judgments (right answers). For the purpose of our studies, we were not interested in all the document types present in the Ad-hoc collection, such as the Federal Register or the Congressional Record. We aimed at analysing only those that could contain factual to topical documents, such as the Wall Street Journal (WSJ), the Associated Press (AP) and the Financial Times Limited (FT). We judged these three collections general enough for our experiments and therefore considered only these sets of documents (WSJ, AP, FT)[9] in our studies in Chapters 4 and 5. Moreover in Chapter 5 we are complementing these datasets with the ones presented in the next section.

#### CLEF 2004

Another set of newspaper articles employed in our studies was created as part of CLEF[10], in particular as part of the CLEF AdHoc-News Test Suites, for the years 2004-2008. This

[7] `http://www.limosine-project.eu/events/replab2012` and `http://www.limosine-project.eu/events/replab2013`

[8] As from `http://trec.nist.gov/faq.html`: "TIPSTER was a DARPA-sponsored project that encouraged the advancement of state-of-the-art technologies for text handling [...], successfully concluded in 1998", while TREC still continues nowadays.

[9] Available at `http://trec.nist.gov/data/docs_eng.html`

[10] `http://www.clef-initiative.eu/`

collection is similar in purpose to the TREC one, containing both topics (queries) and relevance judgments (right answers). Although not relevant to our analysis, we should mention the multi-language nature of this collection, that contains documents as articles from newspapers of different languages (from Bulgarian to Dutch, from Italian to Persian). Due to the fact that these documents are similar to the TREC Ad-hoc ones, we did not employ them in our preliminary experiments, but we rather left them to the last part of our work (Chapter 5), where we used them as a testbed for comparing algorithms of authorship identifications. Following a recent article by Savoy [114] we restricted the collections to be used in our analysis to just two, an English and an Italian set of newspaper articles, namely the Glasgow Herald and the La Stampa[11].

## 3.2 A Novel Corpus

As stated in the previous section, there are few collections of conversational documents available in the literature. Moreover they are also limited in size and topicality. For these reasons we decided to create a novel collection of conversational documents that could serve as reference for the research community, in particular for those interested in author profiling and characterisation. Having joined the organising committee of PAN in 2011 and having organised a subtask of authorship attribution (as explained in Chapter 2), we had the possibility to develop a collection that is suitable for different tasks: the specific Sexual Predator Identification task (as part of PAN 2012), the generic task of authorship attribution (as part of PAN 2013) and the authorship identification task (Chapter 5). The reasons that led us to organise the task and to create the related collection were mainly two: the relatively uncomplicated way of finding a ground truth for the task and the interest of the research community (as in Chapter 2) and beyond[12] for such a sensible and important topic. This interest was also confirmed by the high number of participants in the task (as it is evident from the results available in Appendix B) that made the entire PAN one of the most popular and visited within CLEF in 2012.

### 3.2.1 Requirements

In creating our collection we were animated by the same spirit of TREC and, more recently, CLEF. We wanted to build a *large* collection that could serve as common reference point for researchers of different fields, from IR to NLP, from Text Mining to Machine Learning and which they could employ to compare the performances of their algorithms. Having a large collection is very important and is one of the central aspects of TREC tracks [131] and PAN laboratories [25]. It serves to fill the gap between the research and the industrial applications of the technologies developed in the laboratory.

[11] Available at: `http://catalog.elra.info` as ELRA-E0036 (or ELRA-E0038)

[12] `http://science.slashdot.org/story/12/04/03/1734208/competition-to-identify-sexual-predators-in-chat-logs`

For this reason we created a large collection of hundreds of thousands of conversations with realistic features:

- a small number of true positives (conversations containing a potential "sexual predator"),
- a large number of false positives (people talking about sex or other similar topics to those of the "sexual predator") and
- a large number of false negatives (general conversations between users on different topics).

We believed that in a realistic scenario the percentage of "predator" conversations with respect to the "regular" ones should be very low. In a different field (pedophile queries in peer-to-peer systems) the number of "predator" queries was found to be 0.25% of the total [76]. In our collection we therefore tried to respect that percentage but, in order to make the identification of the predator a feasible investigation, we increased the percentage of one order of magnitude and set this to less than 4%.

### 3.2.2 Sources

When looking for previous work containing "predatory" collections, we found a common source for all the different datasets already employed in the literature, for example in the studies of McGhee et al. [83], Kontostathis et al. [70] and Pendar [96]: the `http://www.perverted-justice.com/` (PJ) website. This is a website where logs of online conversations between convicted sexual predators and volunteers posing as underage teenagers are published. The controversial publication and the preliminary usage of these data have already been discussed in the work of Kontostathis et al. [70], where the authors also give a detailed overview of other collections tools and approaches to cybercrime and online deception detection, that are also treated in the work of Pendar [96] and Yin et al. [139]. We therefore started with the PJ data for building our collection and kept in mind the observations formulated by Pendar [96], who identified two kinds (and different subkinds) of suspicious interactions:

1. Predator/Other interaction, subdivided into:
   (a) Predator/Victim (the victim is underage);
   (b) Predator/Pseudo-Victim (the pseudo-victim is a volunteer posing as a child);
   (c) Predator/Pseudo-Victim (the pseudo-victim is a law enforcement officer posing as a child)
2. Adult/Adult (consensual relationship).

Data of type 1.a (Predator/Victim) and 1.c (Predator/Pseudo-Victim) are difficult to obtain, since it involves the police or law enforcement agencies in the process of data acquisition. To our personal experience, police and law enforcement agencies are reluctant and not very enthusiastic in collaborating on this sensitive topic, therefore we ignored this approach to data acquisition and focused on 1.b (Predator/Pseudo-Victim), which corresponds to the PJ data. PJ data constitute therefore our true positive set.

Regarding interaction of type 2. (Adult/Adult) we initially found several online sources[13] that could have come in useful but we later discarded them because they were based on a single person experience or were not of sufficiently large size (only some hundreds conversations) to be successfully employed in our collection. The documents present in the Omegle repository[14], to the contrary, served our purpose perfectely. The original service Omegle (where the documents come from) is a website that allows two strangers, connected at the same time to the website, to have an anonymous online conversation. The repository presents a random sample of more than 1 million original Omegle conversations and by admission of the provider contains *"abusive language and general silliness online"* and sometimes users *"engage in cybersex"*[15]. The quantity of conversations as well their nature and characteristics made this repository perfect to augment the level of false positives in our collection, thus to make it more challenging and somehow real.

Surprisingly, the major difficulty that we encountered was in crawling "regular" online conversations to complete the false negative set of documents and add a variety of topics of discussion, to hide a possible general topicality of our true positive conversations. We already mitigated the fact that the "predator" conversations are between two users only by introducing the conversations extracted from Omegle, so now we just needed to focus on topics about general discussions. To our surprise, the Internet lacks this kind of conversations: few people share their (private) conversations online and the massive crawling of the public channels of the major IRC networks[16] is neither trivial nor encouraged[17] [88]. Due to this resistance to make an easy access to this kind of data, the problem of retrieving IRC chat logs seemed to be at a low-level networking problem. For this reason, we decided to rely only on IRC logs that included thousands of conversations that were already available on the websites of the IRC channel managers, namely `http://www.irclog.org/` and `http://krijnhoetmer.nl/irc-logs/`. Having a large volume of conversations allowed us to increase the probability of having general discussions, interactions between just few users and a variety of messages in length and duration, despite the topical similarity among these conversations.

[13] See for example: `http://www.oocities.org/urgrl21f/`, `http://www.fugly.com/victims/` or `http://chatdump.com/`

[14] `http://omegle.inportb.com/`

[15] See: `http://inportb.com/2010/02/21/the-omeglean-society/`

[16] See: `http://irc.netsplit.de/`

[17] See: `http://wiki.vorratsdatenspeicherung.de/IRSeeK-en`

| |
|---|
| accessibility activity css developers fx html-wg html5 microformats wai-aria webapps whatwg xhtml |

*(a)* Channels in "Krijn'.

| |
|---|
| aix apache azureus blender c cisco csharp css debian fedora flood freebsd gentoo gentoo-dev gtk hardware html iptables irix java javascript linux-bg macosx mysql netbsd openbsd opensolaris oracle php python qt reactos samba solaris suse tomcat ubuntu vim windows wireless |

*(b)* Channels in "Irc-log".

*Figure 3.3.* List of topics (each word is an IRC channel) for the two most diverse sources of documents included in our collection.

As a last remark, we note that we did not employ any of the previously presented collections of Section 3.1 to generate the false negative set. We wanted, in fact, to avoid potential legal problems or attribution disputes, being the collections sometimes released with unclear licence. Moreover, another problem was the missing information on the topic or set of topics contained in the existing collections. The last issue is represented by the different formatting of the documents in the existing collections, most of the times without timestamp, which would have made the creations of homogeneous conversational treads impossible.

### 3.2.3 Challenges

Besides the problem of obtaining proper false negative examples, we needed to solve other issues due to the different nature and origin of the collections gathered. A first problem occurred when deciding about the semantic definition of *conversation*. In fact, we downloaded files from different sources of different formats, from daily logs of conversations to single transcripts of unique conversations of few lines, and we needed to put them together in a single collection. To make the conversations contained in the different files comparable, we decided to segment all the messages exchanged between the users in the threads. We decided to cut conversations into two threads when there was a break between two consecutive messages of more than 25 minutes. We empirically observed that this was a reasonable threshold for a topic change in the conversation or the starting of a new one. After this step we obtained a consistent collection of hundreds of thousands of conversations (or threads).

### 3.2.4 Properties

Once we segmented the conversations into this homogeneous and consistent set of threads, we noticed by studying their length that the vast majority of conversations

| PJ<br>perverted-justice.com | Krijn<br>krijnhoetmer.nl/irc-logs | Irclog<br>irclog.org | Omegle<br>omegle.inportb.com |
|---|---|---|---|
| ill | bugzilla | 1 | ok |
| hey | hixie | will | ur |
| yeah | html5 | want | haha |
| now | think | work | hello |
| well | archives | need | msn |
| see | I'm | well | just |
| cool | lists | html | lol |
| here | don't | lt | guy |
| know | com | org | herp |
| ya | can | know | derp |
| will | like | gt | rape |
| can | bug | www | boys |
| ur | just | com | like |
| want | w3 | dont | little |
| 2 | public | like | im |
| just | org | href | obama |
| don't | lt | im | asl |
| I'm | gt | just | hey |
| like | html | can | faggot |
| ok | http | http | apos |

*Table 3.2.* Top 20 terms in each datasets.

(from 77% to 99% depending on the source) contained less than 150 messages. We therefore decided to include in the final collection a subset of the conversations that were less than or equal to 150 message exchanges. Finally, we decided to generate a single arbitrary id for each conversation and for each user. We then replaced screen names or user names within each message with the corresponding user ids. Where possible we also replaced real email addresses with arbitrary tags, in order to avoid the potential but less probable identification of real users.

When we released the collection for the Sexual Predator Identification competition, we divided it into two parts, a training set and a testing set. Given the fact that the training set was intended to be a "practicing" set rather than a "training" set as in Machine Learning definition, where a large training set is desired (usually about 70% of the whole collection), we decided to release 30% only of the collection as practicing-training set. The rest of the collection was released as a testing set. This distinction between training and testing set is valid only for the task of Sexual Predator Identification within PAN 2012. In fact, the developed collection was not only meant to be used for the Sexual Predator Identification task but also as a testbed for other experiments on conversational documents. For this reason we employed the full collection for our experiments on authorship attribution (Chapter 5) and we report the main properties of the whole collection and the two different sets in Table 3.3. In Figure 3.3 we il-

| | PJ<br>perverted-justice.com | Krijn<br>krijnhoetmer.nl/irc-logs | Irclog<br>irclog.org | Omegle<br>omegle.inportb.com |
|---|---|---|---|---|
| number of all conversations | 11350 | 50510 | 28501 | 267261 |
| num. conv. length ≤150 | 9076 | 48569 | 21896 | 265747 |
| (% all ) | (80%) | (96%) | (77%) | (99%) |

*(a)* Overall properties.

| | PJ<br>perverted-justice.com | Krijn<br>krijnhoetmer.nl/irc-logs | Irclog<br>irclog.org | Omegle<br>omegle.inportb.com |
|---|---|---|---|---|
| Training set | | | | |
| num. conv. length ≤150 | 2723 | 14571 | 6569 | 43064 |
| num. conv. length ≤150 & with exactly 2 users | 984 | 2420 | 1146 | 41067 |
| (% training) | (36%) | (17%) | (17%) | (95%) |
| unique users | 291 | 2660 | 10613 | 84131 |
| unique predator users | 142 | - | - | - |
| Testing set | | | | |
| num. conv. length≤150 | 5321 | 33998 | 15327 | 100482 |
| num. conv. length ≤150 & with exactly 2 users | 1887 | 5648 | 2673 | 95648 |
| (% testing) | (35%) | (17%) | (17%) | (95%) |
| unique users | 440 | 4358 | 17788 | 196130 |
| unique predator users | 254 | - | - | - |

*(b)* Properties of the released collection.

*Table 3.3.* Properties of the created collection.

lustrate the different IRC channels included into two of the major components of the collections, to stress the numbers of possible topics present in our datasets and that were not present in previous work.

### 3.2.5 Acceptance

The released collection was well accepted by the community and, besides the report papers originated at CLEF 2012 (as in Appendix B), several other journal articles and papers made use of our collection [20, 35, 89, 128]. These works made use of the collection mainly for study related to cybercrime and behaviour of sexual predators online. In this context, the main positive feedbacks are:

- the large size of the collection,
- the limited numbers of predator cases,
- the variety of the conversations within the collection.

These are the aspects we most carefully considered when creating the collection.

## 3.3 Limitations

Despite the collection was well accepted within the community, it still suffered from some limitations, as clearly identified by Morris [86]. A first issue concerns the "number of distinct conversational partners": conversations from the PJ dataset always involve the same couple of users, conversations from the IRC dataset involve different users and conversations from Omegle always involve a unique couple of random users. Morris [86] suggests to address this issue by replacing some Omegle ids with predators ones, thus enlarging the number of conversational partners for predators. A second issue is represented by victims of predators to be only pseudo-victims (i.e. adults posing as underage). However, there do not seem to be easy solutions to this problem due to the difficulties in acquiring predator conversations with real victims. The last issue identified by Morris lies in the distribution of the number of conversations, that seems unbalanced towards the Omegle ones. A proposed solution is the merging of different Omegle authors into the same one, to reduce the number of conversations per user and make it comparable to the one of the other two sources, IRC logs and PJ. An additional issue was also identified by Vartapetiance and Gillam [128] and is related to the quality of the ground truth released with the collection. This was generated by one expert only, while consolidated settings suggest the number of experts to be at least of three.

Not directly related to the specific problem of Predator Identification, we can highlight another general limitation of the collection, that is the type of conversations included. We are, in fact, aware of the fact that the collection is missing representatives of regular one-to-one conversations, for example those private conversations between

users on IRC or conversations taking place on other providers of instant messaging like Skype, Hangout, iMessage, etc. However, as previously mentioned in the chapter, it is not trivial to acquire them[18] . We believe that these limitations might be addressed as indicated above in future work, in particular as updates of the released collection.

## 3.4 Summary

In this chapter we presented the first contribution of our work, the creation of a large collection of conversational documents. This is to fill the gap with existing collections, that are too small to be employed on large scale experiments. Moreover the topicality of the collection created is the widest possible as opposed to the few topics present in the existing collections. Although the developed collection was first employed with the specific task of finding predators in online conversations, it was originally created with a larger focus in mind. In fact, we will employ it in our large experiments for identifying generic authors in online conversations (Chapter 5).

In the first part of the chapter we also presented a list of relevant collections that are already present in the literature. We described all the collections in general terms, especially those that we employed in other parts of the dissertation (for example in Chapter 4 and in Chapter 5). We described those collections that we did not directly employ in our work but that are of a general interest and might be taken into consideration for future work.

[18]Some nice attempts to acquire this kind of data are done from time to time, however with limited success (e.g. `http://www.whatsup-switzerland.ch` [126]).

# Chapter 4

# A Comparative Analysis of Traditional and Short User-Generated Documents

*innovazione = creatività × esecuzione*
Alfonso Fuggetta

In the previous chapter we presented different collections of online user-generated documents, from chats to newsgroups, as well as collections of traditional documents. In this chapter, we will select a representative subset from each collection and analyse it in detail. We will conduct a qualitative analysis observing the most evident features of the subsets of both online user-generated and traditional documents (Section 4.1). We will then perform a quantitative analysis measuring different aspects of the two types of collections (Section 4.2): terms distributions (Sections 4.2.1 and 4.2.2), similarity (Section 4.2.3), burstiness (Section 4.2.4), POS (Section 4.2.5) and emoticons usage (4.2.6).

In the last part of the chapter (Section 4.3) we will highlight the important conclusions derived from the previous analysis and present possible future applications of these observations. Some of the observations will be taken into consideration in the studies of authorship identification reported in Chapter 5.

## 4.1 General Properties of Collections of Online User-generated Documents

As representative of online user-generated documents, we will use the documents from the CAW 2.0 collection, in particular those of chats (Kongregate), message exchanges (Twitter) and blog postings (Myspace and Slashdot). We will compare these collections with the traditional newspaper collection of documents, in particular WSJ, AP

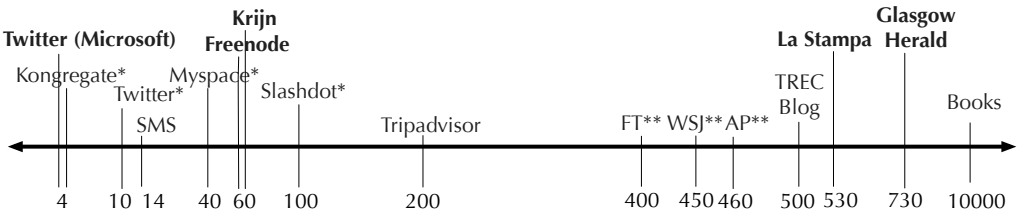



*Figure 4.1.* Average document length (in number of words) for different collections. On the left (*) the datasets of online user-generated documents analysed (CAW 2.0); on the right (**) the standard newspaper datasets (TREC ad-hoc). In **bold** are displayed collections that were presented in the previous chapter and are analysed in the next one. The graph is not on scale and values are placed in a convenient position.

(news articles) and FT (markets and finance articles) from the TREC ad-hoc collection. We note that these collections deal with similar topics than the collections of online user-generated documents used for the comparison, in particular with Myspace and Slashdot. The Myspace dataset covers the themes of campus life, news & politics and movies, while the Slashdot dataset is limited to discussions of politics. The fact that the themes are similar to the news articles is important in order to make meaningful statistical comparisons between the collections. As for the topicality of the Twitter and Kongregate datasets, due to their conversational and more unpredictable nature, we cannot state precisely what their topicality is, as highlighted by Ramage et al.[102], Haichao Dong and He [46], Tuulos and Tirri [125].

In Figure 4.1 we displayed the average document length for some popular collections of documents, including the ones employed in the current analysis: CAW 2.0 and TREC ad-hoc. It is straightforward to identify a common trait of the CAW 2.0 documents, that is the relatively shortness compared to the documents in traditional TREC ad-hoc collection. Below is a list of this and other important features of online user-generated documents. They are:

- Short: their length ranges from few to 100 words per document as opposed to traditional newspaper articles of more than 400 words per document;
- User-generated: because they are produced directly by a person using a particular online service without any review process, as instead happens to documents produced by journalists or professional writers, who to the contrary generate "edited" content;
- "Dirty": as a consequence of being user-generated and most of the time typed as fast as possible on the keyboard in emulating spoken acts, they contain spelling errors, domain specific terms or abbreviations;
- Opinionated and first person: despite being short and dirty, they contain emotions

| Collection | Size # documents | avg. doc. length # words | avg. word length # characters |
|---|---|---|---|
| Kongregate | 144,161 | 4.50 | 7.55 |
| Twitter | 977,569 | 13.90 | 7.30 |
| Myspace | 144,161 | 38.08 | 8.11 |
| Slashdot | 141,283 | 98.91 | 7.88 |
| WSJ | 173.252 | 452,00 | 7,57 |
| AP | 242.918 | 464,23 | 7,53 |
| FT | 210.158 | 401,22 | 7,26 |

*(a)* General properties.

| Collection | Vocabulary | % terms in the vocabulary that are: | | | | |
|---|---|---|---|---|---|---|
| | | stopwords | out-of-dictionary | singleton | common words | rare words |
| Kongregate | 35,208 | 44.90 | 58.94 | 56.65 | 1.39 | 84.65 |
| Twitter | 364,367 | 44.99 | 68.37 | 66.95 | 0.20 | 97.19 |
| Myspace | 187,050 | 50.67 | 69.61 | 53.30 | 0.39 | 96.10 |
| Slashdot | 123,359 | 54.00 | 57.31 | 44.82 | 0.45 | 95.88 |
| WSJ | 226,469 | 41.45 | 67.57 | 34.33 | 0.44 | 96.85 |
| AP | 242,918 | 43.70 | 75.22 | 35.77 | 0.40 | 97.34 |
| FT | 210,158 | 42.45 | 61.22 | 36.45 | 0.36 | 97.23 |

*(b)* Specific properties.

*Table 4.1.* Statistics of datasets. All values were computed before stopwords removal unless indicated.

and sentiment expressions that each user includes when communicating to others on topics he/she likes. This is different from the traditional "edited" documents (newspaper articles), that merely report news or events in an objective way.

Table Table 4.1 shows some basic statistics about these datasets, where these properties can be seen. Particularly evident is the difference in the average document length: online user-generated documents are 5 to 100 times shorter than traditional newspaper articles. In Section 4.2 we examine the implications of this property in terms of documents self-similarity (Section 4.2.3) and burstiness (Section 4.2.4), where we will also explain the role of common and rare words. On the other hand, we are not investigating the aspects of polarity or sentiment in the documents, which is a difficult problem and would have required another dissertation, like the work of Gerani [40].

## 4.2 Analysis of the Datasets

In performing the analysis of the chosen collections (see Table 4.1), we will focus on the text processing block of the IR system present in Figure 2.1 of Chapter 2. We will employ in our study two basic IR laws, Zipf's Law and Heaps' Law, and an elementary documents distance measure, the cosine similarity (introduced in Chapter 2). We will then make use of POS tagger to detect the structure of the documents in terms or their lexical categories (e.g. nouns, verbs, adjectives), as well as emotions and "shoutings".

Some preliminary information about the differences between collection of online user-generated documents and collections of standard documents can be seen when studying the statistics in Table 4.1. To generate those statistics we first indexed the documents in each collection without employing any text processing techniques (e.g without removing any stopwords or using any stemming), then we just used a standard stopwords[1] list to filter them.

For the collections of online user-generated documents we expected fewer terms to be discarded as stopwords, since we assume short documents (in particular the ones used in Kongregate or Twitter) to be written "quick and dirty", with no concern for the syntactical structure of the sentences and using a lot of abbreviations. Surprisingly the quantity of stopwords for online user-generated documents is just slightly above the quantity of standard collections, with an increase for collections representing blogs and fora (Myspace and Slashdot). A better evidence to support our hypothesis can be found when looking at the percentage of terms which occurred only once in the collection, the singleton terms. The collections of online user-generated documents contain definitely more singleton terms, which could be considered spelling mistakes or mistyped words. This is more evident when observing out-of-dictionary terms. These words are not contained in a standard dictionary and are identified as misspelled by a spelling checker algorithm. Although the percentage of out-of-dictionary terms is similar across all datasets, we noticed that for online user-generated documents this value is closer to the number of singleton words (from 2% to 16%), while for traditional TREC collections the value is different (around 33%). This fact may indicate that for online user-generated documents the presence of more singleton words could be considered an indicator of a greater number of mistyped words but also an indicator of unique link identifier, e.g. shortened through services like `https://bitly.com` or `http://tinyurl.com` that were not removed during the indexing procedure. This is not the case of the traditional TREC collections, where the presence of singleton words is less evident and can be explained by the usage of particular terms such as geographical locations, foreign words or person names which are orthographically correct but not used in the spelling checker.

[1] Different standard stopwords lists exist in literature, mostly generated taking into consideration the distribution of terms in classical books or newspaper collections. For the purpose of this chapter we made use of a standard stopwords list from one of the most widely used IR platforms: Terrier (`http://terrier.org`). For a definition of stopwords, see Section 2.1.2 in Chapter 2.

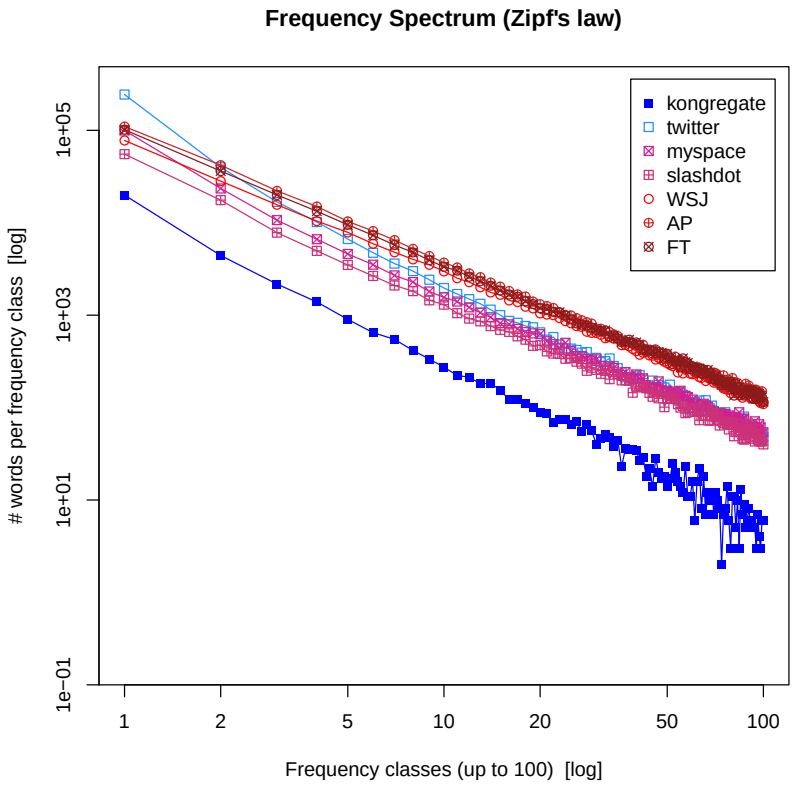



*Figure 4.2.* Zipf's law for collections of online user-generated documents and traditional documents after stopwords removal.

Different conclusions can be drawn when observing the percentage of common and rare terms. Common terms are defined as the most frequent words in the vocabulary, accounting for more than 71% of the text in each collection, while rare terms are the least frequent words in the vocabulary, accounting for just 8% of the text, as indicated by Serrano et al. [116]. Common terms contribute more for Kongregate, which is a collection of conversational documents among online gamers, in which the language of the users repeats a lot and the topical words are fewer, compared to the other collections. On the other hand, the Twitter collection presents fewer common words, a trend that might indicate that Twitter documents are somehow more topical. This means that they also contain useful information for characterising them and better retrieving them, as opposed to Kongregate documents, in which a larger part of the vocabulary contains common words.

### 4.2.1 Zipf's Law (Frequency Spectrum)

As described in the book of Baeza-Yates and Ribeiro-Neto [10] Zipf's Law is an empirical rule that describes the frequency of the text words and states informally that the frequency of any word in a collection is inversely proportional to its rank in the frequency table. In the extended formulation of Mandelbrot it is described as follows [22]:

$$\log f(w) = \log C - \alpha \ \log(r(w) - b) \tag{4.1}$$

where $f(w)$ denotes the frequency of a word $w$ in the collection and $r(w)$ is the ranking of the word (in terms of its frequency), while $C$ and $b$ are collection specific parameters. As can be seen in Figure 4.2, in a log-log scale and for large values of $r(w)$, the relationship between frequency and rank of a word can be approximated with a descending straight line of slope $-\alpha$.

Two properties of Zipf's law are particularly interesting when we study collections of documents. In fact, if we assume that the terms in the collection follow Zipf's law, we can derive the expected proportion of a term in the collection by its rank and we know that few words, generally the least informative ones, occupy a large amount of the vocabulary. The first observation is useful for scoring words, i.e. in the case of ad-hoc retrieval, while the second identifies which words can be discarded before indexing, by identifying stopwords.

If we observe Figure 4.2, for both the online user-generated and traditional documents a linear graph is observed. This is an interesting observation, that shows how the usage of terms for online user-generated documents is comparable to the traditional ones and, therefore, all the assumptions made in this context for the latter might also be valid for the former. Moreover we noticed a dependence between the length of the documents and the slope: the collections containing longer documents tend to have a larger negative slope, which may mean that the words in them are repeated more frequently, while the collections containing shorter documents are less repetitive.

### 4.2.2 Heaps' Law (Vocabulary Growth)

Heaps' law [48] is an empirical rule which describes vocabulary growth as a function of text size, as also described in [10]. Its formulation can be written as follows [27]:

$$v = k \cdot n^{\beta} \tag{4.2}$$

where $v$ is the vocabulary size of $n$ words, while $k$ and $\beta$ are collection-specific parameters.

Heaps' law states informally that the vocabulary of a collection continues to grow with the addition of novel documents, although at a different rate compared to the beginning. Figure 4.3 shows vocabulary growth with respect to the size of the whole collection. We can observe that the vocabulary of online user-generated documents

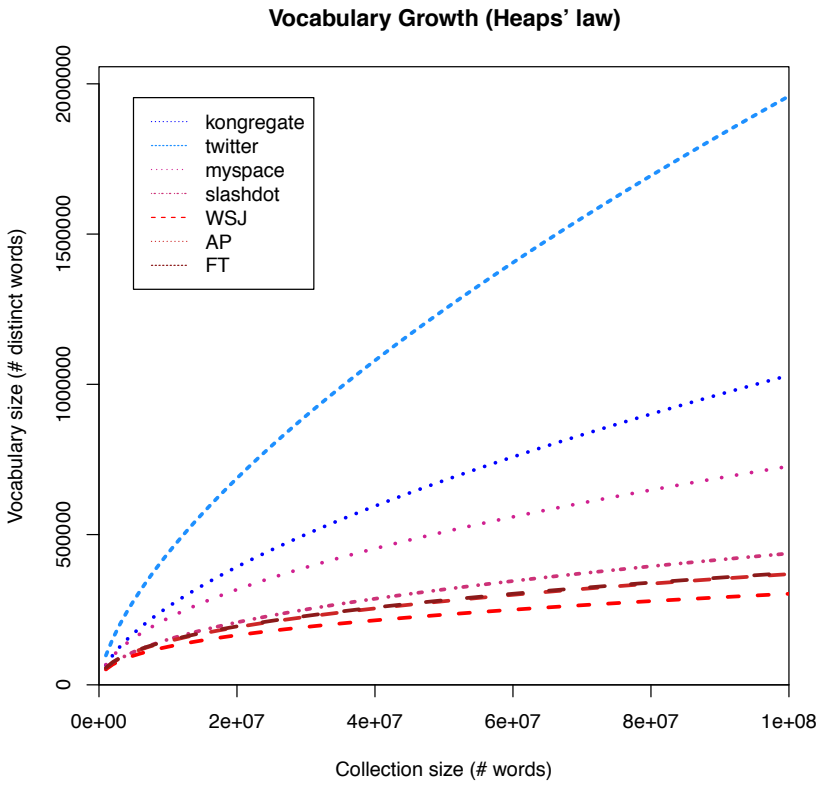



*Figure 4.3.* Heaps' law for collections of online user-generated documents and traditional documents after stopwords removal.

grows much faster in comparison with those containing standard documents. This suggests that conversations between users in Kongregate or broadcast messages of users in Twitter tend to vary greatly with the usage of ever more terms, according to the evolution of topics inside a conversation or the sentiment of a Twitter user. This may be partially explained by the high percentage of singleton, out-of-dictionary, mistyped words, abbreviations or links that are continuously introduced during the production of such documents.

We also noticed a relationship between the decreasing value of the slopes of Zipf's law and vocabulary growth. Twitter has the minimum slope in the case of Zipf's law but the maximum vocabulary growth. To the contrary WSJ has the maximum slope and the minimum vocabulary growth. This could, again, be explained by the high frequency of mistyped terms in the vocabulary of online user-generated documents in comparison to the standard ones.

### 4.2.3 Self-Similarity

Another interesting property that can be used to characterise online user-generated documents over standard collections is the self-similarity between documents. In IR there are different similarity measures that can be used to compute the distance between two texts. A traditional one, that is considered to work better than others, is the cosine correlation similarity. This similarity measure is generally applied to measure the distance between a text in a document ($D$) and a query ($Q$) and is called cosine. The name is due to the measure used to calculate the similarity between the text and the query, which are represented as vector of terms. In determining the distance, the inner product between the two vectors is computed and the result of the product is between 1, if the two vectors are identical, and 0, if they are completely disjoint (no terms in common). The formula of the cosine similarity is expressed by the following equation [27]:

$$cosine(D_i, Q) = \frac{\sum_{j=1}^{t} d_{ij} \cdot q_j}{\sqrt{\sum_{j=1}^{t} d_{ij}^2 \cdot \sum_{j=1}^{t} q_j^2}} \tag{4.3}$$

where $D_i$ is a particular document in the collection with terms $(d_1, d_2, ..., d_t)$ and $Q$ is a query with terms $(q_1, q_2, ..., q_t)$. In our case, we want to compute the similarity between two documents, therefore we should substitute $Q$ with another document in the collection, i.e $D_k$ and $q_j$ become $d_{kj}$. At the numerator, for each matching term $t$ the inner product is computed employing the score associated with each term, while at the denominator a normalisation depending on the length of the two vectors is performed.

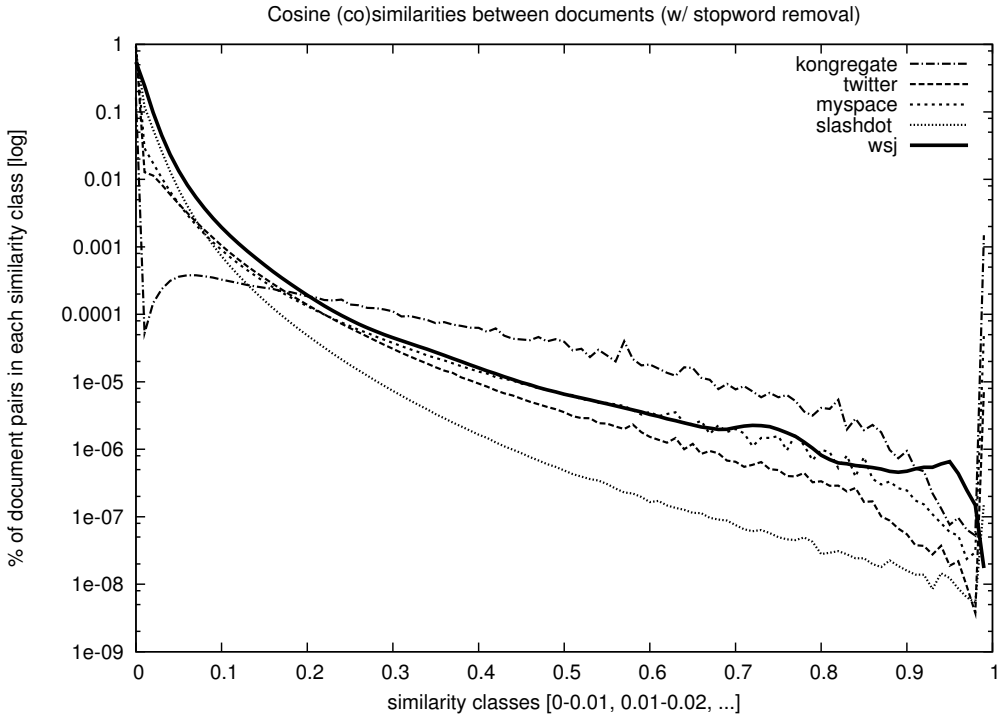
Traditionally the weight associated with each term for the cosine similarity is computed employing the *tf-idf* weighting. The *tf* component considers the relative frequency of a term in the document, while *idf* reflects the importance of the term in the collection.

$$tf_{ik} = \frac{f_{ij}}{\sum_{k=1}^{t} t_{ik}} \tag{4.4}$$

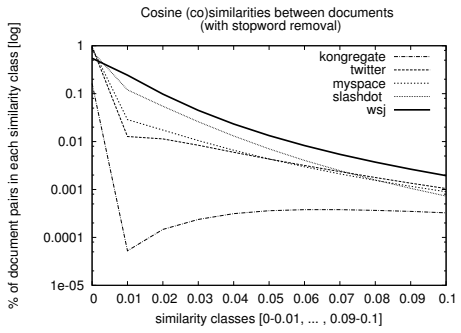
$$idf_k = \log \frac{N}{df_k} \tag{4.5}$$

In Equation 4.4 the formula for the *tf* is displayed, where $f_{ij}$ is the number of occurrences of a term $j$ in a particular document $i$, normalised by the length of the document. In Equation 4.5 the formula for the *idf* is illustrated, where $N$ is the total number of documents in the collection and $df_i$ is the number of documents in which the term $k$ occurs. The final score is obtained by multiplying the two components $tf_{ik} \cdot idf_k$, hence the name *tf-idf* score.

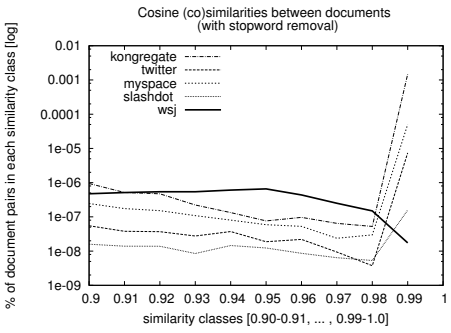
In this experiment, we computed the similarities for all the documents in each collection to study how these documents are similar to each other. The computation of the cosine similarity employing *tf-idf* weighting was done after removing the stopwords from the documents. We decided to display only WSJ as representative of traditional collections, having observed a similar behaviour also for the other collections.




*(a)* Similarity classes




*(b)* First similarity classes




*(c)* Last similarity classes

*Figure 4.4.* Self-similarity between documents after stopwords removal. We normalized the count for document in each similarity class by the total number of comparisons.

| | Kongregate | Twitter | Myspace | Slashdot | WSJ |
|---|---|---|---|---|---|
| Kongregate | 1 | 0.9941 | 0.9969 | 0.9131 | 0.8709 |
| Twitter | 0.9941 | 1 | 0.9988 | 0.9097 | 0.8778 |
| Myspace | 0.9969 | 0.9988 | 1 | 0.9246 | 0.8913 |
| Slashdot | 0.9131 | 0.9097 | 0.9246 | 1 | 0.9914 |
| WSJ | 0.8709 | 0.8778 | 0.8913 | 0.9914 | 1 |

*Table 4.2.* Pearson's product-moment correlation $\rho$ between similarities of different collections ($p < 0.05$).

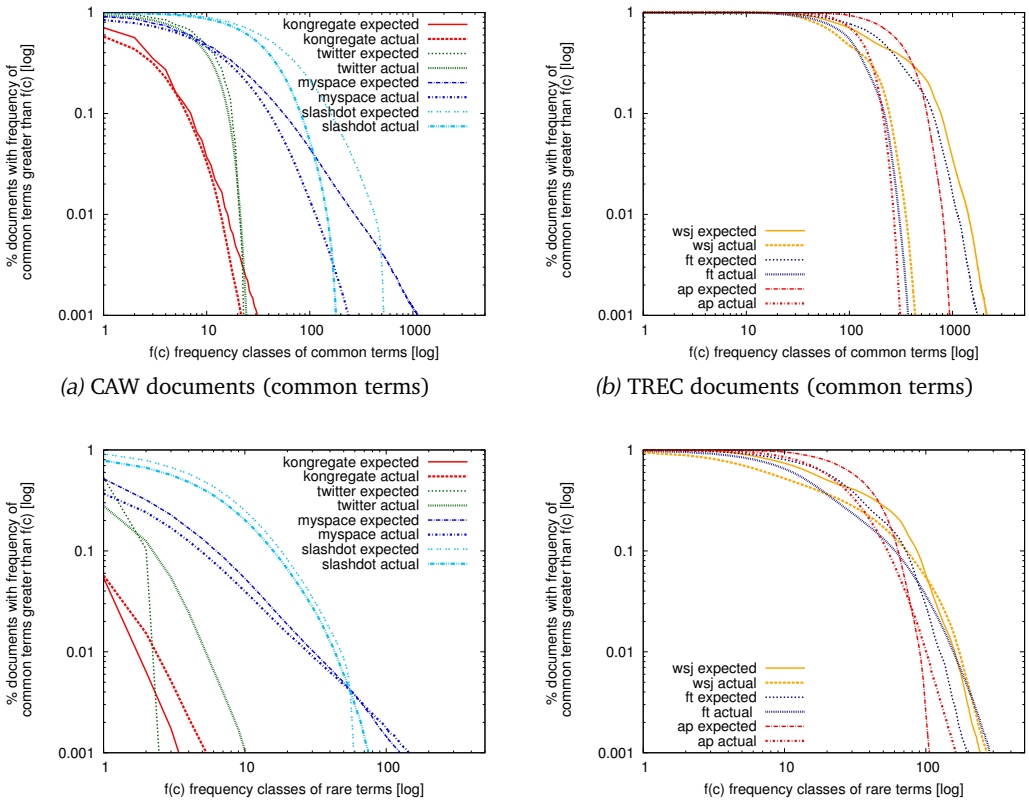
In Figure 4.4 we plot the frequency of each similarity class (from 0 to 1), interpolated by lines for visual purposes. A first observation of the general picture (Figure 4.4a) already allows to identify the most evident differences between online user-generated documents (Kongregate, Twitter, Myspace and Slashdot) and traditional ones (represented by the WSJ) at the extremes of the similarity graph. For this reason we also zoom in to show only the percentage of document pairs with the lowest (Figure 4.4b) and highest (Figure 4.4c) similarity scores. The rest of the graph shows a similar trend for all the collections, although with a lot of variations also among online user-generated documents.

In the first similarity class we observe that online user-generated documents appear less frequently with lower similarity values (0.01-0.09), as they become shorter (from Kongregate to Slashdot). To the contrary, they appear more frequently with higher similarity values (0.9-1.00), in contrast with the behaviour of the documents contained in traditional collections. This latter, in fact, drops down when we consider only the last similarity range (0.99-1.00).

This means that online user-generated documents seem to be more similar among themselves (Kongregate, Twitter, Myspace) than to longer ones (Slashdot, WSJ), as displayed in the correlation Table 4.2. This can be explained with the length of the documents itself: short documents contain fewer words (less "information"). Therefore, given two short documents, there is a higher probability that they appear to be similar even if they are unrelated, just because they are short.

### 4.2.4 Burstiness

In this section we perform another analysis on our collections, where we study the burstiness property of the words. There is not a unique and formal definition of burstiness in literature, but it is generally considered the property of a term to recur more often in a document or in a part of a text where it is already mentioned rather than in other arbitrary places of the same document or text. For this reason we can consider that term to characterize that particular document or part of text. In Figure 4.5 we display this property for a particular set of terms, common and rare, as defined at




*(a)* CAW documents (common terms)

*(b)* TREC documents (common terms)

*(c)* CAW documents (rare terms)

*(d)* TREC documents (rare terms)

*Figure 4.5.* Common and rare term burstiness for user-generated documents (CAW) and traditional ones (TREC).

the beginning of Section 4.2. The plots display the percentage of documents in each collection that contains a certain number of common or rare words.

In each plot we show also the expected number of such documents if the words in the vocabulary were uniformly distributed (according to their overall frequency in the collection) across the documents in the collection. Differences between the curves for actual and expected number of documents indicate that terms in the different classes manifest the burstiness property.

Looking at the common terms plot for the three traditional collections (AP, FT and WSJ), we noticed that the line denoting the actual number of documents with a certain number of common terms in them lies well below the expected number of such documents. This indicates that documents are bursty, since common terms are not spread evenly across the collection of documents, but are concentrated in some documents more than in others. The same is true (although to a less extent) for the rare terms in these collections: the actual number of documents containing a certain number of rare

words lies below the expected curve, again indicating that documents are bursty, since the rare words are not uniformly distributed across documents.

Comparing the plots for online user-generated documents (Kongregate, Twitter, Myspace and Slashdot) with those for traditional collections, we observed that the difference between the expected and the actual number of documents is far less pronounced (especially for the common terms) than it is for the traditional ones. This indicates that burstiness may not be an important issue for online user-generated documents as it is for traditional documents.

The fact that the expected/actual curves for the different collections of online user-generated documents differ greatly from one another, positioning in different parts of the plot, is due to the large difference in average document length in the different collections. The display of these curves, in fact, follows the same order as the average length of documents in each collection. The curves for the traditional collections, instead, line up quite well due to the fact that the average document length is very similar.

### 4.2.5 Part-Of-Speech Distribution

In this section we will analyse the grammatical properties of the terms in each collection, i.e. looking at the number of nouns, adjectives, verbs, etc present in each document. In order to do this, we will employ a posting list where each word is assigned to one grammatical category: the framework GATE[2] and its component ANNIE[3] dedicated to the POS analysis, introduced by Wilcock [136] and Cunningham et al. [28].

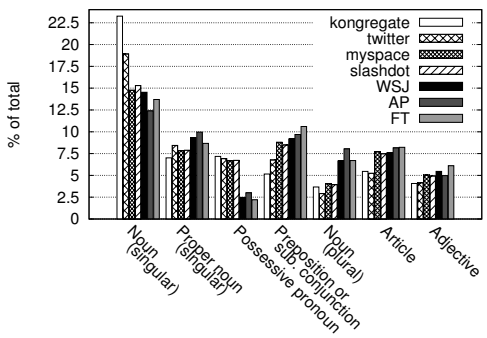
In Figure 4.6 we reported the results of the POS analysis of the full text on 30% of the documents in the collection, selected at random (since we did not find significant variations in the distributions with a higher subset). We used the ANNIE default settings, which include a posting list based on newspaper articles, and reported only the most significant categories[4] in Figure 4.6.

If we observe the results of Figure 4.6 in detail we will notice two different behaviours: first, some inter-collection variations, between the collections of online user-generated documents and the traditional collections, then an intra-collection variation within the collections of online user-generated documents, between chat-style and discussion-style documents. Inter-collection differences can be seen in the usage of proper nouns, possessive pronouns and plural nouns in Figure 4.6a, as well as in the usage of verbs and adverbs in Figure 4.6b. An explanation for this may be found in the nature of the documents contained in each collection: in online user-generated documents the users producing the texts are willing to express their point of view or emotions against the others (high usage of possessive pronouns), qualifying the amount of their sensations (high usage of adverbs), addressing directly in the first person (high

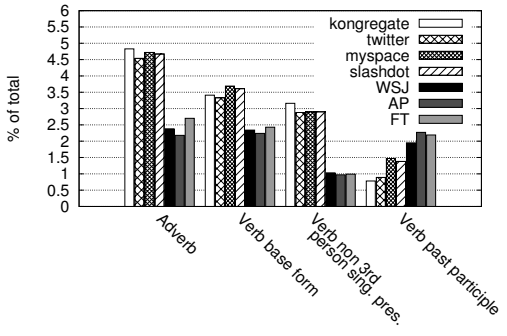
[2]GATE: "General Architecture for Text Engineering", `http://gate.ac.uk/`

[3]ANNIE: "A Nearly-New Information Extraction System", `http://gate.ac.uk/`

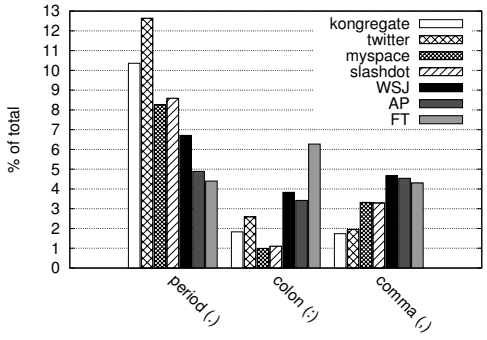
[4]A complete list of the POS tag extracted by ANNIE can be found on `http://tinyurl.com/gate-pos`



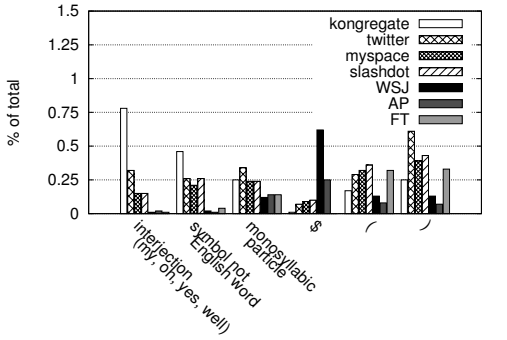

*(a)* nouns, pronouns, articles and adjectives




*(b)* verbs and adverbs




*(c)* high frequency punctuation




*(d)* interjections, symbols and low frequency punctuation

*Figure 4.6.* POS analysis.

usage of verbs not in the third person singular) and referring to actions occurring mostly in the present time (verb in base form). To the contrary, texts that are contained in traditional collections are edited in a professional way and report events occurred in the past (high usage of verbs in past participle), not occurring to the author itself (high usage of the third person in the verb) or taking place in a particular location (higher use of singular proper nouns). Moreover, if we observe the usage of punctuation, interjections and symbols in Figure 4.6c and Figure 4.6d, we will notice how online user-generated documents consist of a more direct, personal and simple communication, given by a more extensive usage of interjections, symbols, monosyllabic particles and periods. Documents in the traditional collections, instead, are more descriptive, due to the usage of colons and commas, which generally link together different concepts inside the same sentence.

Intra-collection differences, on the other hand, can be noticed within the collections

of online user-generated documents, where some datasets (Myspace and Slashdot) appear to be more related to the traditional collections than the others (Kongregate, Twitter), which highlight different properties. These features are a higher usage of proper singular nouns, periods, interjections and symbols, and a less common usage of articles and adjectives, which becomes the least among all the collections for verbs in the past form and commas. These can be seen as attributes of an essential and immediate communication, such as the online-chat (Kongregate) or microblog (Twitter). Despite that, for some POS categories the Myspace and Slashdot datasets are similar to or just in-between with the traditional TREC datasets: this appears for prepositions and subordinative conjunctions, adjectives (Figure 4.6a), verbs in the past partiple form (Figure 4.6b) as well as for periods, commas (Figure 4.6c) and interjections (Figure 4.6d). We therefore label these collections (Myspace and Slashdot) as containing discussion-style documents, a concept introduced by Yin et al. [139], as opposed to the conversational ones (Kongregate, Twitter).

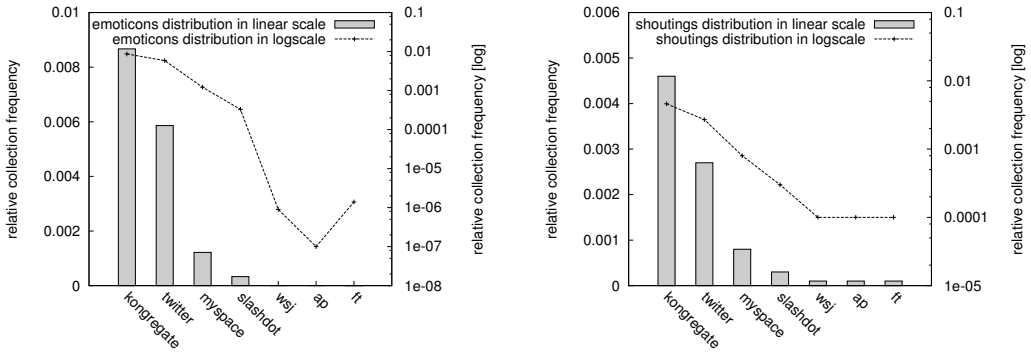
### 4.2.6 Emoticons and "Shoutings" Distribution

In this section we will complement the POS analysis of Section 4.2.5 by investigating the distribution of emoticons and "shoutings" among the different collections. These features, in fact, can be discriminative for identifying user-generated content as illustrated in [11] and in particular conversational data, as pointed out in[46].

We collected a list of the most common emoticons (mostly through Wikipedia) and parsed each document by comparing each token separately with a regular expression, thus identifying and counting only whitespace separated emoticons (such as **:)** and **:P**)[5]. Similarly, we then counted the counted so-called "shoutings", which we define as whitespace separated tokens containing a succession of three-or-more consecutive instances of the same letter (e.g. `zzzz` and `mmmmaybe`). We did not include in this count tokens containing internet addresses (`www` and `WWW`) since they do not provide additional information on the collections being analysed.

In Figure 4.8 we report the distribution of emoticons and shoutings for all the collections. The values represented are the relative collection frequency in both the linear and log scale. The behaviour of the distributions is similar and reflects the nature of the collections. The collections containing user-generated documents (Kongregate, Twitter, Myspace, Slashdot) present a large number of colloquial and informal tokens, such as emoticons and shoutings, that are used to improve the expressiveness of the communication. In the standard collections containing professional edited documents (WSJ, AP, FT), instead, communication remains on a formal and neutral level (showing these collections almost zero counts for emoticons and shoutings and at least 1 order of magnitude less than the others).

[5]We experimented also with matching emoticons within sequences of characters like `hello:)mum` but obtained too many false positives to consider those results valid. For the same reason, we did not count emoticons containing whitespaces such as `: )`.




*(a)* Emoticons distribution

*(b)* Shoutings distribution

*Figure 4.7.* Collection relative distributions for emoticons and shoutings.

As for the POS features analysed in Section 4.2.5, besides the inter-class differences between collections of online user-generated documents and traditional collections, some intra-class differences among the collections of online user-generated documents can be observed: the shorter and more colloquial documents (Kongregate and Twitter) contain more emoticons and shoutings occurrences (on the order of 1 or 2 orders of magnitude) than the documents that are more of a discussion-style (Myspace and Slashdot).

## 4.3 Novel Challenges in Information Retrieval for Social Media

In the previous sections we presented a series of analyses to characterise collections of documents that are new to IR and compared them to traditional collections of newspaper articles or webpages. The collections of online user-generated documents contained documents of different kinds, from online conversations to microblogs, from blogs to fora, and are good representations of the so-called social media. The main question that arises after this analysis is: *what is the real applicability of these results, in particular to the field of IR?*, which is our field of reference. We will try to draw here below some general observations.

- Online user-generated documents are dirty, containing a large and growing number of typos, spelling mistakes, grammatical errors and abbreviations. On the other hand, traditional techniques of IR assume the text in input to the IR system to be clean and consistent. For these reasons it is important to be able to identify such "dirty" components in the texts and be able to "normalise" them. This "normalisation" process can be done at parser level, within the 'text processing' component of the IR system in different ways. One possibility would be the

| | Kongregate | | Twitter | | Myspace | | Slashdot | |
|---|---|---|---|---|---|---|---|---|
| | emoticon | % | emoticon | % | emoticon | % | emoticon | % |
| 1 | :P | 16.89 | :) | 43.35 | :) | 33.13 | :) | 37.26 |
| 2 | XD | 13.09 | ;) | 11.12 | ;) | 12.84 | ;) | 17.75 |
| 3 | :) | 12.72 | :-) | 10.22 | :P | 10.84 | :-) | 14.92 |
| 4 | :D | 10.92 | :D | 8.78 | :D | 8.93 | ;-) | 10.56 |
| 5 | -.- | 5.11 | ;-) | 5.31 | :] | 4.61 | :P | 5.42 |
| 6 | xD | 4.62 | :P | 5.15 | XD | 3.47 | :D | 2.94 |
| 7 | :0 | 3.45 | *:-(* | 1.82 | :p | 2.84 | B) | 1.94 |
| 8 | =D | 2.95 | XD | 1.42 | =P | 2.39 | *:-(* | 1.36 |
| 9 | :p | 2.84 | :p | 1.36 | xD | 2.37 | :p | 1.19 |
| 10 | =P | 2.72 | :-D | 1.10 | :-) | 1.61 | :-P | 1.04 |

*Table 4.3.* Top 10 emoticons in each dataset with their relative frequency as a percentage of all emoticon occurrences. We omitted the few counts for WSJ, AP, FT since they are not informative. Emoticons in *italic* express a negative feeling (sadness), all the others a positive one (happiness, astonishment, smartness, tongue, smiley,...).

correction of the spelling mistakes with the help of a spelling checker. Another possibility might be the substitution of the abbreviations with their longer meaning, according to a posting list of most used expressions. In one last example, one might also decide to preserve those terms and consider them neologisms to be used to characterise a certain portion of a text, a specific document or a particular user associated with them. Emoticons, in particular, might be preserved and later used to detect the sentiment associated to the text they were attached to. We can conclude mentioning a useful tool that was designed by Owoputi et al. [91] specifically to parse and extract POS for Twitter and that is able to parse and recognise, among others, emoticons, abbreviations and urls.

- Online user-generated documents are *short or very short*, in fact so short that they might be considered similar to each other even if they are part of a different context (see Section 4.2.3). This fact is interesting because it means that if we want to analyse online user-generated documents from the novel collections, we cannot simply use the same techniques of traditional IR as if they were standard documents. In this latter case, in fact, if we process a single document with traditional IR techniques, it is often long enough to extract meaningful information from it. In the first case, instead, a single document is generally too short to be able to provide enough information if it is processed with traditional IR techniques. For this reason, these short documents are often aggregated in a single longer document, to be later processed more easily. There are different strategies for combining short documents into longer and more complex ones. One of

```
(Z.Z)  (-.-)Zzz  Zzz :) :-) :-]  :]  :] :> :-> => ^_^  ^-^  (^_^)  ^.^ ;)
;-) ;]  ;-] ;>  ;-> (^_~)  ^_~  ^_* :wink:  :(  :-(  :[  :-[ =(  =[  :< =<
:D  :-D  :D  :ºD =D Xº-D  XºD XD  xD  BD  8D X3  x3 :P  :-P  :-p  :p  =P
=p  :|  :-|  8) 8-)  B)  B-) :'(  :'-( :'[  :'-[  ='(  ='[  :'<  :'-< ='<
T_T  T.T (T_T)  Y_Y  Y.Y  (Y_Y) ç_ç  ç.ç (ç_ç)  ;-;  ;_;  ;.;  :_:  .__.
:S  :-S  =S @_@  :-?  :?  ?_?  ?.?  :\")  :-\")  :/) :-/)  :-0)  :0  :-o) :o
:-O)  :O  =O  =O  =o  =_=  -.-  -.-\  -.-º  o_o  o.o  o.0 0.o  O_o  o_O  º-º
ºº (º-º)  º_º  (º_º)  (oO)  (Oo) ò.ò  o_0  0_o  0.o  o.0 0o  o0 >_<  è.é
è_é  U_U  u_u  (U_U) <3 ()(º_º)() U.U  u.u  (U.U)  U.u  V_V  v_v  (V_V) V.V
v.v  (V.V)  <_<  >_> *_*  (*_*)  (*__*)  (*.*)  *.*  :-*  :*  :-x  :x  :-X
:X  ^*^  \\o/  §_§  (§_§)  x_x  X_X  (X_X)  x.x  X.X  (X.X) #_# (#_#) OwO
(OwO)  (*w*)  *w*  :-Q_  :-Q__ :-Q___  :Q_  :Q__  :Q___  : Q_    (:  (-:
: Q__  : Q___  :-Q :Q  =-Q_  =-Q__  =-Q___  =Q_  =Q__  =Q___  = Q_  = Q__
Y_Y  i.i  i_i = Q___ =-Q  =Q  *ç*  =3  :-3  :3  x3  :-@  :@  >:-@  >:@  >:
>_<  è_é  @>-- @:)  @:-)  @:-]  @:]  @:>  @:->  @=>  @=)  @=]  \\(^_^)/
 *<:o)  ^o^ T.T  Y.Y  T_T
```

*Figure 4.8.* List of Emoticons Used. This is only a partial list; for more complete lists of emoticons and their meaning we suggest to consult online resources like `http://en.wikipedia.org/wiki/List_of_emoticons`.

the simplest methods is concatenating documents according to some proximity. Temporal proximity involves merging documents that were created close-in-time. Semantical proximity involves joining texts with the same approximate content. Proximity based on the authorship implies concatenating documents produced by the same author. These operations, however, are not obvious and they are application dependent. For example, it is relatively easy to concatenate documents of the same author if we want to profile and retrieve these documents based on their authors, like it is done for documents in traditional collections. It is, nevertheless, more difficult to decide which documents to concatenate if we want only those related to a certain topic, as characterising the content is not an easy task. An approach based on simultaneous combination of different strategies might be of help in this case. In the next chapter (Chapter 5) we will propose a method for advancing the concatenation of documents based on author proximity.

- Online user-generated documents are are less bursty, thus *less topically defined*, compared to the ones contained in traditional collections. As briefly presented above, a possibility would be to merge them according to some criteria. The idea is that from a longer text, the topical components would emerge more easily and more sharply than in a shorter text. However, sometimes this is not enough, therefore we should find other ways to better characterise them. One possibility is to expand their semantic content, starting from the few topical words contained

in them and deriving from those additional text fragments from other sources. Having just a few words to start with, the easiest thing to do would be to look up in a standard dictionary and expand the single terms with their descriptions. This is however quite simplistic: a more refined possibility would be to substitute their dictionary entry with the Wikipedia definition, to have a richer and more diverse set of additional terms. Furthermore, one could employ conceptual-semantic networks like Wordnet to navigate along related terms and find new concepts to be inserted in the original text or expanded iteratively. Moreover, since online user-generated documents often contain links to webpages, another possibility would be to concatenate the text from these webpages to the original document, always to obtain a richer description of the underlined topic.

- Standard normalising techniques or scoring measures of IR (like the *tf-idf*) rely on the simple textual content of the documents. These methods can also be applied to online user-generated documents only if some proper preprocessing is employed. The preprocessing techniques include all the steps indicated above, from errors correction, to documents merging and expansions. However, since online user-generated documents are also more expressive than the ones in traditional collections, it might be interesting and effective to combine standard IR scoring with scoring based on other different indicators based on language analysis. We already mentioned emoticons as a way of complementing the standard textual information in a document. Emoticons indicate a particular emotion associated with a text fragment and these can indeed be used to better characterise the text. Besides emotions, other indicators that can be derived and combined with standard textual information are polarity (if the text contains or not opinions), sentiment (if the opinion attached to the text is positive, negative or neutral) as in the work of Gerani [40] or other figurative expressions (like humour or irony), for example, as highlighted by Reyes and Rosso [106].

## 4.4 Limitations

The study conducted in this chapter is bounded to some specific collections, that might be extended to others (e.g. online reviews like Tripadvisor) not analysed herein. Moreover the instruments employed in the analyis are just a portion of the different tools one could utilize, for example other measures of similarity could have been applied or other specific linguistic features, like abbreviations. It is also to be said that we only partially derived the associated laws to the functions displayed in the different parts of the analysis, therefore no statistical observations were made on these functions. This could be done in future work.

## 4.5 Summary

In this chapter we introduced the novel challenges derived by the introduction in the literature of collections of online user-generated documents. The focus was on the analysis of the properties of these and of traditional collections, to compare them both qualitatively (Section 4.2) and quantitatively (Section from 4.2.2 to 4.2.6). Different metrics were employed to compare the collections, including Zipf's and Heaps' laws, cosine similarity, burstiness and both a generic POS analysis and a specialised one to detect emoticons and shoutings. We selected four particular collections as representative of online user-generated documents: conversational documents, microblog documents and documents from fora and blogs. We made use of two standard collections of newspaper articles as representative of traditional documents.

From the studies conducted we observed different properties of online user-generated documents with respect to the ones belonging to the standard collections. In the last part of the chapter (Section 4.3) we highlighted these properties and illustrated techniques for making these online user-generated documents suitable for standard IR systems. In particular we presented methods for dealing with spelling mistakes and emoticons, as part of an unconventional language present in online user-generated documents, as well as for treating short and casual documents in there contained. We concluded with a list of possible indicators to be combined with standard IR metrics to improve the characterisations of documents from social media. In the next chapter we are making use of some of these observations to improve the accuracy of authors identification in conversational documents.

# Chapter 5

# Authorship Identification

*'Quickly' in my mind would be years.*
Jeffrey Bezos

In this chapter we will focus on the problem of authorship identification for conversational documents. As identified in the previous chapter, conversational documents have specific properties that require them to be treated differently with respect to standard documents. In the first part of this chapter we will give an overview of the related work on authorship identification (Section 5.1) and present the state-of-the-art classifiers for authorship identification (Section 5.2). Later we will introduce our proposed approach to improve the state-of-the-art methods and make them suitable also for conversational documents (Section 5.3). After having introduced the experimental settings (Section 5.4), the rest of the chapter (Section 5.5) is dedicated to the experimental validation of our proposed approaches.

## 5.1 Related work

In Section 2.4 of Chapter 2 we already introduced the general problem of authorship attribution and its declination into different subproblems: authorship characterisation, similarity detection and authorship identification. In order to better understand the contribution of this chapter, in this section we will focus on the specific task of authorship identification and provide an overview of the important related work.

Good introductory works on the topic of authorship identification are the book of Juola [64] and the article of Stamatatos [121], in which the two authors highlight the main techniques and applications of authorship identification. These techniques generally apply one of the two different approaches to the classification problem, namely generative (e.g. Bayesian) models and discriminative (e.g. Support Vector Machine - SVM) models [64]. In combination with these classification approaches, different features can be used to characterise the authors: from lexical and character features, to

syntactic and semantic ones [121]. It is important to point out the two basic components of a good method for authorship identification (but it might be extended to any classification problem): the feature selection task and the classification approach. In improving an algorithm for that problem, it is often not necessary to improve both components; one is generally enough. In fact, classification methods are usually well consolidated and might be simply used as-is (as it happens for example in most cases for SVM), while the major room for improvement is generally in the feature selection component. We are following this strategy in our work, similarly to previous and relevant other works presented below.

As mentioned before, online conversations and social media are two means of communication for which little research has been done in the context of authorship identification. Apart from the specific task of predator identification (Chapter 2), there is little research explicitly addressing the problem of authorship attribution [78] or profiling (stylometry) in online conversations [74, 75].
The first paper [78] is important because, to the best of our knowledge, it is the first paper addressing the problem of authorship identification for conversational documents. Moreover, it remarks the importance of features selection in the process of classification. The authors, in fact, focus on an algorithm that makes use of *n*-grams as features and the minimisation of a customised version of the shared root distance as a classifier. While the second component was used as a black-box, the improvement was performed in the feature selection part of the algorithm. The authors, in fact, tried to build author profiles based on the inverse frequency of features, a concept similar to inverse document frequency (illustrated in Chapter 4, Section 4.2.3) but that makes use of character features instead of words, and authors profile instead of documents. Despite the good performance of this approach and it robustness to spelling errors or mistakes typical of conversational documents, we believe it has two main weaknesses. The first weakness is due to the usage of characters features: in many applications, for example security or forensics analysis, it is of primary importance to be able to associate each author with his vocabulary (terms) or topics of interests [8, 7]. This is clearly not possible when using character features. The second weak point in this work is the limited scope of the study: the authors analysed only a small set of authors (50) and a small set of documents (2476, about 50 per authors) generated in a single IRC channel. We already argued against small collections of a limited number of authors and topics (or channels) in Chapter 3 (Section 3.2) and we show later in the chapter how good results for a limited number of authors change when the set of authors and documents is much larger. The second set of papers mentioned [74, 75] is important because it is the first to contain studies on authorship attribution for conversational documents. Despite that, these studies focus on the authorship characterisation problem, that is not central to our work (only marginally, see Appendix A). For this reason we only take note of the features used in these works, that are term-based or style-based. Despite being interesting and of a reasonable size (250k messages and 2500 users), the collection of

documents are unfortunately in a language we could not deal with (turkish).

Other studies on conversational documents are focused on other domains rather than authorship identification. Two examples of these studies are the works on chat disentanglement [133] or segmentation [33, 32], that we already analysed in detail in Chapter 3, when presenting the collections of conversational documents in the literature (Section 3.1.2). On the other hand, if we move our point of view from conversational documents to the novel ones generated online, only few publications exist on the topic of authorship identification. In this context, some studies have already been conducted covering newsgroups [142], blogs [72], microblogs [77, 115, 85, 19, 120] and social media like Netlog[1] [94]. What is to be observed from these works is the common use of a classification strategy based on discriminative methods. In fact, discriminative approaches like SVM have been successfully employed in the great majority of these works [115, 120, 94, 72, 142], while the focus was on the feature selection step of the classification to obtain improvements in the respective algorithms. The only work [77] that did not employ SVM as classifier made use of methods based on customised measures of distance, similarly to [78]. The features in this work are based on *n*-gram characters. There are at this point different clarifications to be done. We already showed in Chapters 3 and 4 how online conversations differ significantly from other social media like blogs, newsgroups or discussion fora (e.g. the length of the messages [57], their style [59]) and for this reason the problem of authorship identification for online conversations should be approached in a different way. Discriminative models like SVM, that seem to work well for other online documents, have two main drawbacks: they estimate a model fitting some example data and "hide" the features needed for the classification into the model. We already highlighted the importance of being able to determine which feature is contributing to which part of the classification [8, 56] and we underline here the unsuitability of SVM or character-based approaches to our problem. Moreover, SVM requires a phase of training based on dedicated data and another phase of classification based on the model derived from the training sample. This limits the possibility of adapting existing models to new sets of data (e.g. new authors) without re-training the models. Even worse, it might be impossible to employ the models at all due to lack of training examples. Generative models, instead, seem to be more flexible and performant in the case of conversational content [74, 75]. This leads us to a recent work [114] where the author explores, in detail, the most common statistical methods for authorship identification, showing their suitability in comparison to other standard generative models (e.g. Naïve Bayes). These methods are as powerful and flexible as the generative models, but benefit further by allowing control over the contribution of each feature (term) in each document (author).

Given all the above reasons, we decided to make use of two of the approaches described in [114] in our work. These are statistical approaches that might scale and perform well even with thousands of authors. We applied these existing statistical

[1] www.netlog.com

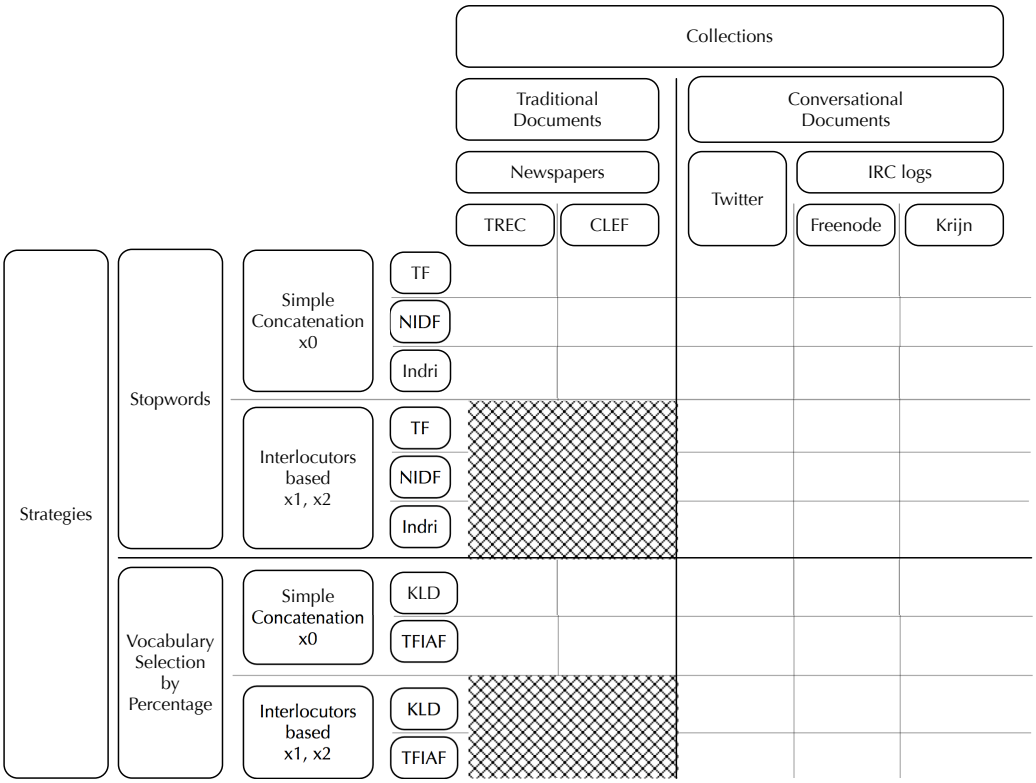



*Figure 5.1.* Experimental plan. Rows contain the name of the collections employed in our experiments while columns report the main experimental settings.

methods with the ones we developed on the datasets we created (Chapter 3) to a large number of authors and documents. This is something unusual in the literature, since experiments are generally conducted on a smaller set of authors (in the order of few dozens) [114, 78]. Despite that, we already mentioned some real applications where the monitoring of a large number of subjects at the same time is desired. Our method, in fact, is designed to return improved results with a large number of conversations and interlocutors per user[2].

In Figure 5.1 we illustrate the experiments conducted. The three statistical methods derived from [114] are applied to each collection of *traditional* and *conversational* documents (columns of the table). They are the classifiers employed in our system: $\chi^2$, Kullback-Leibler Divergence (KLD) and Delta (Section 5.2). For each feature selection strategy: *stopwords* and *vocabulary selection* by percentage we applied the 3 different feature selection methods: *simple concatenation of documents per author* (method x0, presented in Section 5.3.2) and the *interlocutors-based* methods we devel-

[2]We employ the term *user* and *author* interchangeably, since authors of conversational documents are users of the systems or services where these documents are produced.

oped (methods x1 and x2, presented in Section 5.3.3) . These are the rows of the table. For the stopwords strategy we adopted 3 ways of selecting the terms to be considered stopwords: using *Term Frequency* (TF), the *Normalised Inverse Document Frequency* (INDF) and a standard list of stopwords (Indri). These approaches are presented in Section 5.3.1. For the strategy based on the vocabulary selection by percentage we used two methods for ordering the terms: Term Frequency - Inverse Document Frequency (TF-IDF) and KLD used point-wise. These two approaches are extensively presented in Section 5.3.3. The "obscured" area in the table represents the non-applicability of the interlocutors-based approach on the traditional collections of newspapers, because no dialogs are available between authors of the text in these collections. An additional setting is not represented in the figure but later explained in detail (Section 5.4): For each collection we tested all the strategies for a group of 20-randomly selected users and for another large group of hundreds of users. This latter is a challenging situation often avoided for its complexity in the previous literature. However it is of great interest as it represents a realistic scenario. One last remark on the collections employed: we made use of sets of newspaper documents usually employed in the IR or NLP literature (collections from TREC and CLEF workshops) and compared them with sets of conversational documents, Twitter and IRC logs. Twitter and IRC logs are used as representatives of conversational documents in the light of the results of the study conducted in Chapter 4. The presentation of all the datasets characteristics and complete experimental settings can be found in Section 5.4.2. In the next Section, instead, we will start the presentation of our framework by introducing the classifiers employed.

## 5.2 Classifiers

In this section we are presenting the three statistical methods we are using as classifiers for our experiments. Each method computes the "distance" between two distributions of terms, representing author profiles. Each author profile, in fact, is a particular term distribution. The statistical models make use of these terms distributions (terms and term frequencies) to determine the similarity between them. The three statistical models employed are the $\chi^2$ distance (Section 5.2.1), the Kullback-Leibler Divergence (KLD, Section 5.2.2) and Delta (Section 5.2.3). They were first introduced in this form in Savoy's work [114].

### 5.2.1 $\chi^2$ distance

A standard way of measuring is the $\chi^2$ distance

$$\chi^2\big(\mathrm{Q},\mathrm{A}_j\big)=\sum_{i=1}^{m}\frac{\big(\mathrm{q}(t_i)-\mathrm{a}_j(t_i)\big)^2}{\mathrm{a}_j(t_i)} \tag{5.1}$$

originally presented by Pearson [93]. In equation 5.1 we presented $\chi^2$ as in the formulation employed in recent previous work [44, 114], where $t_i$ is the relative term frequency of term $i$ in the "query" document Q and in each of the reference documents $A_j$, while $m$ is the total number of terms used for the computation. In our domain $Q$ is an unknown author profile we want to link to the best matching from the set of known profiles $A_j$.

The origin of $\chi^2$ is in the field of probability theory and statistics, where it is typically used to measure the significance difference between observed data and expected data. The greater the distance, the more the observed data diverge with respect to the expected data. One can then conclude that the two sets of data (the author profiles) are not related. To the contrary, if the distance is small, then also the author profiles are more likely to be similar, thus generated by the same author. In this study we are using the $\chi^2$ distance following the same intuition, using one user profile as a "query" and measuring its distance to each candidate user profile. Each author profile is composed of terms that form a distribution, which can vary from author to author and from setting to setting, depending on the assumption we are making to build each profile. For example, the total number of terms $i = 1 \dots m$ depends on the assumption of the minimum document frequency for each term. In the original formulation [44, 114] this was tested at different levels (2, 5, 10) but we are making different (and more exhaustive) choices. We present these choices in Section 5.3. Having computed the $\chi^2$ distance between a "query" and all the user profiles, we minimise it to find the most likely profile, with the assumption that the distance between the profiles is minimal when they are most similar, that is when they are the same or generated by the same author.

### 5.2.2 Kullback-Leibler Divergence

Another way of measuring is the Kullback-Leibler (KLD) Divergence:

$$\mathrm{KLD}(\mathrm{Q}||\mathrm{A}_j) = \sum_{i=1}^{m} p_q(t_i) \cdot \log_2\Big[\frac{p_q(t_i)}{p_j(t_i)}\Big] \tag{5.2}$$

The Kullback-Leibler Divergence (KLD) (or relative entropy) is an asymmetric measure of disagreement[3] between two probabilistic distributions, which derives directly from the concept of entropy [26]. In analogy to $\chi^2$, if two distributions were generated by the same process, or if two user profiles were generated by the same user, their dissimilarity and thus their KLD distance, would be minimal. For this reason we are computing the KLD between an unknown user profile ("query" document) and all the other profiles. We then minimise it for finding the closest profile and thus the user associated with the unknown profile.

[3] It is not a distance in the strict sense because it is not symmetric and does not respect the triangular inequality property [26].

Furthermore, in previous work it has been demonstrated that KLD is an effective indicator of the similarity between two texts [64] and that it can be used successfully to solve authorship attribution problems [62, 63, 66, 141]. In Equation 5.10 we indicate with $p_q(t_i)$ the probability of a particular term $t_i$ in the "query" document $q$, while $p_j(t_i)$ identifies the probability of the same term $t_i$ in a reference document $j$.

To estimate the probability of a term in a document, we first adopt the Maximum Likelihood Estimation (MLE), under which the probability of a term in a particular document is equal to the frequency of the term in the document ($tf_i$) divided by the total number of tokens in the documents ($n$). Beside the MLE, we also adopt a smoothing technique [81] based on Lindstone's law, as suggested in [114]. This allows null probabilities to be discarded due to the absence of the terms, which happens when working on limited words sets, like in our case, and prevents probabilities from going to infinity due to a null denominator. Being Lindstone's smoothing technique satisfactory for our study, we decided not to investigate other smoothing techniques, such as the Laplace smoothing or Dirichlet smoothing. In equation 5.3 we represent the full formula of the MLE with the smoothing parameter $\lambda$, adjustable as desired, and $|V|$ the vocabulary size of the entire corpus. It should be noted that for $\lambda = 0$, Equation 5.3 represents the formula for computing MLE alone and for $\lambda = 1$, it represents the formula of Laplace smoothing.

$$p(t_i) = \frac{tf_i + \lambda}{n + \lambda \cdot |V|} \tag{5.3}$$

### 5.2.3 Delta

The Delta score is a measure of similarity widely employed in Authorship Identification [53, 1, 114].

Delta is defined as:

$$\Delta(Q, A_j) = \frac{1}{m} \cdot \sum_{i=1}^{m} \left| \mathrm{Z}_{\mathrm{score}}(t_{iq}) - \mathrm{Z}_{\mathrm{score}}(t_{ij}) \right| \tag{5.4}$$

where $Q$ is an author in the testing set, $A_j$ the set of authors from the training set, $m$ the number of terms for which the $\Delta$ is computed, while $\mathrm{Z}_{\mathrm{score}}$ is the score computed for the two authors over the same term $i$.

The $\mathrm{Z}_{\mathrm{score}}$ is defined as:

$$\mathrm{Z}_{\mathrm{score}}(t_{ij}) = \frac{\mathrm{tfr}_{ij} - \mathrm{mean}_i}{\mathrm{stdv}_i} \tag{5.5}$$

for a term $i$ in a document $j$, having document relative frequency of $\mathrm{tfr}_{ij}$ and mean value of $\mathrm{mean}_i$ and standard deviation of $\mathrm{stdv}_i$ in the collection. Like for $\chi^2$ and KLD, also with Delta we are measuring the similarity between terms distributions that represents unknown user profiles and sets of known user profiles, with the explicit goal to match unknown and known profiles with low value of Delta, thus with high similarity.

To successfully employ these classifiers for author identification, there are a set of well established steps that must be followed. The first step is *building author profiles*, which is generally achieved by concatenating all the texts written by an author into a single document. This becomes the author profile. The author profile is then *compared* to all the possible queries, typically other unlabelled user profiles. The comparison is in fact done using $\chi^2$, KLD and Delta, minimising the distances to obtain the best matching author profiles. The *list of features* or terms that "support" the classification process are defined as next step and are presented in more detail in the next section. These include all the standard methods for traditional collections as well as the novel method we proposed for conversational documents.

## 5.3 Features Selection Approaches

In this section we describe the three strategies used for selecting the list of features, later employed for the user classification. In our domain the features are represented by the terms and their frequencies produced by authors in writing the texts. When grouped together by author, all the texts of an author build the author profile. Since the features employed to build author profiles are terms, those are also called "author vocabulary" or simply "vocabulary".

In the traditional domain of authorship identification, the documents that constitute author profiles are long (at least 200 [72], 250 [142] or more [73] terms) since they are representative of standard collections like newspaper articles [114], poems [64], letters [64], emails [2] or blog posts [72]. In the case of conversational documents, however, the situation is different: the length of a text is often below 60 terms. For this reason we claim that the standard feature selection algorithms employed for the traditional documents cannot be successfully employed for the conversational documents. To test and verify this claim, we investigated different techniques of profile building, from the traditional ones (stopwords based) to our novel ones (based on interlocutors influenced vocabulary).

### 5.3.1 Stopwords Vocabulary

The first approach we are testing is the one based on stopwords, which has been proved to work well to classify the profile of author of traditional documents. We decided to employ this approach with conversational documents to demonstrate that standard stopwords lists are not easily transferrable to the collections of conversational documents and that we need different techniques of feature selection for these collections.

According to Manning et al. [81], a standard stopwords list contains between 7 and 300 terms determined by observing the frequency of all terms in the collections and extracting those with higher frequency. We refer to this method as *Term Frequency*

TF: [the, to, i, a, it, html, that, is, http, of, in, and, public, for, you, org, s, t, be, w, on, not, bug, lists, archives, with, if, have, but, bugzilla, as, what, are, can, or, so, just, at, do, this, there, should, an, com, like, we, would, about, was, hixie, from, don, no, m, think, bugmail, they, all, by, new, use, some, spec, one, more, how, www, my, d, yeah, has, me, will, when, whatwg, which, web, get, any, well, people, make, then, does, up, only, now, work, doesn, e, see, could, know, c, ok, time, than, seems, out, re, really, your, element, need, too, way, something, also, other, text, want, good, why, them, p, r, right, because, oh, sure, wiki, philip, ll, using, he, content, css, yes, title, hsivonen, though, planet, thanks, page, list, same, changes, much, actually, attribute, still, case, add, maybe, mozilla, ie, mikesmith, ve, where, irc, guess, support, diffs, here, document, video, annevk, google, even, since, things, might, issue, dom, browser, microformats, test, been, may, mean, data, gsnedders, probably, type, into, wg, say, b, thing, id, better, elements, used, code, jgraham, first, problem, being, opera, point, before, stuff, going, xml, change, those, did, isn, sep, ah, blog, section, lachy, browsers, name, oct, canvas, thread, hmm, file, firefox, already, had, link, aug, who, example, their, xhtml, value, got, user, anything, logs, set, svg, go, different, show, least, didn, very, g, messages, x, script, back, most, yet, anyone, working, idea, last, non, input, js, instead, source, makes, pretty]

NIDF: [the, to, i, a, it, html, that, is, http, of, in, and, public, for, you, org, s, t, be, w, on, not, bug, lists, archives, with, if, have, but, bugzilla, as, what, are, can, or, so, just, at, do, this, there, should, an, com, like, we, would, about, was, hixie, from, don, no, m, think, bugmail, they, all, by, new, use, some, spec, one, more, how, www, my, d, yeah, has, me, will, when, whatwg, which, web, get, any, well, people, make, then, does, up, only, now, work, doesn, e, see, could, know, c, ok, time, than, seems, out, re, really, your, element, need, too, way, something, also, other, text, want, good, why, them, p, r, right, because, oh, sure, wiki, philip, ll, using, he, content, css, yes, title, hsivonen, though, planet, thanks, page, list, same, changes, much, actually, attribute, still, case, add, maybe, mozilla, ie, mikesmith, ve, where, irc, guess, support, diffs, here, document, video, annevk, google, even, since, things, might, issue, dom, browser, microformats, test, been, may, mean, data, gsnedders, probably, type, into, wg, say, b, thing, id, better, elements, used, code, jgraham, first, problem, being, opera, point, before, stuff, going, xml, change, those, did, isn, sep, ah, blog, section, lachy, browsers, name, oct, canvas, thread, hmm, file, firefox, already, had, link, aug, who, example, their, xhtml, value, got, user, anything, logs, set, svg, go, different, show, least, didn, very, g, messages, x, script, back, most, yet, anyone, working, idea, last, non, input, js, instead, source, makes, pretty]

INDRI: [x, y, your, yours, yourself, yourselves, you, yond, yonder, yon, ye, yet, z, zillion, j, u, umpteen, usually, us, username, uponed, upons, uponing, upon, ups, upping, upped, up, unto, until, unless, unlike, unliker, unlikest, under, underneath, use, used, usedest, r, rath, rather, rathest, rathe, re, relate, related, relatively, regarding, really, res, respecting, respectively, q, quite, que, qua, n, neither, neaths, neath, nethe, nethermost, necessary, necessariest, necessarier, never, nevertheless, nigh, nighest, nigher, nine, noone, nobody, nobodies, nowhere, nowheres, no, noes, nor, nos, no-one, none, not, notwithstanding, nothings, nothing, nathless, natheless, t, ten, tills, till, tilled, tilling, to, towards, toward, towardest, towarder, together, too, thy, thyself, thus, than, that, those, thou, though, thous, thouses, thoroughest, thorougher, thorough, thoroughly, thru, thruer, thruest, thro, through, throughout, throughest, througher, thine, this, thises, they, thee, the, then, thence, thenest, thener, them, themselves, these, therer, there, thereby, therest, thereafter, therein, thereupon, therefore, their, theirs, thing, things, three, two, o, oh, owt, owning, owned, own, owns, others, other, otherwise, otherwisest, otherwiser, of, often, oftener, oftenest, off, offs, offest, one, ought, oughts, our, ours, ourselves, ourself, out, outest, outed, outwith, outs, outside, over, overallest, overaller, overalls, overall, overs, or, orer, orest, on, oneself, onest, ons, onto, a, atween, at, athwart, atop, afore, afterward, afterwards, after, afterest, afterer, ain, an, any, anything, anybody, anyone, anyhow, anywhere, anent, anear, and, andor, another, around, ares, are, aest, aer, against, again, accordingly, abaft, abafter, abaftest, abovest, above, abover, abouter, aboutest, about, aid, amidst, amid, among, amongst, apartest, aparter, apart, appeared, appears, appear, appearing, appropriating, appropriate, appropriatest, appropriates, appropriater, appropriated, already, always, also, along, alongside, although, almost, all, allest, aller, allyou, alls, albeit, awfully, as, aside, asides, aslant, ases, astrider, astride, astridest, astraddlest, astraddler, astraddle, availablest, availabler, available, aughts, aught, vs, v, variousest, variouser, various, via, vis-a-vis, vis-a-viser, vis-a-visest, viz, very, veriest, verier, versus, k, g, go, gone, good, got, gotta, gotten, get, gets, getting, b, by, byandby, by-and-by, bist, both, but, buts, be, beyond, because, became, becomes, become, becoming, becomings, becominger, becomingest, behind, behinds, before, beforehand, beforehandest, beforehander, bettered, betters, better, bettering, betwixt, between, beneath, been, below, besides, beside, m, my, myself, mucher, muchest, much, must, musts, musths, musth, main, make, mayest, many, mauger, maugre, me, meanwhiles, meanwhile, mostly, most, moreover, more, might, mights, midst, midsts, h, huh, humph, he, hers, herself, her, hereby, herein, hereafters, hereafter, hereupon, hence, hadst, had, having, haves, have, has, hast, hardly, hae, hath, him, himself, hither, hitherest, hitherer, his, how-do-you-do, however, how, howbeit, howdoyoudo, hoos, hoo, w, woulded, woulding, would, woulds, was, wast, we, wert, were, with, withal, without, within, why, what, whatever, whateverer, whateverest, whatsoeverer, whatsoeverest, whatsoever, whence, whencesoever, whenever, whensoever, when, whenas, whether, wheen, whereto, whereupon, wherever, whereon, whereof, where, whereby, wherewithal, wherewith, whereinto, wherein, whereafter, whereas, wheresoever, wherefrom, which, whichever, whichsoever, whilst, while, whiles, whithersoever, whither, whoever, whosoever, whoso, whose, whomever, s, syne, syn, shalling, shall, shalled, shalls, shoulding, should, shoulded, shoulds, she, sayyid, sayid, said, saider, saidest, same, samest, sames, samer, saved, sans, sanses, sanserifs, sanserif, so, soer, soest, sobeit, someone, somebody, somehow, some, somewhere, somewhat, something, sometimest, sometimes, sometimer, sometime, several, severaler, severalest, serious, seriousest, seriouser, senza, send, sent, seem, seems, seemed, seemingest, seeminger, seemings, seven, summat, sups, sup, supping, supped, such, since, sine, sines, sith, six, stop, stopped, p, plaintiff, plenty, plenties, please, pleased, pleases, per, perhaps, particulars, particularly, particular, particularest, particularer, pro, providing, provides, provided, provide, probably, l, layabout, layabouts, latter, latterest, latterer, latterly, latters, lots, lotting, lotted, lot, lest, less, ie, ifs, if, i, info, information, itself, its, it, is, idem, idemer, idemest, immediate, immediately, immediatest, immediater, in, inwards, inwardest, inwarder, inward, inasmuch, into, instead, insofar, indicates, indicated, indicate, indicating, indeed, inc, f, fact, facts, fs, figupon, figupons, figuponing, figuponed, few, fewer, fewest, frae, from, failing, failings, five, furthers, furtherer, furthered, furtherest, further, furthering, furthermore, fourscore, followthrough, for, forwhy, fornenst, formerly, former, formerer, formerest, formers, forbye, forby, fore, forever, forer, fores, four, d, ddays, dday, do, doing, doings, doe, does, doth, downwarder, downwardest, downward, downwards, downs, done, doner, dones, donest, dos, dost, did, differentest, differenter, different, describing, describe, describes, described, despiting, despites, despited, despite, during, c, cum, circa, chez, cer, certain, certainest, certainer, cest, canst, cannot, cant, cants, canting, cantest, canted, co, could, couldst, comeon, comeons, come-ons, come-on, concerning, concerninger, concerningest, consequently, considering, e, eg, eight, either, even, evens,evenser, evensest, evened, evenest, ever, everyone, everything, everybody, everywhere, every, ere, each, et, etc, elsewhere, else, ex, excepted, excepts, except, excepting, exes, enough]

*Figure 5.2.* Examples of stopwords from TF computation, NIDF computation or Indri list.

*(TF)*. However, in the work of Lo et al. [80] better results were obtained by constructing a stopwords list considering the normalised inverse document frequency (NIDF) of the terms. NIDF was originally defined by Robertson and Jones [108] as:

$$\text{NIDF}_{k\ \text{Norm}} = log\left(\frac{\text{NDoc} - \text{D}_k + 0.5}{\text{D}_k + 0.5}\right) \tag{5.6}$$

where NDoc is the total number of documents in the collection and $\text{D}_k$ is the number of documents containing the term $k$. We refer to this as the *NIDF* method.

Finally, besides these two strategies for generating the list of stopwords, we also considered a standard list taken from the Indri search engine[4]. We refer to this method as the *Indri* one. In Figure 5.2 we reported an example of stopwords computed with TF and NIDF and the list of terms included in the Indri list of stopwords.

### 5.3.2 Simple Author Vocabulary

The second feature selection strategy we considered is the common practice of concatenating all texts generated by a single author together to build the profile for that author [64, 121, 114]. This is a trivial approach that for conversational documents might be applied without any drawbacks given the shortness of these documents. However, it requires more attention to the computational time needed in case of longer documents like newspapers. Given some pre-filtering, we are still able to apply this method to traditional collections of newspapers without incurring major computational problems.

The following is a toy example of the techniques applied to a representative of conversational documents. The document fragment consists in a conversation of 4 lines with 3 users (A, B and C) involved:

```
1  Thu Jan 15 15:28:00 CET      A      that comment clearly missed the point
2  Thu Jan 15 15:29:00 CET      B      anne: which comment?
3  Thu Jan 15 15:39:00 CET      A      in bug 6439
4  Thu Jan 15 15:42:00 CET      C      gotta love public bug systems :)
```

With the *simple author vocabulary* technique, the profile of one user is generated from his messages only. Reported are username, lines of the conversation used to build the user profile, text of the profile. We refer to this strategy of feature selection as *x0* or *baseline*.

**Method x0 -baseline** The profile of one user is generated from his messages only.

```
A [1, 3] that comment clearly missed the point in bug 6439
B [2]    anne: which comment?
C [4]    gotta love public bug systems :)
```

[4] http://www.lemurproject.org/

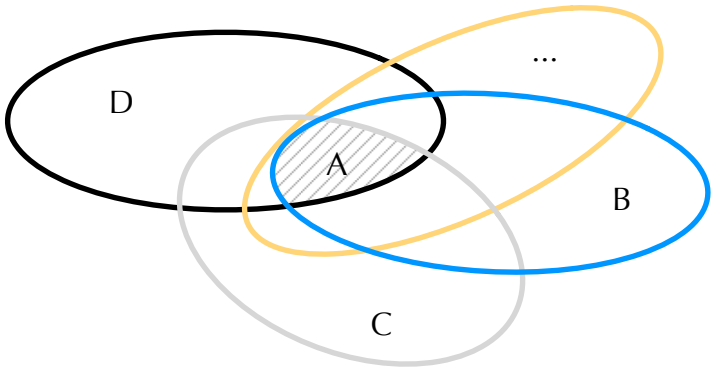



*Figure 5.3.* Specific vocabulary for author A vs. {AB, AC, AD, ...}

### 5.3.3 Interlocutors Influenced Vocabulary

The third and last approach for creating user profiles based on term selection focuses on conversational documents. Our hypothesis is that classical techniques based on stopwords lists or simple author profiles do not apply well to conversational documents, therefore we need to employ novel approaches.

We already presented the properties of conversational documents (Chapter 2). However we are recalling here below the 2 main features that motivate our method:

1. the property that a user's message has an impact on all the future messages in a conversation, and

2. the fact that users need ways of emulating the non verbal expressions that can be found in a regular in-presence conversation, thus creating a more personal and novel language style, no longer related to the common language of all the users.

These two properties also influenced the two different steps for our proposed approach to generate user profiles, that are:

1. **Vocabulary Building**, in which authors participating in the same conversations mutually influence their vocabulary;

2. **Terms Selection**, in which terms belonging to authors vocabularies are ordered by their importance and used in increasing percentage.

### Vocabulary Building

In the first part of the proposed method we make use of the first property of conversations and analyse all the conversations in our dataset to find the group of users participating in the same conversation. For each user we consider the list of the other users he talked to and generate a new list of profiles corresponding to each couple of users with their own respective joint terms usage. This procedure allows the vocabulary of one user to be influenced by the vocabulary of the other users he is talking with. For example, if user A is talking with B and C, we will generate the profiles AB (all the messages of A in that conversation are merged with the messages of B in that conversation) and AC, but also BC since both B and C also participated in the same conversation. In this formulation AB and BA are equal and mutually exchangeable, so we do not distinguish between A sending a message to B or B sending a message to A. In other words we are not considering the temporal aspect of the conversation. Once we have generated all the couples for one author, we create an extended profile for that author. As in the previous example, if for user A we have couples AB, AC, AD etc, then we merge them together and call it A*. The intuition behind this procedure is illustrated in Figure 5.3: all the couples have A in common and our goal is to identify the intersection, i.e. the specific vocabulary of A among all his conversations.

The approach can be applied at different levels, depending on the involvement of the other participants in the conversation in one user profile. We represent the three different modalities of such participants involvements in the example below, built on the fragment of conversation presented in the toy-example of Section 5.3.2.

**Method x1** The profile of one user is generated from his messages and the messages after his ones.

```
A [1, 2, 3, 4] gotta love public bug systems :) in bug 6439 anne: which comment? that
               comment clearly missed the point
B [2, 3]       in bug 6439  anne: which comment?
C [4]          gotta love public bug systems :)
```

**Method x2** The profile of one user is generated from his messages and the messages before and after his ones.

```
A [1, 2, 3, 4] gotta love public bug systems :) in bug 6439 anne: which comment? that
               comment clearly missed the point
B [1, 2, 3]    that comment clearly missed the point in bug 6439  anne: which comment?
C [3, 4]       in bug 6439 gotta love public bug systems :)
```

**Method x3** The profile of one user is generated from all the messages of all the users involved in the conversation.

```
A [1, 2, 3, 4] gotta love public bug systems :) in bug 6439 anne: which comment? that
               comment clearly missed the point
B [1, 2, 3, 4] gotta love public bug systems :) in bug 6439 anne: which comment? that
```

```
                comment clearly missed the point
C [1, 2, 3, 4] gotta love public bug systems :) in bug 6439 anne: which comment? that
                comment clearly missed the point
```

It it easy to see how method x3 leads to a confusion of auhtors profiles, for which it is impossible to distinguish between authors, even in such a simple example. For this reason it was excluded from our investigation, that focused on the other two examples instead. Methods x1 and x2 are the two features selection methods that we therefore employed in our study. They represent two different levels of profile expansions, for which the interlocutors play a bigger (x2) or smaller (x1) role.

### Terms Selection

The second step of the method takes inspiration from the second properties of the conversations, the one that implies the creation of a "novel language" by authors of conversational documents. We believe, in fact, that each user of conversational services (like IRC chats or Twitter) uses a very specific language that needs to be identified, as opposed to text produced in traditional media for which the stopwords are good features. To identify such a specific vocabulary of a user, we should be able to order the terms inside each user profile according to its specificity. In literature the *TF-IDF* scoring (introduced in Section 4.2.3) has been proved to work quite well for such tasks. However we believe that having to compare distributions of terms, a better choice would be the adoption of an entropy-based measure, such as the KLD (Section 5.2.2). To verify this intuition, we use both the TF-IDF and KLD for the terms ordering and selection in all experiments.

**TF-IDF** We first recall the TF-IDF formula, which is composed by the product between the relative frequency of a term in the document (TF) and the importance of the term in the collection (IDF). In this specific application we consider as documents the different authors profiles: we therefore measure the relative frequency of terms in each author profile and their importance in all the collection, constituted by author profiles.

$$tf_{ik} = \frac{f_{ij}}{\sum_{k=1}^{t} t_{ik}} \tag{5.7}$$

$$idf_k = \log \frac{N}{df_k} \tag{5.8}$$

In Equation 5.7, $f_{ij}$ is the number of occurrences of a term $j$ in a particular author profile $i$, normalised by the length of the author profile. In Equation 5.8, $N$ is the total number of documents in the collection of author profiles, while $df_i$ is the number of author profiles in which the term $k$ occurs. The final score is obtained by multiplying the two components together $tf_{ik} \cdot idf_k$.

**KLD** To compute the point-wise KLD score for each term in each author profile, we calculate the divergence between each author profile A* and the collection of all user profiles Γ.

$$\text{KLD(A*}||\Gamma) = p_{A*}(t) \cdot \log_2\Big[\frac{p_{A*}(t)}{p_\Gamma(t)}\Big] \tag{5.9}$$

In Equation 5.9, for all the terms $t$ belonging to the user profile A* we compute the KLD, where $p_{A*}(t)$ is the probability of term $t$ in the user profile A* and $p_\Gamma(t)$ is the probability of the term $t$ in the collection $\Gamma_{A*}$. The probability is estimated through MLE with Lindstone smoothing as in Equation 5.3, with $\lambda = 0.1$ and $|V| = |\Gamma|$.

```
KLD:    [reallybigindex, books, tutorial, archive, java, thx, download, sun, os, docs, windows,
         ok, one, href, my, can, html, com, in, is, it, i, http, a]

TF-IDF: [reallybigindex, books, tutorial, sun, docs, java, archive, html, com, thx, download,
         os, http, windows, ok, href, one, my, can, a, in, is, it, i]
```

```
KLD: [ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx,
     oposite, windowsxp, strange, hmmm, nick, protected, pc, sep, stopped, asks, ideas, showing,
     connections, smb, appear, group, cant, iptables, remove, isnt, share, anybody, computer,
     via, box, samba, having, under, doesnt, write, network, few, directory, test, conf, open,
     o, log, possible, things, doing, seems, idea, wrong, he, trying, still, same, help, other,
     them, user, hi, lol, by, something, now, yes, windows, want, linux, problem, am, was, set,
     no, this, just, has, what, file, not, t, for, s, of, see, me, you, any, as, a, have, one,
     my, and, is, on, can, that, but, it, in, to, i, the]

TF-IDF: [ringostarr, ucit, folders, shared, dr, drwxrwsrwt, bodyrider, comuter, authorised, rwx,
     strange, oposite, pc, hmmm, windowsxp, nick, samba, ideas, o, group, cant, sep, box, lol,
     protected, stopped, asks, iptables, remove, smb, windows, see, linux, write, connections,
     showing, having, appear, am, possible, user, share, isnt, one, anybody, computer, conf,
     log, via, network, under, directory, doesnt, test, set, open, file, few, problem, any, he,
     has, things, seems, wrong, idea, doing, trying, but, same, still, as, them, my, other,
     help, me, by, can, something, hi, now, in, yes, on, the, want, was, to, that, no, have, i,
     this, just, it, what, and, s, not, t, is, of, for, you, a]
```

*Figure 5.4.* Two examples of author profiles for which the terms have been ordered using KLD and TF-IDF.

The results of the author profile scoring with TF-IDF and KLD is illustrated in Figure 5.4. In the figure we report two examples of author profiles for which the terms have been ordered using the two scoring algorithms. The first terms of each list (i.e. *reallybigindex, books, tutorial* and *ringostarr, unit, folders*) are the most relevant to the profile, the last terms (e.g. *is, it, i, http,a, for you, that, ...*) the least relevant.

As can be seen by the two lists, both scoring algorithms are able to separate stopwords (common terms) and specific terms (topical terms) but with differences. These differences seem to be more evident for the stopwords than for the topical terms. For

example, the last five terms in each list, which are the most common words, are all different if we observe TF-IDF and KLD, while the first five words, the topical words, are almost the same for both TF-IDF and KLD. This is an interesting observation. To investigate these differences in greater detail, we analysed the user profiles in the collections at different percentage of vocabulary, starting from the most specific (topical) terms to the least ones. Our intuition is that specific terms are the most distinctive for author profiles of conversational documents, while the common terms are effective, as in the literature, for collections of traditional written documents.

In the next Section 5.4 we describe the experimental settings and all the design choices taken during the evaluation of the different methods presented, in particular the proposed one, the *Interlocutors Influenced Vocabulary* (as in Section 5.3.3).

## 5.4 Experimental Settings and Evaluation

In this section we are presenting the settings employed in our experimental framework: we will first specify the datasets used (Section 5.4.1), we will then describe the general settings employed for each collection and each method (Section 5.4.2), finally we will describe the evaluation function employed (Section 5.4.3). In the last part of the chapter (Section 5.5) we will present the results derived from the experiments.

### 5.4.1 Datasets

| |
|---|
| aix apache azureus blender c cisco csharp css debian fedora flood freebsd gentoo gentoo-dev gtk hardware html iptables irix java javascript linux-bg macosx mysql netbsd openbsd opensolaris oracle php python qt reactos samba solaris suse tomcat ubuntu vim windows wireless |

*(a)* Channels in “Freenode”

| |
|---|
| accessibility activity css developers fx html-wg html5 microformats wai-aria webapps whatwg xhtml |

*(b)* Channels in “Krijn”

*Figure 5.5.* List of all the channels of each collection of IRC logs used in the authorship identification study.

The focus of our investigation is on conversational documents and for this reason in our experiments we employed a subset of the collections of online conversations presented in Chapter 3. The collection includes 4 different datasets of IRC logs (“perverted justice”, “Krijn”, “irclogs” and “omegle”) and was originally designed to solve problems related to sexual predator identification. In this chapter, however, we are not interested

in exploring those kinds of problems: we want to be able to identify authors in conversations rather than profiling them into specific categories (predator/non-predator). However, not all datasets were suited for our experiments: for one dataset ("omegle") in the collection, every author had only one document produced and of a limited length, while for another one ("perverted justice") the conversations were between two users only. We believe these datasets are too specific and not general enough to be employed in an experimental study like ours. For these reasons, we decided not to use those datasets as testbed in our study and employ in the experiments only the two remaining ones: "Krijn" and "irclogs". We renamed the last datatset ("irclogs") into "Freenode", from the name of the specific IRC network the documents were downloaded from. We individually analyzed the *Krijn* and *Freenode* datasets and perform individual experiments on each of them. Although our study does not focus directly on topicality, we noticed that the topicality of the two datasets was to some extent homogeneous. Documents from Krijn, in fact, are centered on topics related to HTML 5 (e.g. html5, css, micro formats, accessibility, ...) while the ones from Freenode are somewhat more diverse, ranging from java, gentoo and macosx to php, oracle and samba (see Figure 5.5 for the complete list of topics). Despite this topical homogeneity, users often engaged in conversations that diverged from the expected topic. They introduce, in fact, discussions about family, general interests and sometimes even anger. This is interesting because it makes the two datasets more diverse and representative of the conversations that take place online. In addition to this collection of "pure" conversational documents we decided to test our algorithm also on a "less conventional" collection. In fact, we decided to employ the collection of Twitter messages already segmented in conversations (the "Microsoft Conversation in Twitter Collection" presented in Chapter 3). Due to the fact that we observed in Chapter 4 some properties in common between the IRC logs and the Twitter messages, we wanted to test how these could be of help in a real scenario like this one.

Besides these collections of conversational documents, we also employed in our experiments three collections of traditional documents: *Associated Press* (AP) from the Tipster Collection and *The Glasgow Herald* plus *La Stampa* from the CLEF collection. These three collections were already presented in Chapter 3. The reason to employ them here is mainly to prove that state-of-the-art and simple methods for authorship attribution on traditional collections work well but do not scale when we increase the number of authors under investigation. This is more evident when we enlarge the test set and try to identify hundreds of authors. One last remark has to be done: since these collections contain documents that are newspaper articles not involving any discussion, we could not test the novel methods designed for conversational documents on them (Figure 5.1).

In Table 5.1 we reported the statistics for each of the collection employed and noted the differences between the statistics of Twitter and IRC logs (Krijn and Freenode) documents. These latter ones, in fact, are similar in terms of average profile length, despite

the fact that one contains manifestly more users and documents than the other. Twitter on the other hand has definitely more users and documents, but the average profile length is one order of magnitude smaller than the other two. Despite that, we still consider Twitter a sort of conversational media, like the IRC logs, but we are aware of possible significative differences between the two. Traditional newspaper documents, on the other hand, contain definitely fewer documents and fewer authors than the other collections. The length of documents, however, is comparable within the three collections but, as expected, is much higher than in the collections of conversational documents.

### User profiles

To test the different methods (*stopwords* and *vocabulary selection* by percentage) with the two different sets of collections (*traditional documents* and *conversational documents*) as in the experimental plan presented in Figure 5.1), we employed two different sets of users profiles. The first set of user profiles reflects the current state-of-the-art approaches [114, 78, 77] in which the number of user profiles is relatively small, in the order of few dozens, and generally no more than 50 [112]. In the work of Layton et al. [78, 77], for example, 10 to 50 authors have been considered, while in Savoy's one Savoy [114] they are 20. For these reasons, we also set to 20 in our first experiments the number of user profiles to be identified. We therefore extracted a subset of 20 random users in both collections of conversational and traditional documents. Moreover, for conversational documents we selected the 20 users trying to preserve an average profile length which could be as much as possible comparable to standard documents.

The second set of user profiles is a less established one and aims at studying the applicability of traditional methods (based on stopwords) to large sets of users. Despite the fact that this setting is less frequent in case of traditional documents, it becomes of great interest in case of the collections of conversational documents, for which hundreds of users are active at the same and for which different real-case scenarios can be of interest, in particular in the filed of security or advertisement. For this reason the second set of users employed in our experiments is composed of hundreds or thousands of users of conversational systems or authors of newspaper articles. We randomly selected the number of users employed as "test set" to make it big enough to be realistic but small enough to be computed in a reasonable time (in particular for Twitter). The users not inserted in the test set were used anyway as "noise" to make the identification of the right author more challenging but at the same time more realistic.

**Interlocutors and Conversations** In Section 5.3.3, when describing our novel approach *Interlocutors Influenced Vocabulary* we already mentioned the central role of the interlocutors for each of the two methods of profile expansion x1 and x2. In Figure 5.6 we displayed the distribution of interlocutors per user and we noticed some dif-

| Dataset | Number Documents | Total Users | Selected Users | Avg. Profile Length Tokens | Avg. Profile Length Singletons |
|---|---|---|---|---|---|
| AP | 119,495 | 2,348 | 1,776 | 531.87 | 266.13 |
| La Stampa | 35,076 | 1,210 | 943 | 524.27 | 293.36 |
| Glasgow Herald | 26,113 | 780 | 536 | 726.59 | 338.57 |
| Freenode | 93,327 | 4,646 | 705 | 58.07 | 39.05 |
| Krijn | 78,605 | 19,046 | 2,866 | 60.80 | 39.90 |
| Twitter | 2,718,101 | 295,996 | 2,416 | 3.83 | 1.63 |

*(a)* Statistics of the datasets. The first group of collections is representative of traditional documents (newspapers) while the second group is representative of conversational documents (IRC logs and Twitter). For each collection is indicated the total number of documents and users and the number of randomly selected users employed during the testing. It is also reported the average user's profile length as number on total terms generated (token) as well as unique terms (singleton).

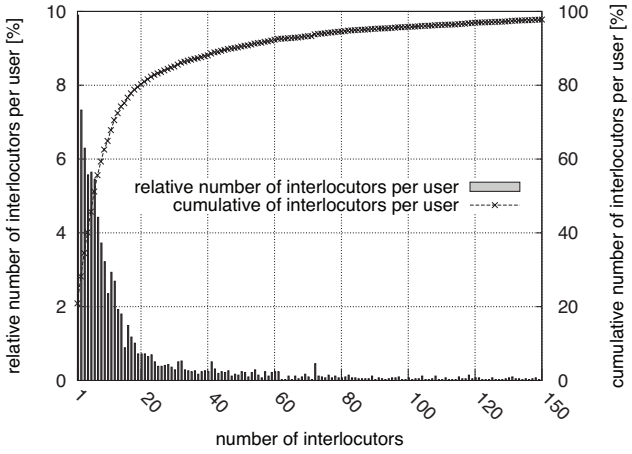
| Dataset | Number Documents | Users | Avg. Profile Length Tokens | Avg. Profile Length Singletons |
|---|---|---|---|---|
| AP | 4,187 | 20 | 693.54 | 325.33 |
| La Stampa | 4,273 | 20 | 665.75 | 366.34 |
| Glasgow Herald | 5,416 | 20 | 714.67 | 343.20 |
| Freenode | 13,282 | 20 | 139.02 | 81.68 |
| Krijn | 3,247 | 20 | 136.74 | 78.01 |
| Twitter | 2,831 | 20 | 124.03 | 64.32 |

*(b)* Statistics of the subset of 20 randomly selected users for each dataset.

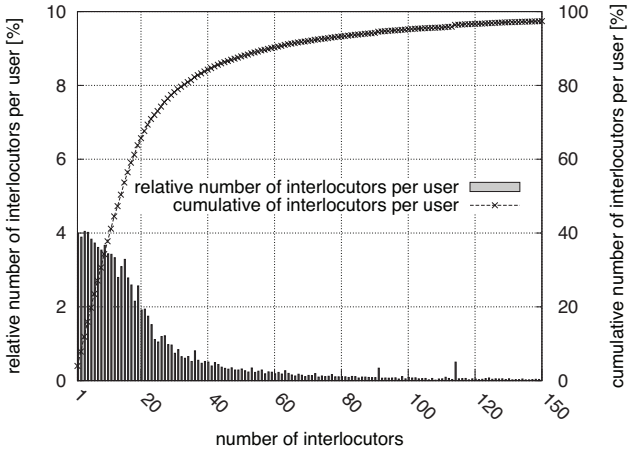
*Table 5.1.* Datasets statistics

ferences among the collections: despite having the biggest number of documents and users, Twitter has a very low number of interlocutors per users. In fact, around 90% of its users have less than 10 interlocutors, i.e. a person whom they direct a message to. The other two collections, Freenode and Krijn, on the other hand, show different properties: they are more similar to one another, in fact 90% of their users have between 40 and 60 interlocutors.

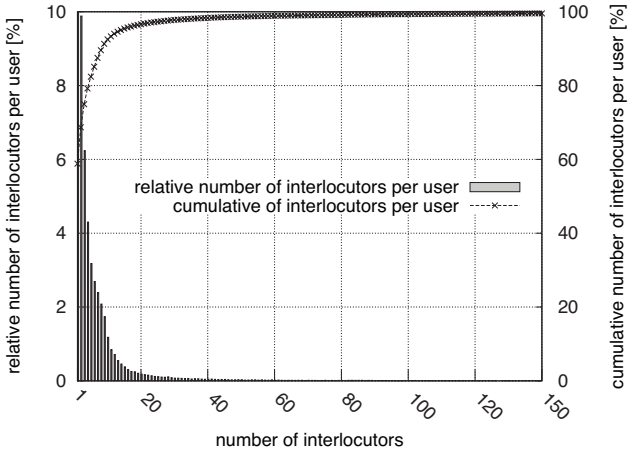
A different behaviour can be observed in the plot of conversations per user in Figure 5.7: a lot of users of Krijn and Freenode engage in few conversations (90% of users have between 8 and 15 conversations) while in Twitter there is the opposite situation, having 90% of the users up to 25 conversations each. A general remark regarding this distribution is that in Twitter people do not engage in conversations with a lot of users but interact much more in different "sessions" of message exchanges, while in IRC chats (Freenode and Krijn) the users tend to engage in fewer conversations but with a lot of other users. These observations are useful to read the results of the experiments (Section 5.5).




*(a)* Freenode dataset.

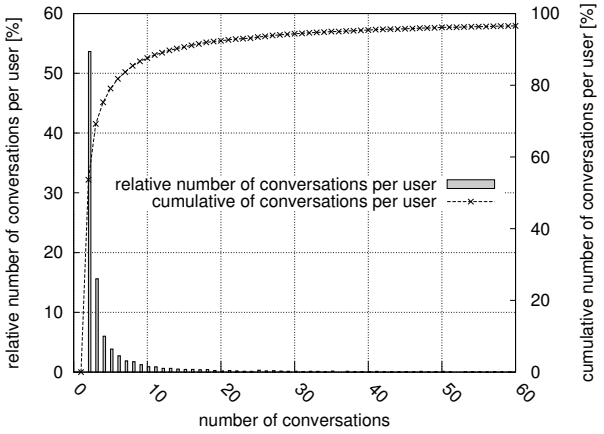



*(b)* Krijn dataset.

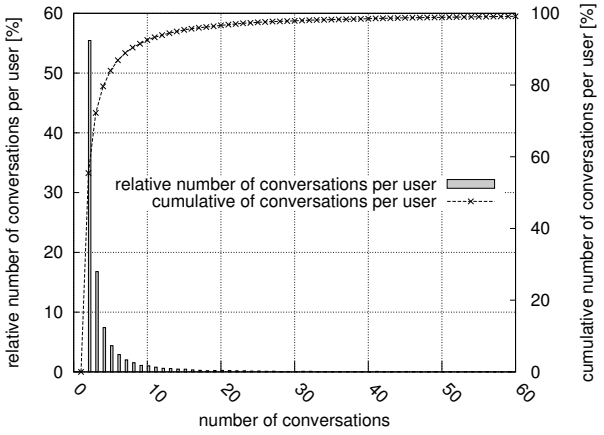



*(c)* Twitter dataset.

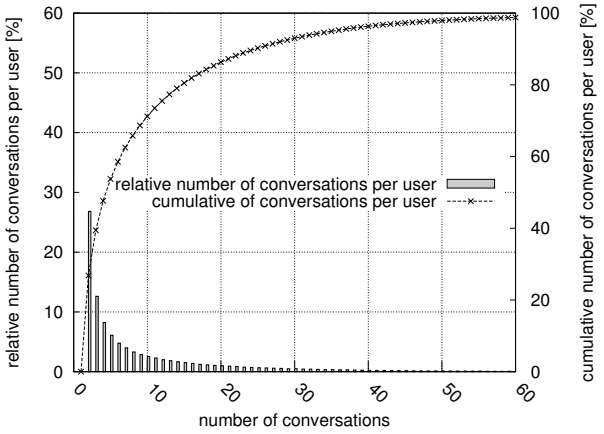
*Figure 5.6.* Relative frequency and cumulative (in percentage) of the number of interlocutors per user in each collection.




*(a)* Freenode dataset.




*(b)* Krijn dataset.




*(c)* Twitter dataset.

*Figure 5.7.* Relative frequency and cumulative (in percentage) of the number of conversations per user in each collection - method independent.

### 5.4.2 General Settings

To identify users among conversations we adopted the scoring and classifying methods presented in Section 5.2 and Section 5.3.3, following the experimental plan of Figure 5.1 and for two different sets of users: 20 users and hundreds of users, randomly selected.

The first step was to randomly divide the conversations in each collection into two sets, identified as *testing* set and *training* set. From each set we derived two lists of users, for which we computed profiles with the different methods: simple concatenation (x0) and interlocutors based (x1 and x2). This operation was performed one time only and it produced as a result three profiles (x0, x1, x2) for each user in each sets (training and testing). We then removed from the test set those users with only one conversation or one interlocutor and those not being represented in the training test. We finally randomly sampled from the remaining users in the test set to obtain the final lists of 20 users and of hundreds of users. For methods x1 and x2 the terms in the users profiles were ordered with the different stategies (KLD and TF-IDF) and employed at different percentage against the users in the training set. For method x0 no ordering was needed.

As a result of the test set construction procedure, the experiments with hundreds of users were conducted using half of all the users in the IRC logs datasets (Freenode and Krijn) and 1/16 of all the users in the Twitter collection. This was also done to keep the computational time to a reasonable order of magnitude (hours instead of days[5]). For traditional collections of newspapers, we employed all the documents and all the users. For the experiments with the subset of 20 users, we employed all the documents associated to each user.

We indexed the author profiles using Lucene[6] with the embedded *SimpleAnalyzer* parser, which tokenises the text by using all non-letter characters as separators and lowercases it [47]. We also experimented with the other predefined Lucene analysers. However, we found that *WhitespaceAnalyzer* was introducing too much noise while *StopAnalyzer* and *StandardAnalyzer* were removing the stopwords which we wanted to preserve. We also preserved the original spelling of each term, not applying any stemming at all. The choice of not removing the stopwords and not performing any stemming has two justifications: i) given the nature of the conversational datasets under investigation, any spelling variation (including mistyping and spelling errors) can be used as indicator of a particular user and ii) traditionally in authorship identification the most frequent words are used to identify the desired user. Following this last remark, in Section 5.5 we will present the results from the first experimental configuration, first the one based on the stopwords, then those based on authors vocabulary

[5]For all experiments we employed a Server DELL R820, 4 x Intel Xeon E5-4650 Processors with 8 Cores each at 2,70GHz frequency and 8 X 32GB RAM memory. As for disks, we used a virtual partition of 4 SATA disks RAID5 for a total of 9TB, plus 2 SSD disks for running the OS (CentOS 6.6).

[6]Lucene version 4.0, a standard indexing and search engine library available at `http://lucene.apache.org/core/`

selection.

It is useful to recall the three settings employed in the experiments using the stopwords (as in Section 5.3.1):

**TF** Term Frequency;

**NIDF** Normalised Inverse Document Frequency;

**Indri** Standard list of stopwords.

Below are the three methods employed in the vocabulary selection part:

**Method x0** and baseline, profile of a user built with just his messages (Section 5.3.2);

**Method x1** profile of a user built considering his messages (x0) + the messages of the users writing one line after him (Section 5.3.3);

**Method x2** profile of a user built considering his messages (x0) + the messages of the users writing one line after him (x1) + messages of the users wiring one line before him (Section 5.3.3);.

For each of the three methods (x0, x1 and x2) of the vocabulary selection part we wanted to study the influence that specific and common terms have on the user identification. For this reason we analysed the different methods employing the user profiles at different "length". As described in Section 5.3.3, after computing the user profile with any of the methods x0, x1 or x2, we prioritise the terms contained in each profile based on their specificity. Being the terms ordered, we first considered the 5% most specific terms in a profile as representative of the user style, then the 10% of the most specific terms, then the 15% until the 100% of the terms in the profiles.

In Figure 5.8 we reported an example of this procedure, showing the different profile "length" corresponding to the different percentage of terms considered. For brevity, in the example we omitted the "central" profiles (from 45% to 75%), focusing on the profiles containing the most specific terms (5% to 35%) and those where the most common words are present (from 70% to 100%). This procedure aims at verifing our hypothesis that in conversational documents the most specific terms are more helpful than the common terms in the user identification and that this is more evident when the profile of an author is enriched with terms from his interlocutors (methods x1 and x2).

### 5.4.3 Evaluation functions

In literature [114], one of the most widely employed evaluating functions of the problem of author identification is *average accuracy*. However (average) accuracy, defined as the number of corrected classified authors over the total number of authors in the test set, was found not completely appropriate in this context. In fact, it allows to evaluate

```
 5%: [ringostarr, ucit, folders, shared, drwxrwsrwt]

10%: [ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx, oposite]

15%: [ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx, oposite, windowsxp, strange, hmmm,
    nick, protected]

20%: [ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx, oposite, windowsxp, strange, hmmm,
    nick, protected, pc, sep, stopped, asks, ideas]

25%: [ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx, oposite, windowsxp, strange, hmmm,
    nick, protected, pc, sep, stopped, asks, ideas, showing, connections, smb, appear, group]

30%: [ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx, oposite, windowsxp, strange, hmmm,
    nick, protected, pc, sep, stopped, asks, ideas, showing, connections, smb, appear, group, cant, iptables, remove, isnt,
    share, anybody]

35%: [ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx, oposite, windowsxp, strange, hmmm,
    nick, protected, pc, sep, stopped, asks, ideas, showing, connections, smb, appear, group, cant, iptables, remove, isnt,
    share, anybody, computer, via, box, samba, having]

...


75%: [ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx, oposite, windowsxp, strange, hmmm,
    nick, protected, pc, sep, stopped, asks, ideas, showing, connections, smb, appear, group, cant, iptables, remove, isnt,
    share, anybody, computer, via, box, samba, having, under, doesnt, write, network, few, directory, test, conf, open, o, log,
    possible, things, doing, seems, idea, wrong, he, trying, still, same, help, other, them, user, hi, lol, by, something, now,
    yes, windows,  want, linux, problem, am, was, set, no, this, just, has, what]

80%: [ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx, oposite, windowsxp, strange, hmmm,
    nick, protected, pc, sep, stopped, asks, ideas, showing, connections, smb, appear, group, cant, iptables, remove, isnt,
    share, anybody, computer, via, box, samba, having, under, doesnt, write, network, few, directory, test, conf, open, o, log,
    possible, things, doing, seems, idea, wrong, he, trying, still, same, help, other, them, user, hi, lol, by, something, now,
    yes, windows,  want, linux, problem, am, was, set, no, this, just, has, what, file, not, t, for, s]

85%: [ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx, oposite, windowsxp, strange, hmmm,
    nick, protected, pc, sep, stopped, asks, ideas, showing, connections, smb, appear, group, cant, iptables, remove, isnt,
    share, anybody, computer, via, box, samba, having, under, doesnt, write, network, few, directory, test, conf, open, o, log,
    possible, things, doing, seems, idea, wrong, he, trying, still, same, help, other, them, user, hi, lol, by, something, now,
    yes, windows,  want, linux, problem, am, was, set, no, this, just, has, what, file, not, t, for, s, of, see, me, you, any]

90%: [ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx, oposite, windowsxp, strange, hmmm,
    nick, protected, pc, sep, stopped, asks, ideas, showing, connections, smb, appear, group, cant, iptables, remove, isnt,
    share, anybody, computer, via, box, samba, having, under, doesnt, write, network, few, directory, test, conf, open, o, log,
    possible, things, doing, seems, idea, wrong, he, trying, still, same, help, other, them, user, hi, lol, by, something, now,
    yes, windows,  want, linux, problem, am, was, set, no, this, just, has, what, file, not, t, for, s, of, see, me, you, any,
    as, a, have, one, my]

95%: [ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx, oposite, windowsxp, strange, hmmm,
    nick, protected, pc, sep, stopped, asks, ideas, showing, connections, smb, appear, group, cant, iptables, remove, isnt,
    share, anybody, computer, via, box, samba, having, under, doesnt, write, network, few, directory, test, conf, open, o, log,
    possible, things, doing, seems, idea, wrong, he, trying, still, same, help, other, them, user, hi, lol, by, something, now,
    yes, windows,  want, linux, problem, am, was, set, no, this, just, has, what, file, not, t, for, s, of, see, me, you, any,
    as, a, have, one, my, and, is, on, can, that, but]

100%:[ringostarr, ucit, folders, shared, drwxrwsrwt, bodyrider, dr, comuter, authorised, rwx, oposite, windowsxp, strange, hmmm,
    nick, protected, pc, sep, stopped, asks, ideas, showing, connections, smb, appear, group, cant, iptables, remove, isnt,
    share, anybody, computer, via, box, samba, having, under, doesnt, write, network, few, directory, test, conf, open, o, log,
    possible, things, doing, seems, idea, wrong, he, trying, still, same, help, other, them, user, hi, lol, by, something, now,
    yes, windows, want, linux, problem, am, was, set, no, this, just, has, what, file, not, t, for, s, of, see, me, you, any,
    as, a, have, one, my, and, is, on, can, that, but, it, in, to, i, the]
```

*Figure 5.8.* Example of a profile at different vocabulary lengths (in percentage), with terms ordered from the most specific (topical) to the most common (stopwords) ones.

well the traditional authorship identification problems, where the number of authors to identify is limited (around 20 and generally no more than 50) and the authors profile length is quite long (around 250 terms on average). Under these conditions, finding the author as the one with the minimum distance from a short list of possible candi-

dates was demonstrated to work quite well. The same does not apply to the problem of authorship identification where the users are hundreds or thousands and their profile length quite limited (around 50 terms). Having hundreds of authors with a limited vocabulary can lead to the problem of having a very low accuracy, due to the fact that an author might not be found at the very first place (the one with minimum distance). However, it might be located at the second place or among the first 10 or 20. When having hundreds of authors this is a results that might be as good as finding the author at the first place. For example, in the case of suspect prioritisation within the framework of law enforcement a police agent might decide to manually inspect the top-10 retrieved suspects rather than just the first suspect. For these reasons, to evaluate the quality of the methods under evaluation, and in particular to test x1 and x2 against the baseline x0, we employed the Mean Reciprocal Rank.

### Mean Reciprocal Rank

Mean Reciprocal Rank (MRR) as defined for evaluating the problem of Question Answering in TREC-8 has similar characteristics to the problem of identifying authors. First, there is only one relevant answer (or author) among the retrieved ones. Then it might also happen that an answer (or author) is not found. In the case of authorship identification this happens when the author profile is so short that no inference on similarity can be assessed. In the original formulation of MRR it is said that "an individual question received a score equal to the reciprocal of the rank at which the first correct response or 0 if none of the five responses contained the correct answer. The score for a submission was then the mean of the individual question's reciprocal rank" [130]. The only difference between the original and the authorship attribution setting is that in the latter the maximum number of returned answers is not bounded to only five but might take at most the whole list of authors in the testing set (some hundreds up to thousands, depending on the collections) and that we never assign 0 to any match but at least 1 over the number of authors in the testing set. This is due to the fact that we impose that all the users in the testing set have also a corresponding profile in the training set (closed class problem), while we do not allow profiles in the test set to be labelled as unknown (open class problem).

To summarise MRR with a formula:

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i} \tag{5.10}$$

where $Q$ is the set of authors in the testing set and $\mathrm{rank}_i$ the position in the ranking of the matching author from the training set.

From the formula, it is also easy to see that the accuracy corresponds to the MRR in the very optimistic case in which all the $Q$ users in the testing set have $\mathrm{rank}_1$. In other words, this only happens when we have 100% accuracy. In all other cases MRR

would always have a greater value than the average accuracy, due to the fact that it always assigns a non-zero value to a testing profile, while accuracy assigns a 0-score if a testing profile has not minimum distance with the correct one.

### 5.4.4 Statistical Significance

After setting the MRR as evaluating function for the experiments, we need a way to assess wether our proposed methods x1 and x2 are performing better than the baseline method x0 or if among the three methods there are no significative differences, being equivalent. We designed our experiments to have a predefined set of users (testing set) that was employed in every experiment. For this reason, we could perform a *paired test*. From every experiment we considered only one independent variable at a time, which was the MRR score, and we compared for statistical significance only the results of two methods (or two different experiments) at a time. Having these characteristics and being our data non-parametric[7], the most appropriate test to assess statistical significance resulted to be the *Wilcoxon test*. This was established following the book of Field and Hole [36]. We used the default package included in the R statistical environment [101] for computing the Wilcoxon (paired) statistical test.

For simplicity and readability, in reporting the results of our experiments in Section 5.5 we just indicated the measured mean value of MRR for each experiment and wether there is statistical significance difference among two experimental results, without explicitly mention the exact p-value or exact value of Wilcoxon test. We tested all results with a significance level of $p < 0.05$. Detailed tables with all the exact values and p-values are available in Appendix C.

## 5.5 Experimental Results

In this section we will present the result of the experimental evaluation. In Section 5.5.1 we will first present the experimental results obtained building the users profiles with different stopwords lists. In Section 5.5.2 we will present, instead, the results obtained building the user profiles by first ordering the terms in the user profiles by their importance and then using a different percentage (from 5% to 100%) of the vocabulary to compute the distance between profiles.

### 5.5.1 Stopwords

The first experiments conducted are those reproducing the state-of-the-art approaches of authorship identification based on stopwords. In Table 5.2 we report the results of the first experiments, divided into two groups: in the first group we reported the

[7] Our data, in fact, were not normally distributed in any experiments, as we could test with *Shapiro-Wilk* tests [36].

results obtained employing only 20 users for all the collections (Table 5.2a), while in the second group (Table 5.2b) the results obtained employing a larger number of users, up to thousands (the detailed numbers were presented in Table 5.1).

### 20 Users

As expected, employing a limited number of users (20) on collections of traditional documents led to almost perfect results: experiments on collections of newspaper articles from AP, La Stampa and Glasgow Herald for two settings (TF and NIDF) with method x0 returned 100% MRR. It is useful to recall that TF and NIDF are lists of stopwords computed based on the most frequent terms within those of the authors' profiles, while Indri is a pre-computed list of generic stopwords. This explains the lower values of Indri with respect to the other two settings. As for the classifiers, KLD seemed to perform better than $\chi^2$, while Delta is not really effective, with less than 20% MRR. As a last remark, it should be said that for these collections we could not apply the two proposed methods x1 and x2 because they are not conversational. For the conversational documents we observed a similar behaviour, although with different values of MRR. With all the three methods (x0, x1 and x2) the settings TF and INDF returned higher values than the Indri one and Delta classifier was still performing worse than KLD and $\chi^2$, with few exceptions for the Freenode collection. However, for the two collections of IRC logs, Freenode and Krijn, the best results were around 80% MRR. Our method x1 performed statistically significantly better than the baseline x0 and both methods x1 and x2 applied on Krijn were comparable to x0, not being statistically different from x0 itself. Twitter, on the other hand, manifested a different behaviour. The results were much worse than the other two collections, Freenode and Krijn, and all the three methods x0, x1 and x2 returned the same results of 46% MRR. This latter behaviour can be due to a limited sample of users not sharing conversations, therefore not profiting from the improvements of the methods x1 and x2. The overall worst performance can be explained by the nature of the collection itself, which contains documents that are at least 25% shorter than the Freenode or Krijn ones (see Table 5.1b).

### Hundreds of Users

When repeating the experiments for a larger set of users, in the order of some hundreds up to thousands, we immediately noticed a degradation of performances, with MRR decreasing by more than 50%, reaching values between 39% and 45% for collections of newspaper articles (AP, La Stampa and Glasgow Herald) and less than 10% for collections of conversational documents (Freenode, Krijn, Twitter). For Twitter in particular, it was almost impossible to correctly identify any users. Moreover, the employment of methods x1 and x2 did not help in improving the MRR of the users identification. These experiments are novel in the literature: to the best of our knowledge nobody explored sets of hundreds of users for different types of documents like the traditional

| Collection | Class. | method x0 | | | method x1 | | | methodx2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | TF | NIDF | Indri | TF | NIDF | Indri | TF | NIDF | Indri |
| AP | KLD | **1.00** | **1.00** | 0.99 | - | - | - | - | - | - |
| | $\chi^2$ | 0.84 | 0.84 | 0.75 | - | - | - | - | - | - |
| | Delta | 0.18 | 0.18 | 0.19 | - | - | - | - | - | - |
| La Stampa | KLD | **1.00** | **1.00** | 0.96 | - | - | - | - | - | - |
| | $\chi^2$ | 0.89 | 0.89 | 0.68 | - | - | - | - | - | - |
| | Delta | 0.18 | 0.18 | 0.19 | - | - | - | - | - | - |
| Glasgow Herald | KLD | **1.00** | **1.00** | 0.98 | - | - | - | - | - | - |
| | $\chi^2$ | 0.92 | 0.92 | 0.86 | - | - | - | - | - | - |
| | Delta | 0.18 | 0.18 | 0.18 | - | - | - | - | - | - |
| Freenode | KLD | 0.72 | 0.72 | 0.56 | **0.80**† | **0.80**† | 0.57† | 0.79† | 0.79† | 0.55 |
| | $\chi^2$ | 0.51 | 0.51 | 0.26 | 0.53 | 0.53 | 0.38 | 0.53 | 0.53 | 0.38 |
| | Delta | 0.57 | 0.57 | 0.49 | 0.49 | 0.49 | 0.37 | 0.52 | 0.52 | 0.39 |
| Krijn | KLD | **0.81** | **0.81** | 0.63 | 0.79 | 0.79 | 0.70 | 0.80 | 0.80 | 0.69 |
| | $\chi^2$ | 0.59 | 0.59 | 0.41 | 0.62 | 0.62 | 0.47 | 0.65 | 0.65 | 0.44 |
| | Delta | 0.20 | 0.20 | 0.19 | 0.23 | 0.23 | 0.20 | 0.23 | 0.23 | 0.21 |
| Twitter | KLD | **0.46** | **0.46** | 0.43 | **0.46** | **0.46** | 0.41 | **0.46** | **0.46** | 0.42 |
| | $\chi^2$ | 0.39 | 0.39 | 0.38 | 0.41 | 0.41 | 0.41 | 0.42 | 0.42 | 0.37 |
| | Delta | 0.22 | 0.22 | 0.19 | 0.20 | 0.20 | 0.16 | 0.23 | 0.23 | 0.19 |

*(a)* 20 selected users.

| Collection | Classifier | x0 | | | x1 | | | x2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | TF | NIDF | Indri | TF | NIDF | Indri | TF | NIDF | Indri |
| AP | KLD | **0.42** | **0.42** | 0.32 | - | - | - | - | - | - |
| | $\chi^2$ | 0.22 | 0.22 | 0.09 | - | - | - | - | - | - |
| | Delta | 0.04 | 0.04 | 0.02 | - | - | - | - | - | - |
| La Stampa | KLD | 0.34 | 0.34 | **0.39** | - | - | - | - | - | - |
| | $\chi^2$ | 0.19 | 0.19 | 0.11 | - | - | - | - | - | - |
| | Delta | 0.04 | 0.04 | 0.02 | - | - | - | - | - | - |
| Glasgow Herald | KLD | **0.45** | **0.45** | 0.41 | - | - | - | - | - | - |
| | $\chi^2$ | 0.24 | 0.24 | 0.15 | - | - | - | - | - | - |
| | Delta | 0.06 | 0.06 | 0.03 | - | - | - | - | - | - |
| Freenode | KLD | **0.08** | **0.08** | 0.06 | 0.06 | 0.06 | 0.05 | 0.05 | 0.05 | 0.04 |
| | $\chi^2$ | 0.03 | 0.03 | 0.05 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.02 |
| | Delta | 0.07 | 0.07 | 0.05 | 0.05 | 0.05 | 0.04 | 0.05 | 0.05 | 0.04 |
| Krijn | KLD | 0.06 | 0.06 | 0.04 | 0.03 | 0.03 | 0.02 | 0.03 | 0.03 | 0.02 |
| | $\chi^2$ | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| | Delta | **0.07** | **0.07** | 0.04 | 0.05 | 0.05 | 0.03 | 0.05 | 0.05 | 0.03 |
| Twitter | KLD | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | $\chi^2$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.01** |
| | Delta | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

*(b)* Hundreds of users.

*Table 5.2.* Results using the different stopwords strategies for building users profiles. *TF*: stopwords list based on collection term frequencies; *NIDF*: stopwords list based on normalised inverse document frequency; *Indri*: stopwords from a standard pre-computed list. In **bold** the best results for each collection (1 is the best possible). †: statistical significance of methods x1 or x2 w.r.t. method x0.

and the conversational ones to understand the impact of stopwords in performing user identification. The relatively good performance observed for collections of newspaper articles is due to the length and style of the documents themselves, which are relatively long and written in a structured manner, following grammatical and syntactical rules. This allowed to identify the writing style of the users, although with some additional noise, compared to the small portion of 20 users.

Conversational documents, on the other hand, are written in a sort of colloquial and speaking-like manner, with syntactical construction of the sentences most of the time arbitrary, thus the words employed are written in the wrong form and sometimes also not existing in the standard dictionary (as also verified in Chapter 4). For this reason we believe that for conversational documents different strategies need to be employed. In Section 5.5.2 we display the results obtained under different conditions, employing these different strategies.

A last remark on the evaluation of the presented results: in this study the goal was not to measure the performance of the different classifiers, therefore we did not assess the statistical significance among the different classifiers (KLD, $\chi^2$, Delta) but only among the different methods (x0, x1, x2).

### 5.5.2 Vocabulary Selection by Percentage

We first considered the approach based on *Simple Author Vocabulary* (method x0, Section 5.3.2) as vocabulary selection strategy, then the other two based on profile expansion through *Interlocutors Influence* and terms reordering (method x1 and method x2, Section 5.3.3). These three methods are employed in this Section with a complete study on the influence of each portion of the authors vocabulary (topical terms, common terms) on the authors identification.

#### 20 Users

When analysing the results for 20 users employing the vocabulary selection strategies, we can derive similar conclusions for the stopwords strategy: on the collections of traditional documents (Figure 5.9, Figure 5.10, Figure 5.11) the identification of users is almost perfect. The KLD classifiers were able to return 100% MRR even with few portions of the users vocabulary (already with only 5% of profiles length) for AP, La Stampa and Glasgow Herald, not necessarily considering stopwords, contained in the last portion of users vocabulary (80% to 100% profile length). This is the case for the $\chi^2$ classifiers, which returned 100% MRR only when introducing the last 5% of terms, the stopwords. The Delta classifier, on the other hand, was performing worse than the other two, not reaching 40% of MRR. This was observed also in the previous experiments, in which we employed stopwords only (Section 5.5.1).

The same behaviour for Delta could be observed for the collections of conversational documents: with one exception only, Delta never went above 50% MRR (Figure

5.12, Figure 5.13, Figure 5.14), while the other two classifiers, KLD and $\chi^2$, were able to obtain MRR above 80%, starting with only 5% of users vocabulary for Freenode and Krijn. This last result is interesting for several reasons. First, in the previous experiments with stopwords, 81% of MRR was the best result among all the conversational collections, in particular for collection Krijn using KLD as classifier and the baseline method x0. Nevertheless, in the current experimental settings the same result (80% MRR) represents only the worst result for the same collection Krijn, classifier KLD and baseline method x0, that is up to almost 100% MRR with method x1. Second, if we focuse on Krijn collection (Figure 5.13), we will notice that the best results were obtained with the classifier KLD (regardless of the scoring approaches TF-IDF or KLD) with the two methods x1 and x2, that are performing as well as, or significantly better, than the baseline x0. Third, the scoring approach TF-IDF returned the best results with only 20% of the profiles length. This is important to consider if interested in size of memory or speed of computation: the lower the number of terms, the faster the computation and the lower the memory needed to store the data. Although the data related to the profiles of 20 users are limited, this could become an issue when considering hundreds or thousands of users, as in the next set of experiments.

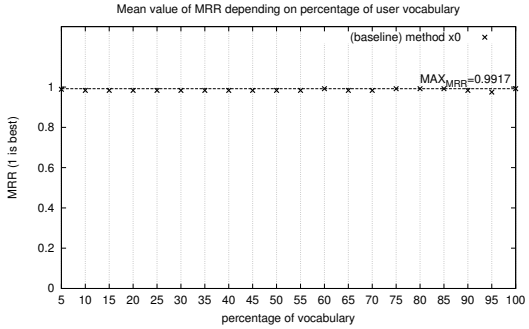
These observations are important because they are a first confirmation of our initial hypothesis: we assumed that strategies based on stopwords would not work well for conversational documents and, instead, specific words obtained through profile expansion and vocabulary analysis would work better for identifying users in such collections. The role of stopwords seems to be beneficial under particular conditions, for example when employing the full profile of users with a particular classifier: in both the collections Krijn and Freenode, we could observe that for the classifier $\chi^2$ the MRR increased dramatically with the addition of the last 5% of the terms for user profiles, that is the stopwords. A better confirmation of our hypothesis can be found in the results for the classifier KLD on the collection Freenode (Figure 5.12). As it was for Krijn, also in Freenode both methods x1 and x2 are significantly better than the method x0. Moreover they both allowed for almost 100% MRR when employed with the KLD classifier, already at low level of profile length (around 30%). If we compare these results with the ones obtained with the stopwords only (Table 5.2a), we can easily see an improvement of almost 20% for Freenode and Krijn with the same classifier KLD. This is not only due to the usage of more "evidence" (more terms), like the full profile length, but also to better terms, obtained with methods x1 and x2. The best results, in fact, were obtained with either method x1 or x2 being significantly better than the baseline x0 with profiles expanded of about 30% better terms.

Different observations can be done for Twitter: although for this collection methods x1 and x2 do not beat the baseline x0, the use of specific words allowed to reach a MRR of more than 75%, which is much better compared to the 46% obtained with the stopwords approach. Similarly to Freenode and Krijn, also for the Twitter collection KLD and $\chi^2$ performed better than Delta but with fewer difference than in the other two

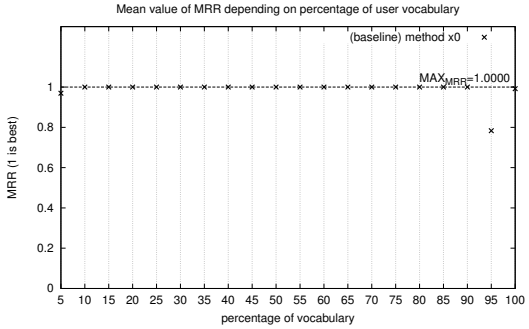
collections. However, we could observe these results already at 25%-30% of profiles length. In some cases these results were better than those obtained employing the 100% of profiles length. This is something interesting, because the specific terms that we hypothesised to work better for conversational documents lie in the first part of users profiles. We noticed a similar behaviour also in the next experiments, with hundreds and thousands of users.

### Hundreds of Users

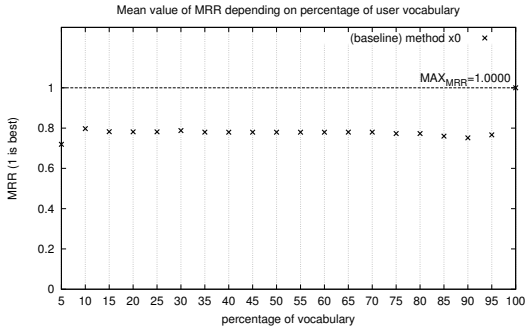
With hundreds and thousands of users, the MRR decreased for collections of newspaper articles (Figure 5.15, Figure 5.16, Figure 5.17), as it did for the stopwords: from 100% MRR obtained with 20 authors only, to values around 60% only. For Associated Press the best MRR is 67%, for La Stampa 58% and for the Glasgow Herald 60%. All these values were obtained employing KLD both as classifier and as methods for scoring terms within the authors profiles and making use of full author profiles (that is, 100% of vocabulary). The usage of the full vocabulary for each author might seem quite heavy, however there are other two aspects to consider: first, in the last part of the profiles (in particular in the last 5%) lie the stopwords that are known to work well for user identification of traditional collections. This can be observed also with the KLD and $\chi^2$ classifiers, for all the three collections. Second, the results employing the first 5%-10% of terms with the KLD classifier and KLD scoring are similar to the best ones obtained with the stopwords only (Table 5.2b). This is interesting because it suggests a reasonable and close relationship between the authors of the newspaper articles and the topics contained in the articles. The study of this relationship is out of the scope of our work, that is instead on conversational documents. However it might be considered as a component for the future work. This observation can be verified also for the Delta classifier and it is the only remark to be done for this classifier, that is performing incredibly lower than the other ones. Similar observations can be done for collections of conversation documents: when considering hundreds of users the MRR dropped from almost 100% for Freenode and Krijn to 24% and 17% respectively (Figures 5.18 and 5.19), while Twitter dropped from 76% to 12% (Figure 5.20). This is clearly better than what was achieved with the stopwords only (as in Table 5.2b: 8%, 7% and 0%) but with some remarks. In fact, the best performance for Freenode and Krijn were obtained with the method x0 and 100% of the profiles vocabulary. While this behaviour was not expected, we could observe the desired behaviours immediately after: methods x1 or x2 to perform as second or third best at lower level of profiles vocabulary. For Freenode, in fact, 23% of MRR was achieved with method x2 and 30% of vocabulary only, while for Krijn 14% and 12% were obtained with method x1 at 100% and 30% of vocabulary. Despite that, for Twitter the best results were achieved with method x2 at 100% of vocabulary (12% MRR) and 20% of vocabulary (10.5% MRR), while the best performance of method x0 is for 100% vocabulary at 9.9% MRR. In Table 5.3 we reported a summary of this information.



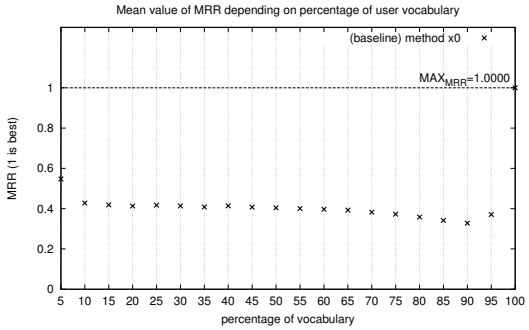

*(a)* Classifier KLD, terms ordering with KLD



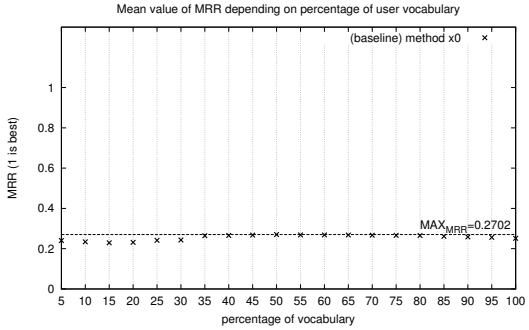

*(b)* Classifier KLD, terms ordering with TF-IDF



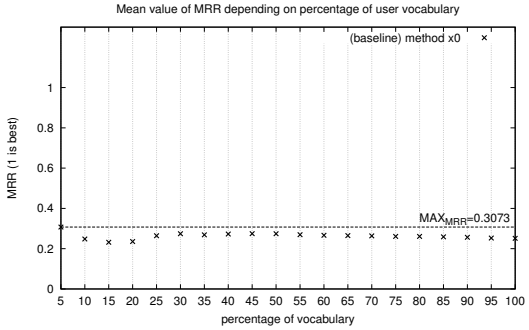

*(c)* Classifier $\chi^2$, terms ordering with KLD




*(d)* Classifier $\chi^2$, terms ordering with TF-IDF

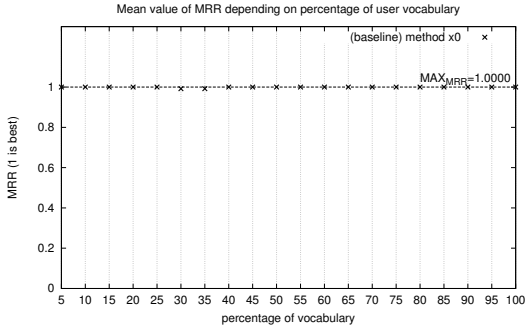



*(e)* Classifier Delta, terms ordering with KLD

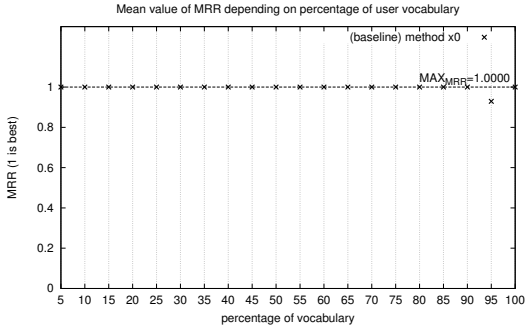



*(f)* Classifier Delta, terms ordering with TF-IDF

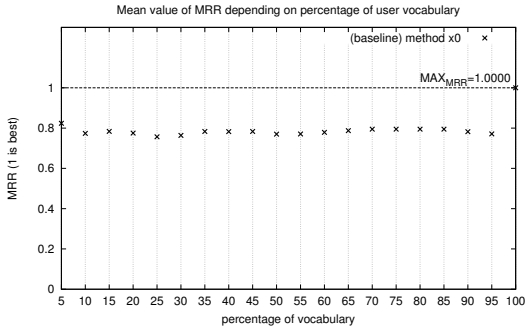
*Figure 5.9.* Experimental results employing the vocabulary selection strategy on 20 users for the Associated Press (AP) collection.



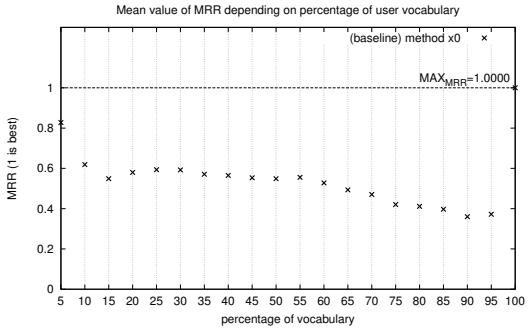

*(a)* Classifier KLD, terms ordering with KLD



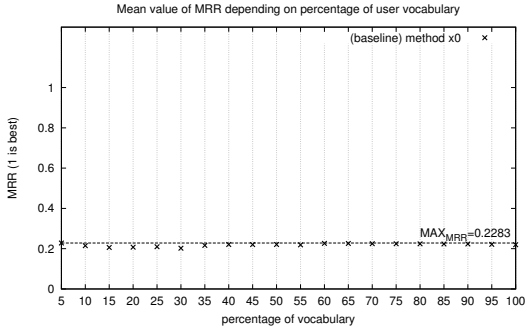

*(b)* Classifier KLD, terms ordering with TF-IDF



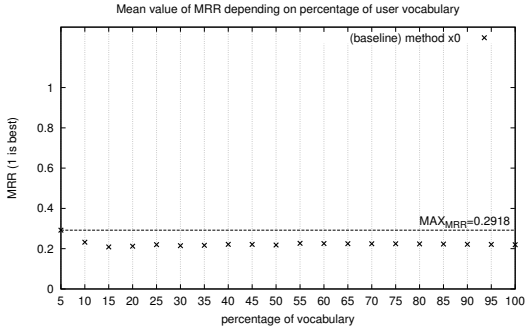

*(c)* Classifier $\chi^2$, terms ordering with KLD




*(d)* Classifier $\chi^2$, terms ordering with TF-IDF

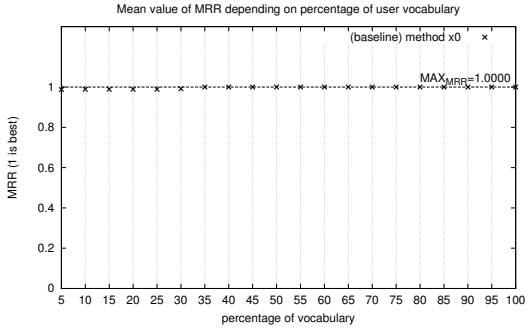



*(e)* Classifier Delta, terms ordering with KLD

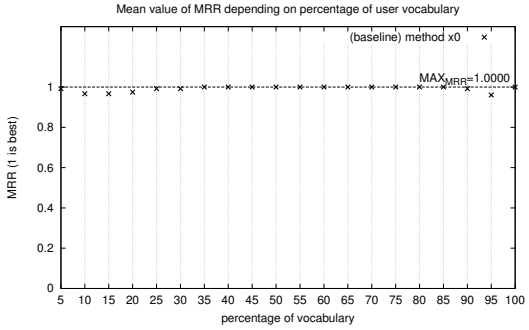



*(f)* Classifier Delta, terms ordering with TF-IDF

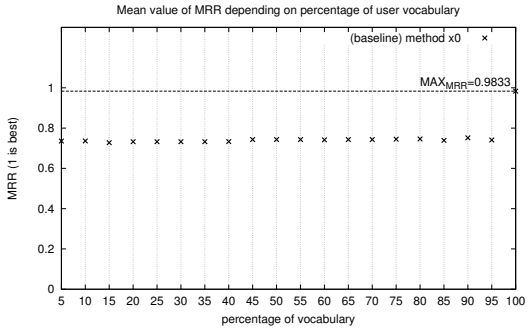
*Figure 5.10.* Experimental results employing the vocabulary selection strategy on 20 users for La Stampa collection.



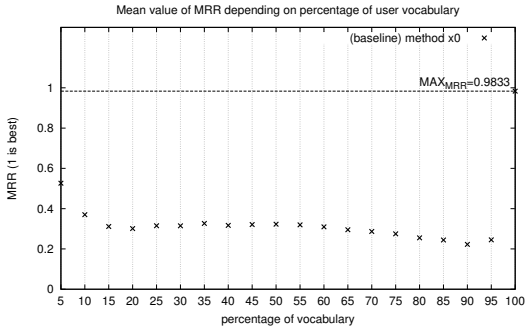

*(a)* Classifier KLD, terms ordering with KLD



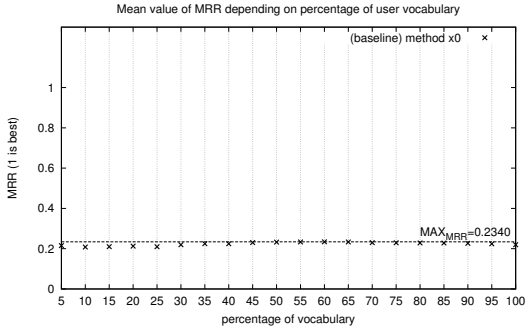

*(b)* Classifier KLD, terms ordering with TF-IDF



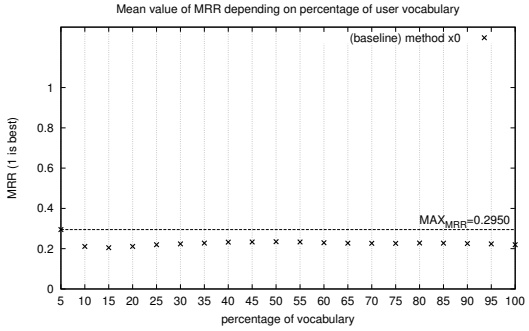

*(c)* Classifier $\chi^2$, terms ordering with KLD




*(d)* Classifier $\chi^2$, terms ordering with TF-IDF

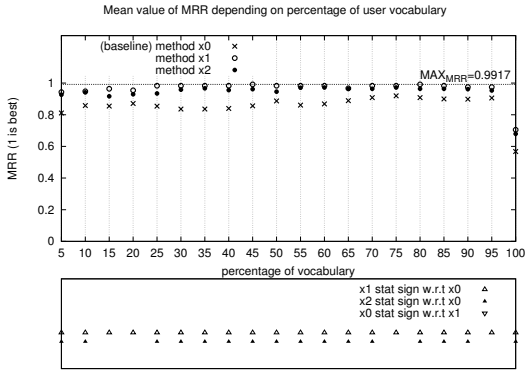



*(e)* Classifier Delta, terms ordering with KLD

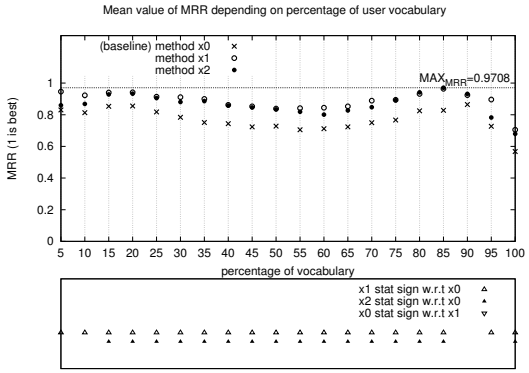



*(f)* Classifier Delta, terms ordering with TF-IDF

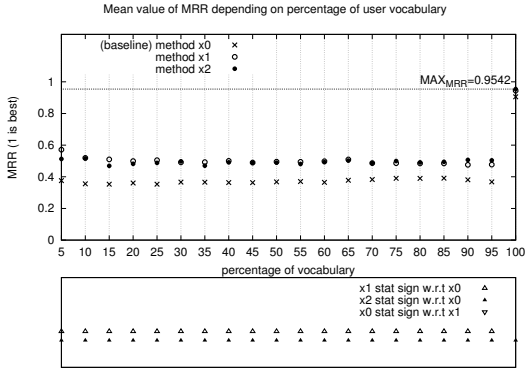
*Figure 5.11.* Experimental results employing the vocabulary selection strategy on 20 users for Glasgow Herald collection.



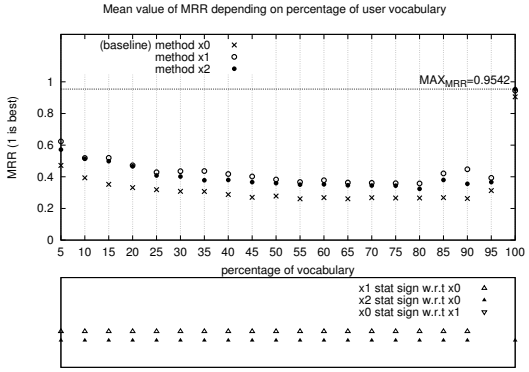

*(a)* Classifier KLD, terms ordering with KLD



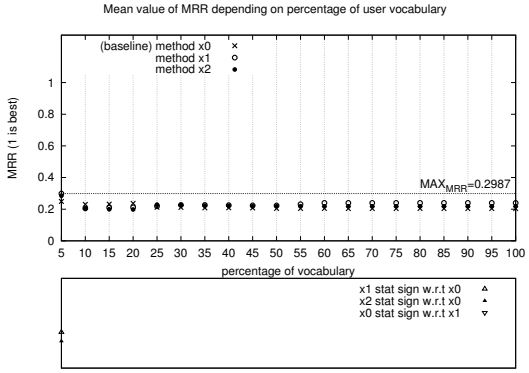

*(b)* Classifier KLD, terms ordering with TF-IDF



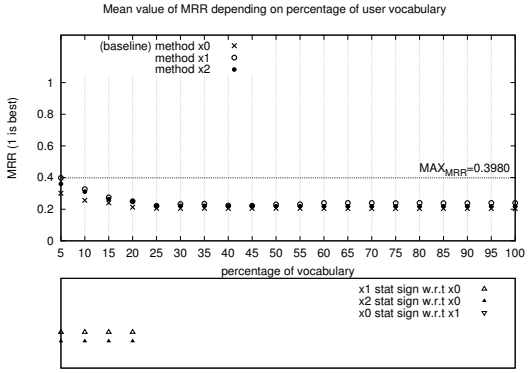

*(c)* Classifier $\chi^2$, terms ordering with KLD




*(d)* Classifier $\chi^2$, terms ordering with TF-IDF

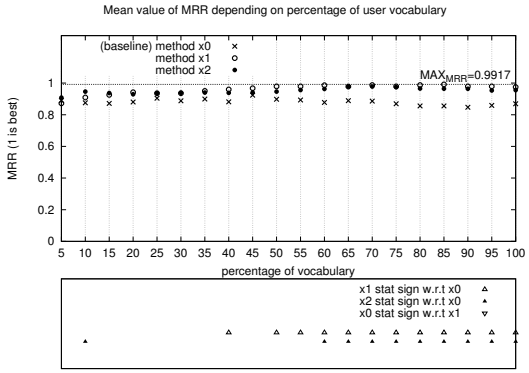



*(e)* Classifier Delta, terms ordering with KLD

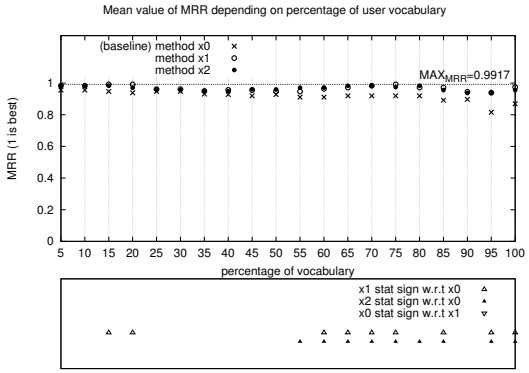



*(f)* Classifier Delta, terms ordering with TF-IDF

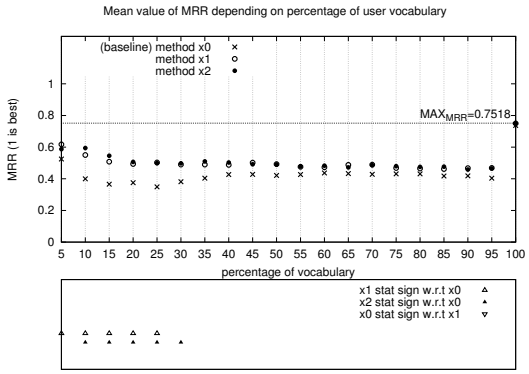
*Figure 5.12.* Experimental results employing the vocabulary selection strategy on 20 users for the Freenode collection.



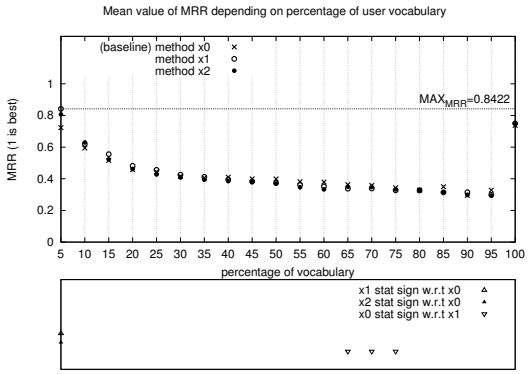

*(a)* Classifier KLD, terms ordering with KLD



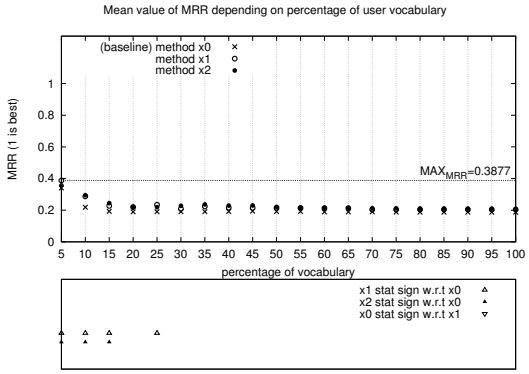

*(b)* Classifier KLD, terms ordering with TF-IDF



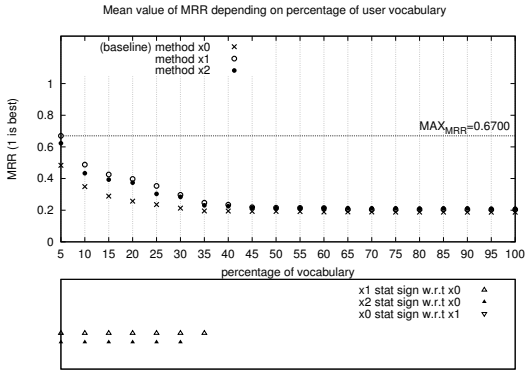

*(c)* Classifier $\chi^2$, terms ordering with KLD




*(d)* Classifier $\chi^2$, terms ordering with TF-IDF

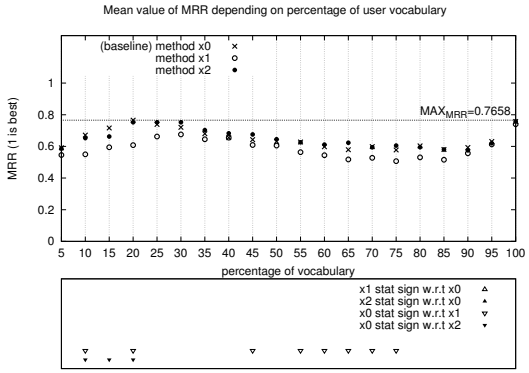



*(e)* Classifier Delta, terms ordering with KLD

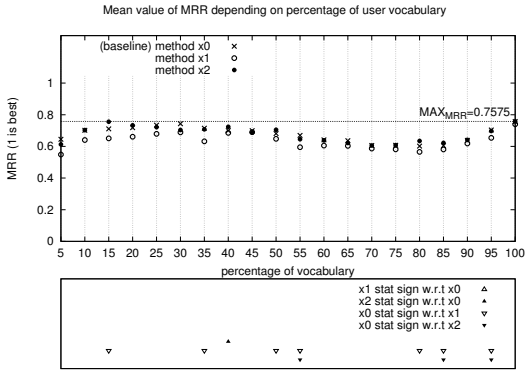



*(f)* Classifier Delta, terms ordering with TF-IDF

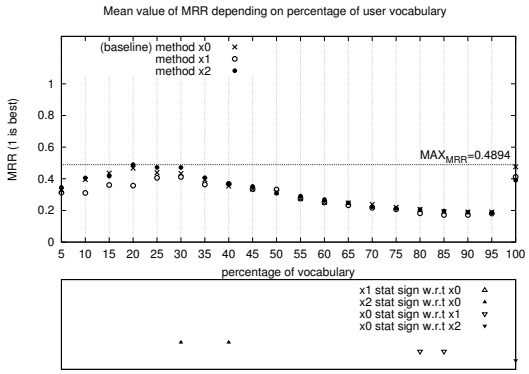
*Figure 5.13.* Experimental results employing the vocabulary selection strategy on 20 users for the Krijn collection.



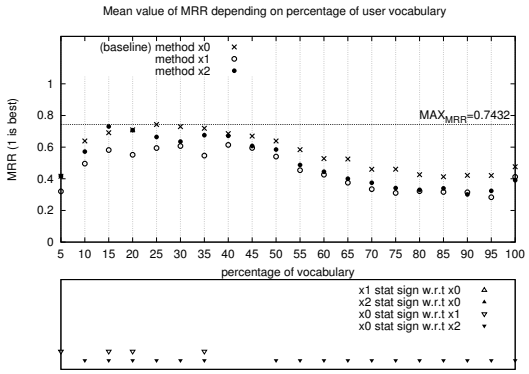

*(a)* Classifier KLD, terms ordering with KLD



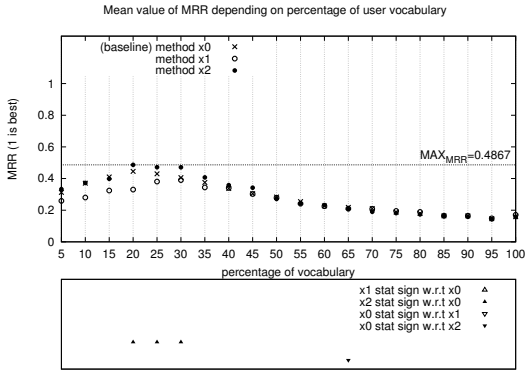

*(b)* Classifier KLD, terms ordering with TF-IDF



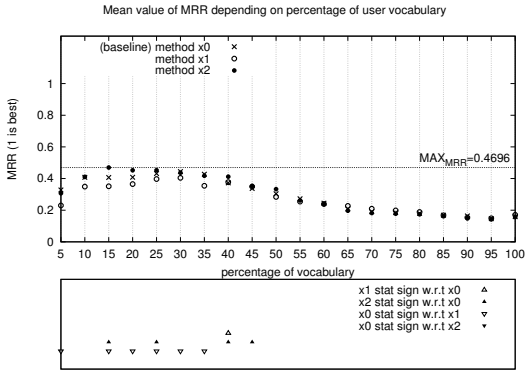

*(c)* Classifier $\chi^2$, terms ordering with KLD




*(d)* Classifier $\chi^2$, terms ordering with TF-IDF




*(e)* Classifier Delta, terms ordering with KLD




*(f)* Classifier Delta, terms ordering with TF-IDF

*Figure 5.14.* Experimental results employing the vocabulary selection strategy on 20 users for the Twitter collection.

Although the full employment of the vocabulary is a clear signal that it provides a great contribution to the best performance of the different methods x0, x1, x2, the fact that good results could be observed already at 20% or 30% of profile length and for methods x1 and x2 it is an indication of still space of improvement in this area. This is also a confirmation that our initial hypothesis might still be valid, with some additional work. An example of this could be the combination of terms in the first (specific words) and last level of percentage (common words) only or the opposite analysis at different percentages, from common to specific words, to observe the performance of the different classifiers. Moreover, if we think of some possible applications where the employment of 100% of the vocabulary for all the profiles of all the users is not possible, a small loss of precision in the MRR (e.g. from 24% to 23% in case of Krijn) could be acceptable, with the advantage of profiles shorter of about 30% (see Tables 5.4, 5.5 and 5.6) at lower vocabulary percentage (from 100% to 30%).

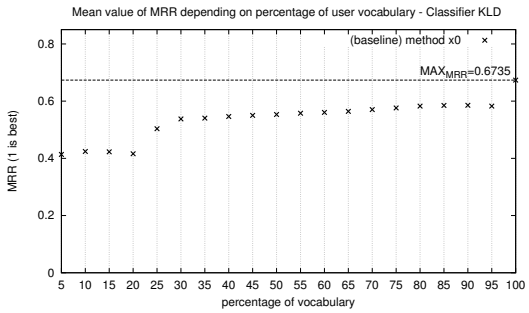
Finally, as for the classifier, it seems that KLD was working better for Freenode and Krijn, while $\chi^2$ was providing better results for Twitter. Delta, on the other hand, was not performing as well as KLD and $\chi^2$ as in the other experiments. As for the terms ordering strategy, KLD seems to work better in combination with the same KLD classifier (although unrelated and applied at different times), while TF-ID had a better employment with $\chi^2$.

| Collection | Best results method x0 | | | Best results interlocutors approach | | | | |
|---|---|---|---|---|---|---|---|---|
| | value | % | class. | method | value | % | class. | scoring |
| Freenode | **0.2391**† | 100 | KLD | x2 | 0.2336 | 30 | $\chi^2$ | TF-IDF |
| Krijn | **0.1686**† | 100 | KLD | x1 | 0.1444 | 100 | KLD | KLD |
| Twitter | 0.0999 | 100 | $\chi^2$ | x2 | **0.1215**† | 100 | $\chi^2$ | TF-IDF |

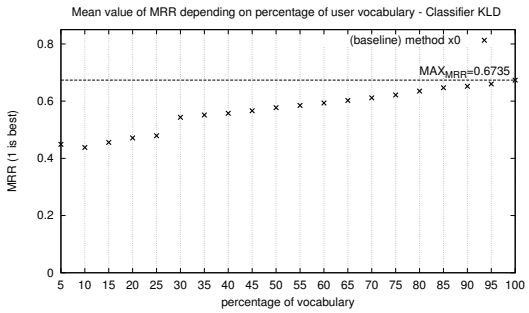
*Table 5.3.* Summary of the main results for the vocabulary selection strategy with hundreds of users on the conversations documents. For each collection we report the best results achieved by method x0 and the second best by either method x1 or x2. †: indicates the result is statistically significant w.r.t. the corresponding one in the other column.

### Profiles Length

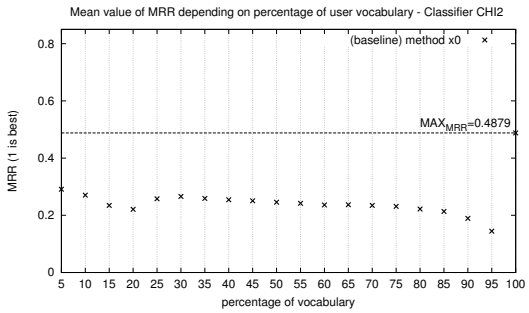
To better understand the impact of the different vocabulary percentages on the length of user profiles, we illustrate on Table 5.4, Table 5.5 and Table 5.6 the average profiles length for each collection (Freenode, Krijn and Twitter), for the different methods (x0, x1 and x2) and experimental settings (20 users and hundreds of users). This values differs from those of Table 5.1 because they are obtained after applying the pre-processing of Section 5.4.2, that eliminates small profiles and those without interlocutors. The scoring methods TF-IDF and KLD simply rearrange the terms within each



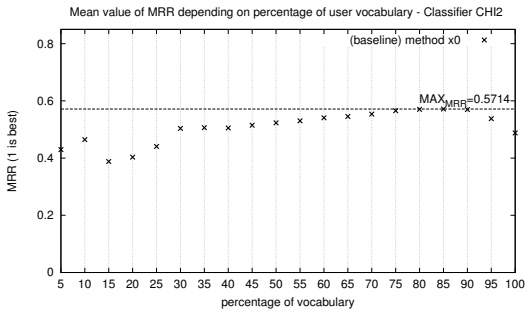

*(a)* Classifier KLD, terms ordering with KLD



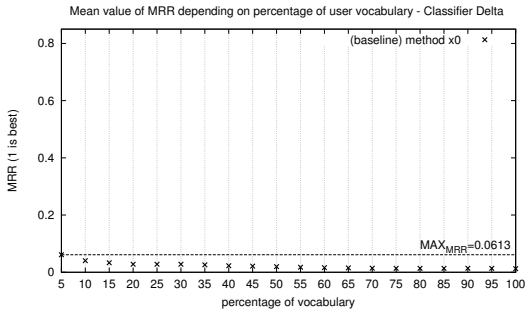

*(b)* Classifier KLD, terms ordering with TF-IDF



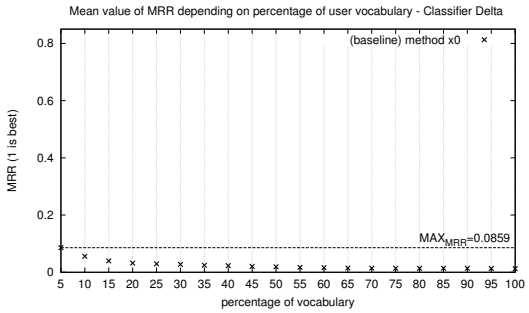

*(c)* Classifier $\chi^2$, terms ordering with KLD




*(d)* Classifier $\chi^2$, terms ordering with TF-IDF

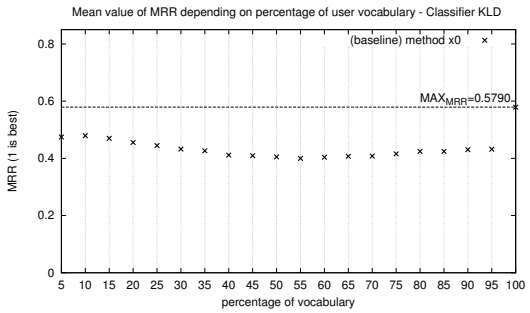



*(e)* Classifier Delta, terms ordering with KLD

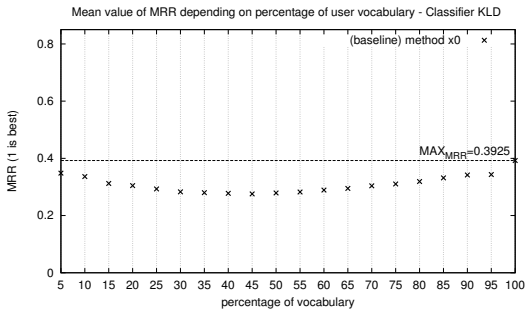



*(f)* Classifier Delta, terms ordering with TF-IDF

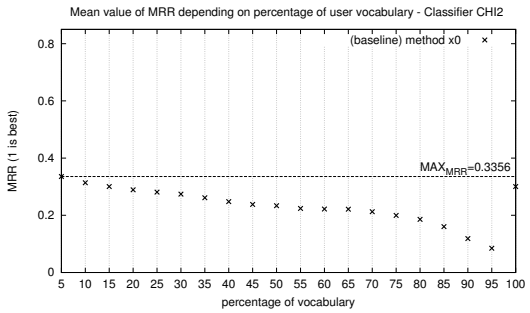
*Figure 5.15.* Experimental results employing the vocabulary selection strategy on hundreds of users for the Associated Press (AP) collection.



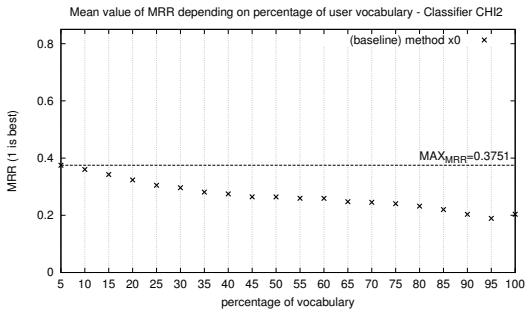

*(a)* Classifier KLD, terms ordering with KLD



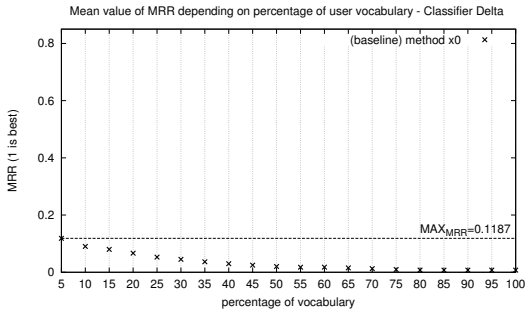

*(b)* Classifier KLD, terms ordering with TF-IDF



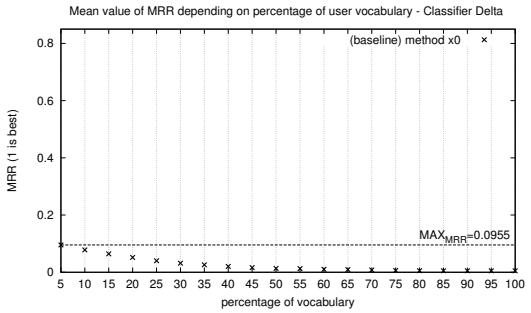

*(c)* Classifier $\chi^2$, terms ordering with KLD




*(d)* Classifier $\chi^2$, terms ordering with TF-IDF

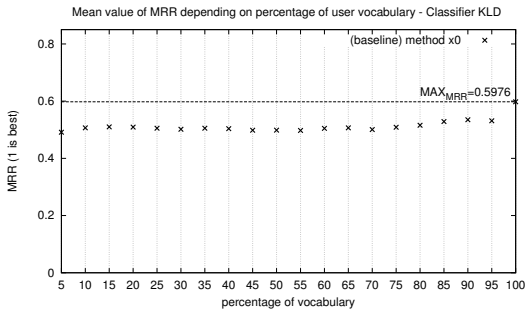



*(e)* Classifier Delta, terms ordering with KLD

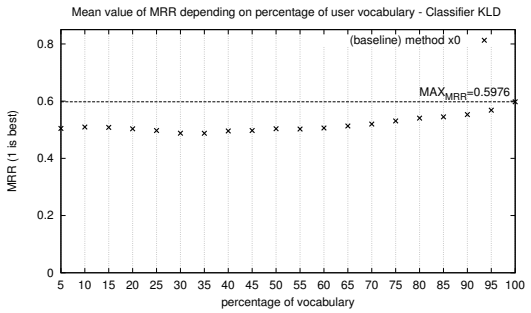



*(f)* Classifier Delta, terms ordering with TF-IDF

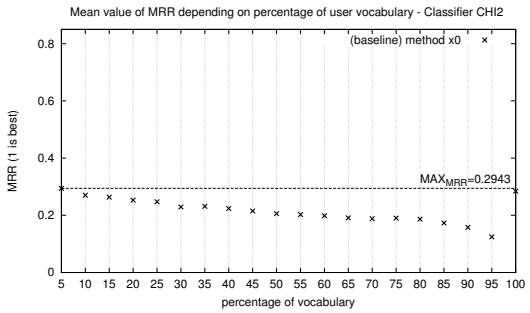
*Figure 5.16.* Experimental results employing the vocabulary selection strategy on hundreds of users for the La Stampa collection.



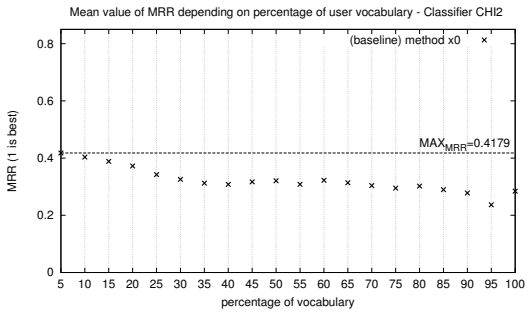

*(a)* Classifier KLD, terms ordering with KLD



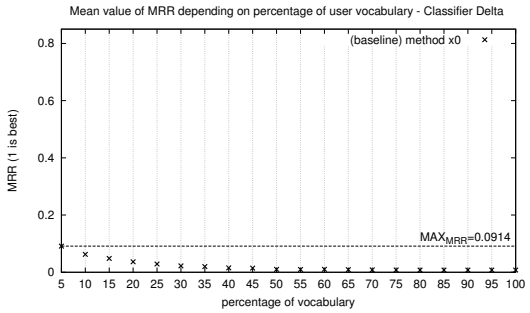

*(b)* Classifier KLD, terms ordering with TF-IDF



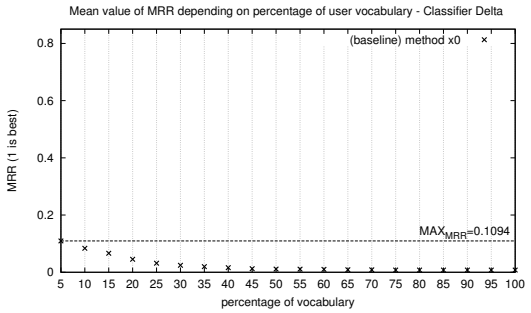

*(c)* Classifier $\chi^2$, terms ordering with KLD




*(d)* Classifier $\chi^2$, terms ordering with TF-IDF

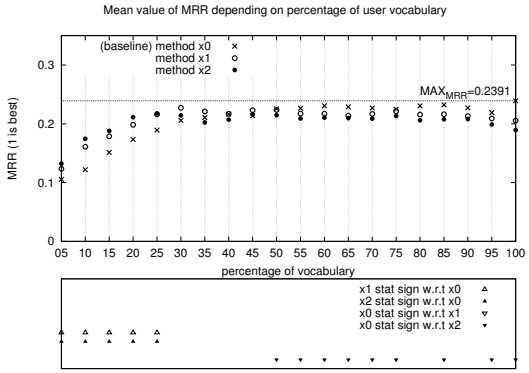



*(e)* Classifier Delta, terms ordering with KLD

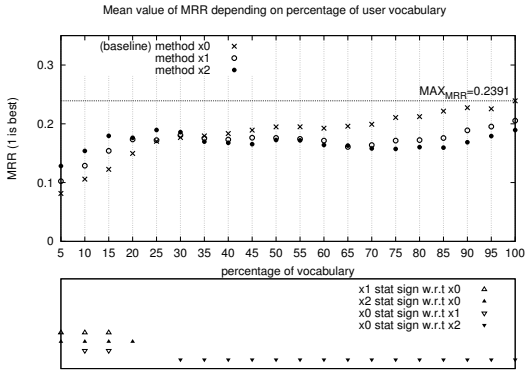



*(f)* Classifier Delta, terms ordering with TF-IDF

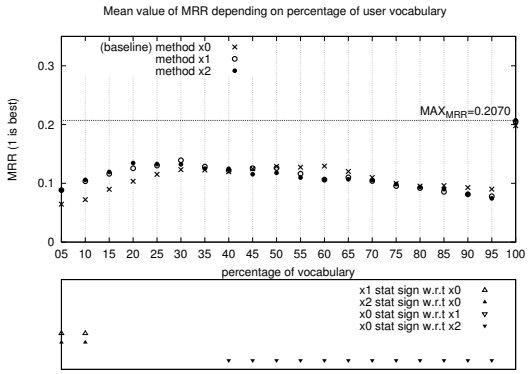
*Figure 5.17.* Experimental results employing the vocabulary selection strategy on hundreds of users for the Glasgow Herald collection.



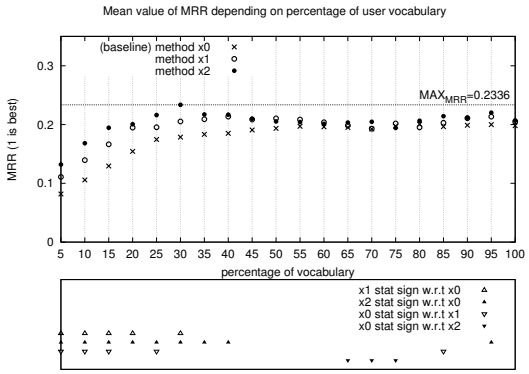

*(a)* Classifier KLD, terms ordering with KLD



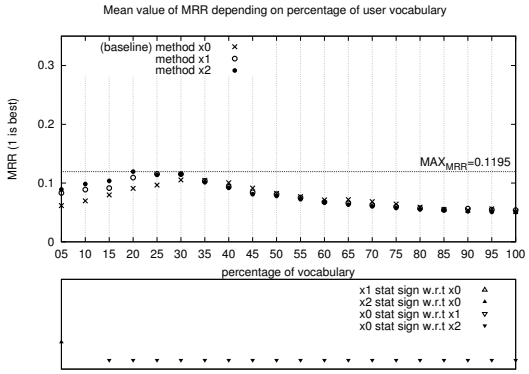

*(b)* Classifier KLD, terms ordering with TF-IDF



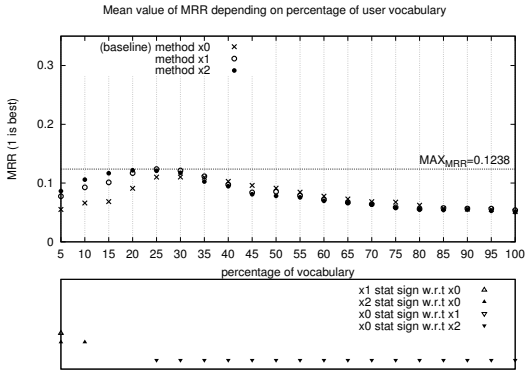

*(c)* Classifier $\chi^2$, terms ordering with KLD




*(d)* Classifier $\chi^2$, terms ordering with TF-IDF

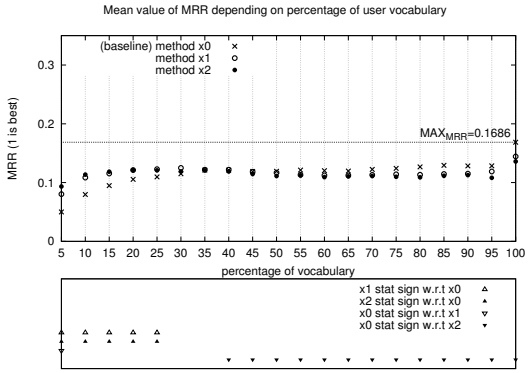



*(e)* Classifier Delta, terms ordering with KLD

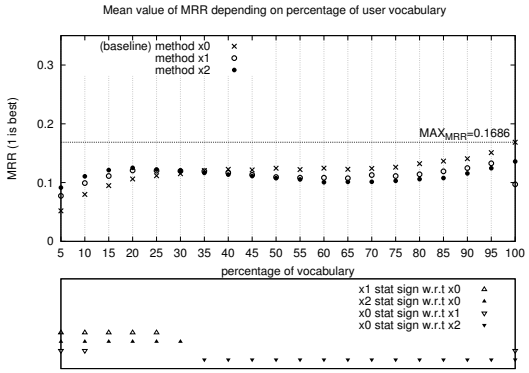



*(f)* Classifier Delta, terms ordering with TF-IDF

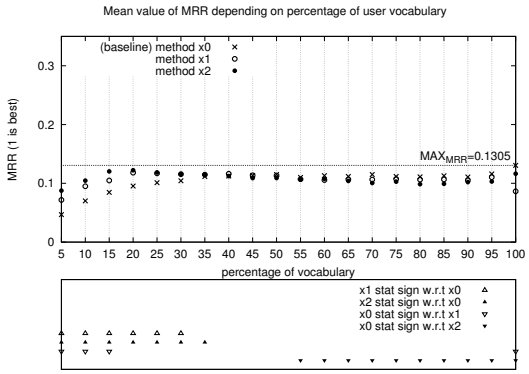
*Figure 5.18.* Experimental results employing the vocabulary selection strategy on hundreds of users for the Freenode collection



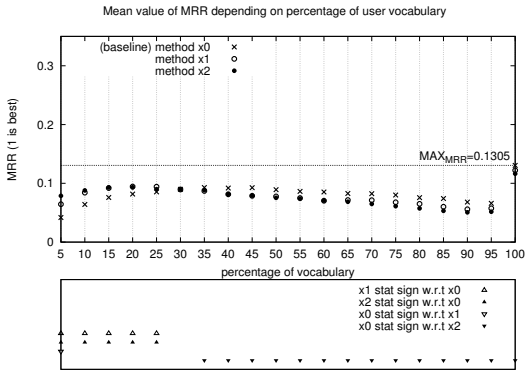

*(a)* Classifier KLD, terms ordering with KLD



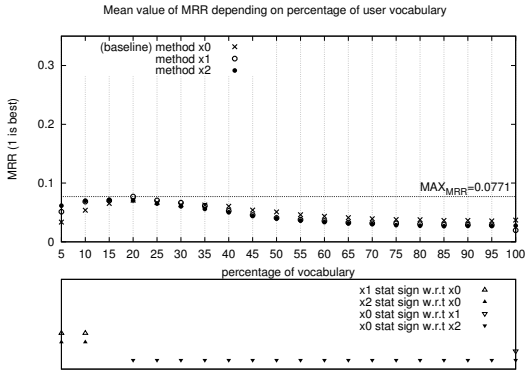

*(b)* Classifier KLD, terms ordering with TF-IDF



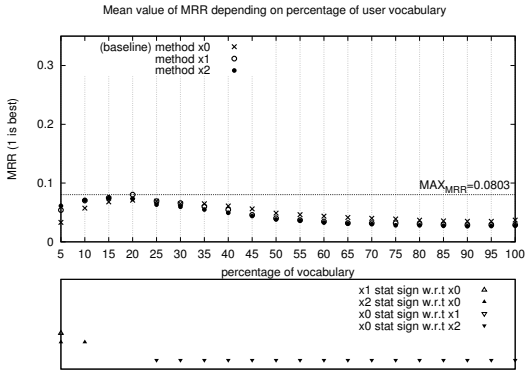

*(c)* Classifier $\chi^2$, terms ordering with KLD




*(d)* Classifier $\chi^2$, terms ordering with TF-IDF

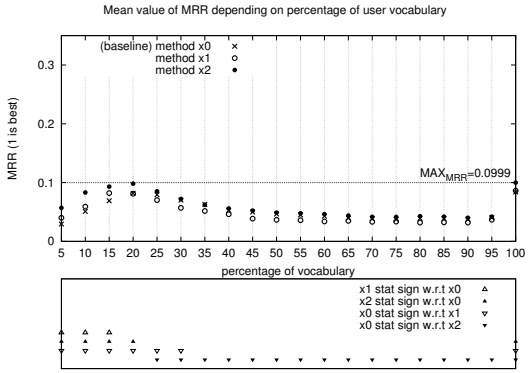



*(e)* Classifier Delta, terms ordering with KLD

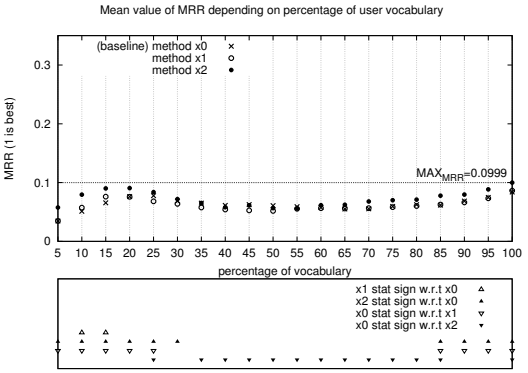



*(f)* Classifier KLD, terms ordering with TF-IDF

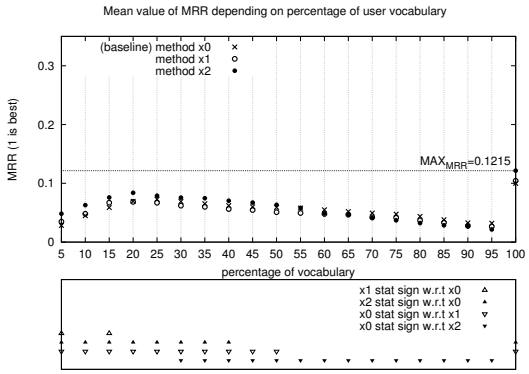
*Figure 5.19.* Experimental results employing the vocabulary selection strategy on hundreds of users for the Krijn collection



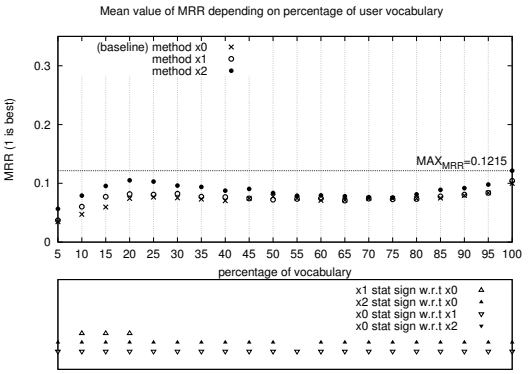

*(a)* Classifier KLD, terms ordering with KLD



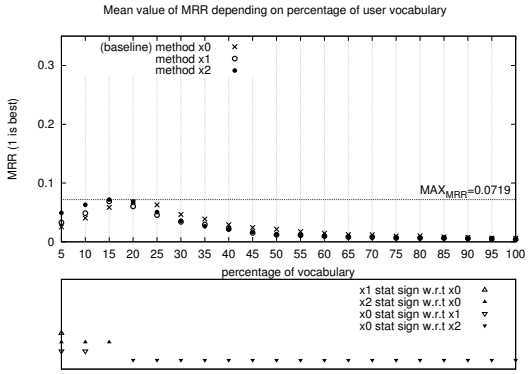

*(b)* Classifier KLD, terms ordering with TF-IDF



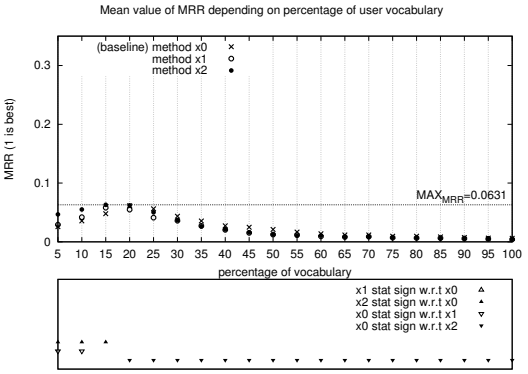

*(c)* Classifier $\chi^2$, terms ordering with KLD




*(d)* Classifier $\chi^2$, terms ordering with TF-IDF




*(e)* Classifier Delta, terms ordering with KLD




*(f)* Classifier Delta, terms ordering with TF-IDF

*Figure 5.20.* Experimental results employing the vocabulary selection strategy on hundreds users for the Twitter collection

profile, therefore they do not have any influence on the determination of the profiles length.

It can be observed that the profile expansion performed by methods x1 and x2 increases significantly the user profiles respect to the baseline x0. This has several implications. First, one might think that the MRR improvements obtained with methods x1 or x2 at lower percentages (around 30%), as illustrated in the previous Section and in Table 5.3, should be attributed to the longer profiles only (more terms) and not to the quality of the profiles (terms) introduced by the expansion. If this is true, methods x1 and x2 should always outperform the baseline x0, but this happens in some cases only. Moreover, profiles obtained with methods x1 and x2 at 30% of vocabulary are significantly shorter than those obtained with all methods at 100% of vocabulary, in particular for baseline x0. This means that, even if methods x1 and x2 introduce more features (terms) at similar profile percentages, they should always be preferred over longer profiles when performing better. For example, for collection Freenode (Table 5.3) method x2 at 30% (about 140 terms) performs better than baseline x0 at 100% (about 189 terms).

| % Voc. | Method | | |
|---|---|---|---|
| | x0 | x1 | x2 |
| 5 | 137.42 | 183.23 | 204.43 |
| 10 | 275.30 | 367.07 | 409.40 |
| 15 | 413.08 | 550.77 | 614.32 |
| 20 | 551.01 | 734.68 | 819.27 |
| 25 | 688.83 | 918.45 | 1024.25 |
| 30 | 826.61 | 1102.15 | 1229.08 |
| 35 | 964.37 | 1285.83 | 1434.10 |
| 40 | 1102.28 | 1469.70 | 1639.08 |
| 45 | 1240.07 | 1653.42 | 1843.92 |
| 50 | 1378.65 | 1837.42 | 2049.05 |
| 55 | 1515.70 | 2020.93 | 2253.75 |
| 60 | 1653.62 | 2204.83 | 2458.73 |
| 65 | 1791.30 | 2388.52 | 2663.57 |
| 70 | 1929.20 | 2572.27 | 2868.63 |
| 75 | 2067.09 | 2756.12 | 3073.57 |
| 80 | 2204.89 | 2939.85 | 3278.55 |
| 85 | 2342.68 | 3123.58 | 3483.35 |
| 90 | 2480.51 | 3307.35 | 3688.32 |
| 95 | 2618.34 | 3491.12 | 3893.23 |
| 100 | 2756.49 | 3675.33 | 4098.58 |

*(a)* 20 selected users.

| % Voc. | Method | | |
|---|---|---|---|
| | x0 | x1 | x2 |
| 5 | 13.95 | 19.71 | 22.91 |
| 10 | 28.40 | 39.94 | 46.32 |
| 15 | 42.82 | 60.13 | 69.69 |
| 20 | 57.32 | 80.42 | 93.14 |
| 25 | 71.80 | 100.65 | 116.58 |
| 30 | 86.13 | 120.77 | 139.88 |
| 35 | 100.53 | 140.94 | 163.22 |
| 40 | 115.05 | 161.22 | 186.69 |
| 45 | 129.42 | 181.36 | 209.99 |
| 50 | 144.07 | 201.79 | 233.62 |
| 55 | 158.24 | 221.77 | 256.80 |
| 60 | 172.75 | 242.04 | 280.26 |
| 65 | 187.13 | 262.18 | 303.57 |
| 70 | 201.58 | 282.39 | 326.95 |
| 75 | 216.08 | 302.68 | 350.44 |
| 80 | 230.48 | 322.84 | 373.81 |
| 85 | 244.84 | 343.00 | 397.11 |
| 90 | 259.31 | 363.22 | 420.51 |
| 95 | 273.71 | 383.41 | 443.88 |
| 100 | 288.61 | 404.08 | 467.75 |

*(b)* Hundreds of users.

*Table 5.4.* Average profiles length (unique terms) for collection Freenode, at different level of vocabulary percentage and for the different methods (x0, x1, x2).

| % Voc. | Method | | |
|---|---|---|---|
| | x0 | x1 | x2 |
| 5 | 51.99 | 69.75 | 78.35 |
| 10 | 106.32 | 139.98 | 157.20 |
| 15 | 160.45 | 210.15 | 235.97 |
| 20 | 215.38 | 280.43 | 314.83 |
| 25 | 269.57 | 350.62 | 393.67 |
| 30 | 323.55 | 420.80 | 472.40 |
| 35 | 378.05 | 490.98 | 551.15 |
| 40 | 432.37 | 561.20 | 630.02 |
| 45 | 486.41 | 631.38 | 708.75 |
| 50 | 541.82 | 701.77 | 787.85 |
| 55 | 595.06 | 771.78 | 866.45 |
| 60 | 649.69 | 842.03 | 945.37 |
| 65 | 703.35 | 912.18 | 1024.05 |
| 70 | 758.11 | 982.40 | 1102.87 |
| 75 | 812.72 | 1052.70 | 1181.75 |
| 80 | 866.66 | 1122.80 | 1260.55 |
| 85 | 921.19 | 1193.02 | 1339.23 |
| 90 | 975.26 | 1263.22 | 1418.07 |
| 95 | 1029.36 | 1333.42 | 1496.85 |
| 100 | 1086.16 | 1404.08 | 1576.15 |

*(a)* 20 selected users.

| % Voc. | Method | | |
|---|---|---|---|
| | x0 | x1 | x2 |
| 5 | 6.08 | 9.39 | 11.36 |
| 10 | 12.66 | 19.29 | 23.21 |
| 15 | 19.20 | 29.15 | 35.03 |
| 20 | 25.84 | 39.09 | 46.94 |
| 25 | 32.43 | 49.00 | 58.79 |
| 30 | 38.90 | 58.79 | 70.56 |
| 35 | 45.46 | 68.64 | 82.36 |
| 40 | 52.09 | 78.60 | 94.27 |
| 45 | 58.58 | 88.40 | 106.03 |
| 50 | 65.37 | 98.50 | 118.09 |
| 55 | 71.69 | 108.14 | 129.71 |
| 60 | 78.33 | 118.09 | 141.61 |
| 65 | 84.81 | 127.90 | 153.37 |
| 70 | 91.41 | 137.79 | 165.22 |
| 75 | 98.05 | 147.75 | 177.14 |
| 80 | 104.58 | 157.60 | 188.94 |
| 85 | 111.07 | 167.38 | 200.71 |
| 90 | 117.66 | 177.30 | 212.57 |
| 95 | 124.19 | 187.15 | 224.38 |
| 100 | 131.23 | 197.49 | 236.69 |

*(b)* Hundreds of users.

*Table 5.5.* Average profiles length (unique terms) for collection Krijn, at different level of vocabulary percentage and for the different methods (x0, x1, x2).

| % Voc. | Method | | |
|---|---|---|---|
| | x0 | x1 | x2 |
| 5 | 28.62 | 34.67 | 37.13 |
| 10 | 58.56 | 69.75 | 74.87 |
| 15 | 88.49 | 104.91 | 112.42 |
| 20 | 118.79 | 140.07 | 150.20 |
| 25 | 148.85 | 175.35 | 188.00 |
| 30 | 178.56 | 210.34 | 225.55 |
| 35 | 208.59 | 245.40 | 263.17 |
| 40 | 463.53 | 280.65 | 300.85 |
| 45 | 238.68 | 315.71 | 338.52 |
| 50 | 268.48 | 351.09 | 376.43 |
| 55 | 299.16 | 385.95 | 413.81 |
| 60 | 328.45 | 421.11 | 451.58 |
| 65 | 358.61 | 456.26 | 489.17 |
| 70 | 388.31 | 491.36 | 526.82 |
| 75 | 418.48 | 526.64 | 564.66 |
| 80 | 448.71 | 561.69 | 602.23 |
| 85 | 478.49 | 596.75 | 639.91 |
| 90 | 508.52 | 631.94 | 677.49 |
| 95 | 568.26 | 667.00 | 715.21 |
| 100 | 599.72 | 702.60 | 753.27 |

*(a)* 20 selected users.

| % Voc. | Method | | |
|---|---|---|---|
| | x0 | x1 | x2 |
| 5 | 2.58 | 4.18 | 5.28 |
| 10 | 5.64 | 8.85 | 11.06 |
| 15 | 8.66 | 13.47 | 16.80 |
| 20 | 11.77 | 18.20 | 22.63 |
| 25 | 14.85 | 22.87 | 28.42 |
| 30 | 17.81 | 27.44 | 34.11 |
| 35 | 20.84 | 32.07 | 39.83 |
| 40 | 23.97 | 36.79 | 45.67 |
| 45 | 26.93 | 41.37 | 51.35 |
| 50 | 30.20 | 46.24 | 57.34 |
| 55 | 33.02 | 50.66 | 62.88 |
| 60 | 36.13 | 55.38 | 68.71 |
| 65 | 39.11 | 59.95 | 74.40 |
| 70 | 42.19 | 64.63 | 80.17 |
| 75 | 45.29 | 69.36 | 86.02 |
| 80 | 48.32 | 73.98 | 91.74 |
| 85 | 51.28 | 78.56 | 97.43 |
| 90 | 54.36 | 83.23 | 103.22 |
| 95 | 57.37 | 87.85 | 108.95 |
| 100 | 60.89 | 92.97 | 115.18 |

*(b)* Hundreds of users.

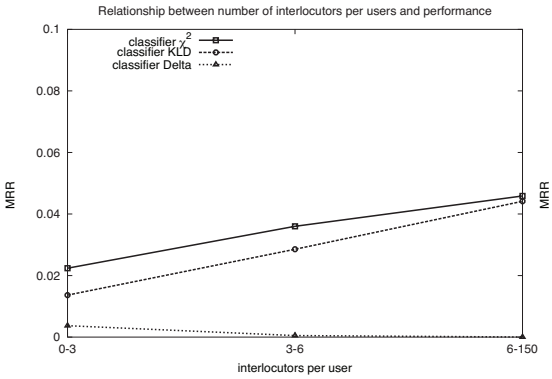
*Table 5.6.* Average profiles length (unique terms) for collection Twitter, at different level of vocabulary percentage and for the different methods (x0, x1, x2).

### Influence of Interlocutors

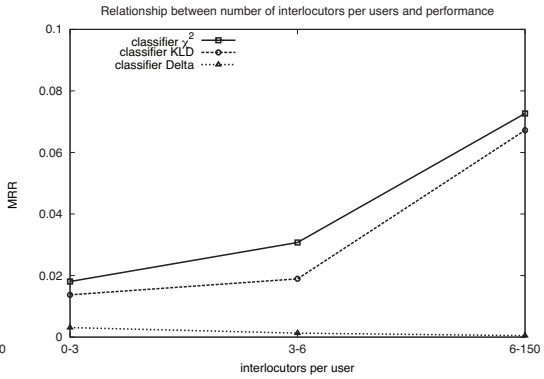
One of the goals of our analysis was to understand the influence of the interlocutors on the performance of our methods x1 and x2. To better understand the influence of interlocutors on the performance of our classifiers we grouped the number of interlocutors per user into bins of comparable size (i.e. same number of users with that range of interlocutors) and computed the classification accuracy considering the users in that bin only. This analysis only made sense on the large set of hundreds and thousands of users, rather than on the limited set of 20. Since there was not a particular level of vocabulary to test with, we decided to set it to the full profiles length (100%). We believe this setting to be general enough to draw general conclusions. Moreover it is independent from the function (KLD or TF-IDF) employed to score the terms.

We are displaing the behaviour of interlocutors per users in Figure 5.21 and in Table 5.7 for the three datasets of conversational documents. If we observe the graphs, we might note a linear relationship between the number of interlocutors and the performance of the KLD and $\chi^2$ classifiers. For Delta this relationship seems linear but negative, i.e. the larger the number of interlocutors, the worse the performances of the classifiers. This is not desired, but we recall that Delta was also the worst performing classifier, thus not really of interest. For the Krijn collection, instead, both KLD and $\chi^2$ for the method x2 are linear dependent to the number of interlocutors per user (see Adjusted $R^2$, F-statistic and p-value in the captions of sub-images in Figure 5.21) i.e. the larger the number of interlocutors, the better the performance of the classifiers. This happens also for Twitter with the KLD classifier for method x1. This is encouraging and partially supports the suitability of our methods x1 and x2 and the intuition they are based on (i.e. the more the interlocutors, the better the performance). Moreover, since for methods x1 and x2 good MRR were also obtained with short profiles length, we also studied the relationship between the number of interlocutors and the performance of KLD and $\chi^2$ also at 20% of profiles length. We could observe new linear relationships (at $p < 0.05$) for Krijn with KLD and method x1 and for Freenode with KLD and method x2. This reinforces the conclusion that there is a linear relationship (although not always significant) between the number of interlocutors and the performance of some classfiers, at least of the KLD classifier, that was almost always the best one in almost all the experiments we conducted.

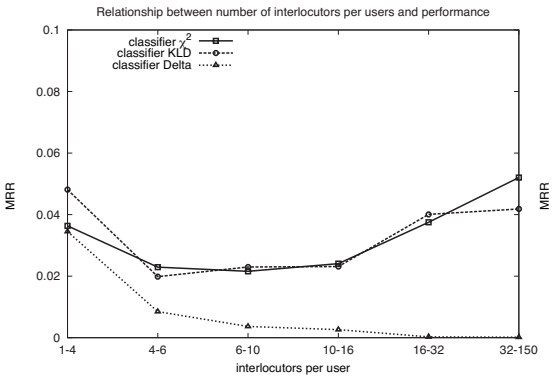
The last remark is on the number of interlocutors in the plot: to maintain a balanced number of users per number of interlocutors, we had to group them into different numbers. The only effect is in the change of the granularity of the graph: with a fewer (or greater) bins the increase would not have been so clear (although it would have been present) as with a balanced number of users per bin.



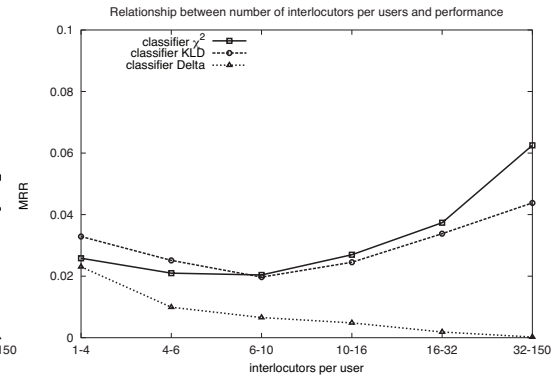

*(a)* Twitter with method x1. KLD: Adjusted R-squared: 1 F-statistic: Inf, p-value: $< 2.2e\text{-}16$



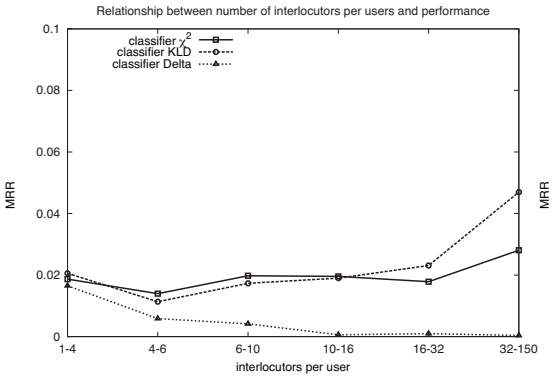

*(b)* Twitter with method x2



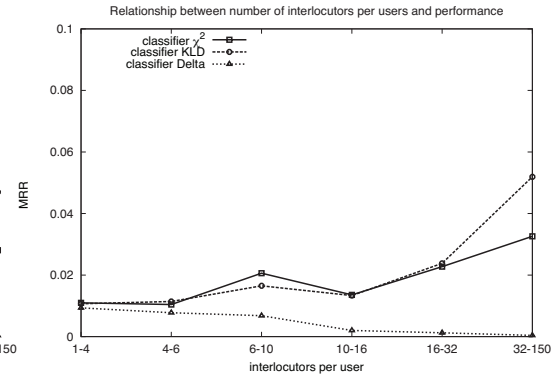

*(c)* Freenode with method x1




*(d)* Freenode with method x2




*(e)* Krijn with method x1




*(f)* Krijn with method x2. $chi^2$: Adjusted R-squared: 0.6844 F-statistic: 11.84, p-value: 0.02625; KLD: Adjusted R-squared: 0.5733 F-statistic: 7.718, p-value: 0.04992

*Figure 5.21.* Relationship between macroaverage MRR and number of interlocutors per user. Each category (interlocutors per users) contains a comparable number of users.

| interl. | num. users | classifier $\chi^2$ | KLD | Delta |
|---|---|---|---|---|
| 1-4 | 227 | 0.036 | 0.048 | 0.034 |
| 4-6 | 121 | 0.023 | 0.020 | 0.008 |
| 6-10 | 113 | 0.022 | 0.023 | 0.004 |
| 10-16 | 86 | 0.024 | 0.023 | 0.003 |
| 16-32 | 88 | 0.038 | 0.040 | 0.000 |
| 32-150 | 70 | 0.052 | 0.042 | 0.000 |

*(a)* Freenode collection with method x1

| interl. | num. users | classifier $\chi^2$ | KLD | Delta |
|---|---|---|---|---|
| 1-4 | 140 | 0.026 | 0.033 | 0.023 |
| 4-6 | 124 | 0.021 | 0.025 | 0.010 |
| 6-10 | 143 | 0.020 | 0.020 | 0.007 |
| 10-16 | 101 | 0.027 | 0.025 | 0.005 |
| 16-32 | 102 | 0.037 | 0.034 | 0.002 |
| 32-150 | 95 | 0.063 | 0.044 | 0.000 |

*(b)* Freenode collection with method x2

| interl. | num. users | classifier $\chi^2$ | KLD | Delta |
|---|---|---|---|---|
| 1-4 | 671 | 0.019 | 0.021 | 0.017 |
| 4-6 | 520 | 0.014 | 0.011 | 0.006 |
| 6-10 | 630 | 0.020 | 0.017 | 0.004 |
| 10-16 | 422 | 0.020 | 0.019 | 0.001 |
| 16-32 | 371 | 0.018 | 0.023 | 0.001 |
| 32-150 | 242 | 0.028 | 0.047 | 0.000 |

*(c)* Krijn collection with method x1

| interl. | num. users | classifier $\chi^2$ | KLD | Delta |
|---|---|---|---|---|
| 1-4 | 332 | 0.011 | 0.011 | 0.009 |
| 4-6 | 454 | 0.010 | 0.011 | 0.008 |
| 6-10 | 683 | 0.021 | 0.017 | 0.007 |
| 10-16 | 506 | 0.014 | 0.013 | 0.002 |
| 16-32 | 529 | 0.023 | 0.024 | 0.001 |
| 32-150 | 362 | 0.033 | 0.052 | 0.000 |

*(d)* Krijn collection with method x2

| interl. | num. users | classifier $\chi^2$ | KLD | Delta |
|---|---|---|---|---|
| 1-3 | 1027 | 0.022 | 0.014 | 0.004 |
| 3-6 | 750 | 0.036 | 0.029 | 0.000 |
| 6-150 | 639 | 0.046 | 0.044 | 0.000 |

*(e)* Twitter collection with method x1

| interl. | num. users | classifier $\chi^2$ | KLD | Delta |
|---|---|---|---|---|
| 1-3 | 976 | 0.018 | 0.014 | 0.003 |
| 3-6 | 726 | 0.031 | 0.019 | 0.001 |
| 6-150 | 714 | 0.073 | 0.067 | 0.000 |

*(f)* Twitter collection with method x2

*Table 5.7.* Relationship between macroaverage MRR and number of interlocutors per user.

## 5.6 Limitations

The main limitations of the study conducted in this chapter are related to the design choices and type of approaches considered. First, we employed as conversational documents only those originated from IRC and Twitter, but other sources might have been considered, for example email conversations and fora. The same can be said for the traditional documents, that we selected from newspaper articles only, although other sources already employed in the literature, like poems or books, could had been taken into consideration. Second, we decided to work at features level and worked on improving them, rather than working on improving the classifiers. Future work might extend the current one addressing this point.

Another limitation lies in the number of users per fold employed in the study: we decided to employ 1/2 only of the authors for each collection in each fold, 1/16 for Twitter. This was done to handle the different experiments within a reasonable time frame given the hardware at our disposal. The machine at our disposal is considered in any case an excellent machine with the best hardware to date (see footnote in Section 5.4.2). Moreover we wrote the code to test methods x0, x1 and x2 to make use of all the available cores of the machine. Examples of running times can be found in Appendix D. We believe that the next step to improve the number of users considered in the study would be to move to a distributed network of machines or, simply, allow for more time to wait.

Finally, we did not investigate the author profiles "backwards", employing different percentages of vocabulary, from the most common terms to the most specific. We also left to future study the analysis of combinations of different portions of the authors vocabulary (e.g. the combination of a portion of the most common words and a portion of the most specific words). Another open point is the selection of the interlocutors' text to be used to expand an author profile. We decided to employ the texts immediately after (method x1) or immediately before and after (method x2) the individual messages of each author. However, these are only two possible strategies that future studies might complement with different approaches to determine the quantity (i.e. the number of messages preceding or following the author ones) of text needed to expand author profiles.

## 5.7 Summary

In this chapter we studied the problem of authorship identification for conversational documents. We first introduced the problem with the associated related work and then presented the plan of our study. We explained the role of classifiers and feature selection algorithms in the field of authorship identification and presented the baselines and challenges of this problem. We suggested a method for improving the classification accuracy when dealing with conversational documents, such as IRC chat logs or Twitter,

and compared this approach with the standard methods employed in the literature with collections of newspaper articles, with a conventional number of authors (20) and with an augmented number (hundreds or thousands) of them. With the use of well known classifiers, KLD, $\chi^2$ and Delta, we experimentally validated our methods and observed a dependency in the accuracy of the classification with the number of interlocutors per user. Our initial hypotheses were only partially confirmed and we could identify room for improvement and future work.

# Chapter 6

# Conclusions and Future Work

*The greater the struggle, the more glorious the triumph.*

Mendez, director of the Butterfly Circus

In this chapter we will summarise the work presented in the dissertation and present the final conclusions. We will also give an outlook to future work emerging from our analysis.

## 6.1 Main Results

When we started our work, documents generated on the Internet were starting to capture the interest of the research community. Despite the fact that some investigations were already conducted for some specific documents (e.g. blogs or reviews), what was missing in the literature was a complete overview of the properties of these novel documents and a comparison with traditional documents. For this reason we asked:

**RQ1** *What are are the differences between traditional collections of documents employed in Information Retrieval (e.g. newspaper articles) and recent collections of documents generated online (e.g. Internet Relay Chat -IRC- logs, Twitter)?*

In Chapter 4 we addressed this problem with different experiments. We were able to demonstrate that recent collections follow the same Zipfian distributions as standard IR collections, thus being suitable for standard indexing and retrieval algorithms of IR. On the other hand, the collections of conversational or discussion documents manifest a higher vocabulary growth, which implies the presence of less conventional and more diverse terms in their vocabulary. The terms, however, do not constitute discriminant features when it comes to the identification of the topics of the documents, due to their poor bursty distribution. On the other hand, some of these terms are emoticons, abbreviations, interjections or spelling mistakes that constitute an important part of the

vocabulary of these collections of documents and that are totally absent from the collections traditionally used in IR. These features, instead of topics, can be employed to identify the authors of the documents, thus are very useful in the problem of author characterisation or identification. Finally, the length of the documents within the recent collections is really different from that of standard collections. This affects, together with their vocabulary, the level of similarity of the online user-generated documents when compared one against the other. This has also implications on the classical IR systems, for example in the retrieval phase: given the high number of false positives, the list of relevant retrieved documents might be very imprecise. In the same chapter we proposed methods for dealing with this problem (e.g. document expansion) and to improve the quality of the retrieval or the accuracy of the classification of such documents. For example, in Chapter 5 we employed techniques of documents expansion for conversational documents to build the profile of one author. In the intermediate step of our method we merged into a single profile not only all the messages of one user, but also the messages the user exchanged with his interlocutors. This resulted in an enlarged profile that could be used to extract more relevant terms than from the original user profile.

Within RQ1 we considered other two minor aspects of the collections of online user-generated documents. Since the classes of documents generated on the Internet are broad, we asked:

- *Is there any difference within these collections of online user-generated documents: i.e. are Twitter documents different from IRC logs or from blogs?*
- *What are the specific characteristics of conversational documents (e.g. IRC logs)?*

In Chapter 4 we also found answers to the above two questions. From the analysis conducted we could identify two classes of documents: conversational and discussion documents. We included in the first class short or very short documents (like Twitter or the IRC logs) which present a high number of emoticons and other markers of non-verbal communication. Despite that, from the studies in Chapter 5 we noted small differences also within this class (between IRC logs and Twitter), which makes its analysis more interesting and challenging. Discussion documents, on the other hand, contain documents (e.g. from fora, blogs, newsgroups) which for their length and style and, in particular for their POS, are closer to the class of standard IR documents (newspaper articles) than conversational documents.

Based on the answers to RQ1, we decided to select only one class of documents to be analysed in detail in our work. We opted for the more challenging group of documents, the conversational ones. These were in fact less studied in the literature than the discussion ones. Moreover, the conversational nature of these documents made them short and containing a specific vocabulary, thus increasing the difficulty in processing and analysing them.

The first question at the beginning of the analysis of conversational documents was:

**RQ2** *Is there a good representative collection for conversational documents, with a large number of documents and a variety of topics?*

Although the question might have seemed trivial, in Chapter 3 we illustrated the previous situation in the literature for collections of conversational documents. For Twitter there had been some previous studies, while for IRC logs the situation was clearly different. There were in fact, few datasets available for studying the behaviour of hundreds of users producing textual documents. Furthermore, these previous datasets contained only a limited number of documents and users and were mostly monothematic. For our studies we needed a more robust, reliable and ample dataset of online conversations. In order to motivate us in creating this kind of dataset, we decided to organise the Sexual Predator Identification competition, as part of PAN 2012 within CLEF (presented in Chapter 2), that allowed 16 teams from all over the world to take part in the competition. The results of the competition are presented in Appendix B, where the different methods and features employed by the different teams are presented and the most successful ones are commented. These results also helped us with the next research questions, in deciding which strategy could be used to approach the problem of author identification in conversational documents.

Although the organisation of such a task might be seen as purely instrumental, the collection created and released represents a significant improvement of the state of the art in the analysis of conversational documents, because such a complete collection was until then missing. Moreover, the collection allowed participants in the competition to test their algorithm on a common ground. Finally, part of the collection was employed also in PAN 2013 as part of the author profiling task and in our own experiments on author identification.

After having understood the properties of the conversational documents and having created a collection containing this kind of documents, we investigated them with the following question:

**RQ3** *Are traditional methods of authorship identification suitable also for conversational documents?*

We decided to investigate authorship identification instead of authorship characterisation because the latter needs a specific application in mind (like the Sexual Predator Identification) or a better ground truth, containing, for example, additional data on the authors like age, gender, preferences, which we did not have and could not retrieve. Moreover, it is easy to see that once the authorship identification problem is solved, we might group sets of users within certain categories, thus reducing the problem to a user characterisation one. In Chapter 5 we presented the results of our studies in the field of authorship identification for conversational documents. In developing methods to solve this problem, we advanced the state of the art by applying three statistical classifiers on different sets of features selection algorithms, two traditional and two innovative that we proposed. It was the first time in the literature, to the best of our knowledge,

that statistical classifiers had been applied to the problem of authorship identification for conversational documents. This was also one of the few studies in the literature on the general topic of authorship identification for conversational documents. We first verified that traditional algorithms for features selection were not working for conversational documents. As an example of conversational documents, we employed a subset of our collection composed by IRC logs and a dataset of Twitter messages for which the conversations were already identified. Despite the fact that in a traditional setting with few users all the methods seemed to work quite reasonably, even for conversational documents, we detailed better the authorship identification problem with the following additional questions:

- *Are traditional methods of authorship identification suitable also for hundreds or thousands of authors, like in the case of conversational documents?*
- *How does the number of interlocutors of an author impact on the performance of author identification?*

In fact, while for standard collections of newspapers the settings with few users might be realistic, for conversational documents this is not true anymore and we needed a wider scenario, with hundreds or thousands of users. For this reason we opened up the problem to a bigger number of users.

Under these settings we verified that for hundreds of users the traditional methods were not performing well with the use of stopwords only. This behaviour changed with the use of all possible features (e.g. 100% of author profiles length), in particular for the basic method of simple profile building. The methods we proposed, based on the profile expansion with interlocutors terms and vocabulary scoring, were only partially verified: the two methods were not able to beat the standard one with full profile length for hundreds of users, but they were better and close to the best for shorter profiles (although expanded) and doing definitely better for few (20) users instead. This is encouraging and we believe with additional work and study the proposed methods might in the future overtake the simple ones.

Finally, as for the number of users, we could observe that for both IRC logs and Twitter, the more interlocutors an author has, the better the performance of the proposed methods. This is due to the process of document expansions, that took advantage from users interacting together in conversations and allowed for a better selection of discriminative terms for single users.

To conclude, the work reported in this thesis had an impact on different fields of research (IR, Text Mining, NLP) in its first part, dedicated to the study of the general properties of short user-generated documents in the Internet. In the second part, we advanced the state of the art by introducing a novel collection of conversational documents and proposing novel algorithms for authorship identification on that collection of conversational documents.

### 6.1.1 Domain Specific Applications

We are presenting here two domain-specific applications that might directly benefit from the findings of the work presented in this dissertation.

**Software Engineering** The popularity of methods for extracting useful information from unstructured code repositories or development websites [6] can be an interesting application field for the proposed methods. In fact, with the proposed method for authorship identification we might detect those developers that are the most effective or active in solving bugs or signalling possible solutions to known problems. They might also be useful to identify trends among the users engaged in conversations about problems or good features of programs, that are often discussed online by experts or power users.

**Sexual Predator Identification** With our proposed methods one could also improve the results obtained by participants in the competition of Sexual Predator Identification. Following the idea of identifying the common features of a group of users, one can employ our algorithm for user identification to first label unknown profiles and then can group together those users suspected to be possible predators. This way it is possible to measure the accuracy of predator detection according to the metrics presented in Appendix B.

In the next section we will present some future work emerging from the contributions presented in the dissertation and summarized above.

## 6.2 Future Research Directions

In this section we will identify the future work emerging from this dissertation, for conversational documents in general and for the authorship identification problem more in detail.

**Other classifiers and additional settings for the authorship identification problem** Further improvement of the work on authorship identification might include the use of a larger portion of the datasets, up to the full list of users and documents. Other extensions might include the study of other metrics to select discriminant terms for each user, e.g $\chi^2$ or the cosine similarity. It would be interesting to perform a reverse analysis from the most common terms to the most specific ones for the authors profiles and to test in depth combinations of different kinds of terms (e.g. the top 30% most significant combined with the least 10%). Finally, it should be investigated the quantity (i.e. the number of messages preceding or following the author ones) of text needed to expand author profiles.

**Other conversational or discussion media** In this work we considered as representative of conversational documents only those that were strictly matching the definition given in Chapter 2 and the properties identified in Chapter 4. However, other kinds of documents exist, like fora, newsgroups, blogs or emails, in which people are also involved in "conversations". For this reason it would be interesting to apply the methods proposed and tested to work well on IRC logs also on this kind of documents. For example, the Enron Email Dataset might be a good testbed for conversations in emails and one should find also a representative dataset for messages from fora or newsgroups. The issues in analysing these kinds of documents might be in the pre-processing stage, where one should be careful in identifying correctly the persons and the conversational threads within the general discussions.

**Temporal Model of the Conversations** In presenting this method for authorship identification, we stated that within a conversation there were no differences between A sending a message to B, or vice-versa. This was however a simplification and more refined models might be employed to capture this behavior as well as the temporal evolution of users vocabulary. In fact, along the same conversation or within different conversations spread over some days or weeks, a user might change topics, writing style, ideas or might be subject to different emotional statuses that influence its conversation. For these reasons, interesting future work might explore these evolutions in users behavior within his history of conversations.

**Topic Identification (and Tracking)** This is related to the temporal modeling of the user's behavior within a conversation. It might be the case that tracking all the conversations of a user over time is computationally heavy and expensive. For this reason, a solution might be to track just particular features, for example user's interests or topics of discussion in the conversation. In doing this, one could reduce the amount of data and might associate semantic concepts to the different conversations or parts of the conversations. The identified topics might also be used in different ways, for example for automatic labelling of a conversation of for mining users interests for the purpose of marketing, advertisement or security. Moreover one might create a hierarchy of topics and be able to characterise a conversation or a user profile by different levels of granularity.

**Summarisation** Summarisation methods could provide another dimension to the conversational documents. Methods for summarising conversations to provide an "abstract" of a long discussion might help in jumping into a conversation without the need of reading all the previous messages but just a summary. Moreover, in case of verbose or too long user profiles, obtained through methods of document expansion, techniques used in summarisation might help in identifying the best representative features (e.g. terms) for a specific profile.

**User Characterisation** All the extensions presented so far might help in moving from the authorship identification problem, in which we aim at identifying single users, to the problem of characterising or profiling users based on their interests (topic) or the evolution of these interests over time (temporal modelling). If one is able to identify single groups of users with the same characteristics, then summarisation techniques might provide an aggregated view of the data associated with each class of users.

**Sentiment Analysis** Another useful dimension that might be explored is the one related to the sentiment and opinion detection for each message in the conversation or for each conversation or for groups of conversations. Being able to understand the emotions and mood of a user through his messages, might allow us to better track its interactions and be able to forecast his behaviour or future actions. Moreover, in combination with topic detection and tracking, we could enrich his profile and observe the relationship between profile, topics and mood.

**Visual Representation** Finally, all the presented applications and studies might take advantage of any visual representation: from the analysis of the relations between users with graphs, to colour variations when modelling sentiment, to the histograms of different words for different topics in different instants of time, to bubbles of features for user profiles characterisation or text summary. Or any combinations of these and other techniques.

## 6.3 Summary

In this chapter we highlighted the main contributions of the dissertation, that answered the research questions presented in Chapter 1 and that guided us through this work. We also mentioned two possible fields of application for the proposed algorithm of authorship identification and we concluded presenting some possible extensions and future work related to the dissertation.

# Appendix A

# Authorship Characterization for SocialTV

*To achieve great things, two things are needed: a plan and not quite enough time.*

Leonard Bernstein

In this appendix we will present a framework to generate auxiliary rich TV content metadata by processing social networks data. We employed Twitter as an example of social media and created a method to compute their informative value based on simple criteria to identify authoritative social media sources. We extracted dozens of features from short Twitter messages, mostly inspired by the analysis conducted in Chapter 4. With the help of these features we could characterise the messages from Twitter in terms of quality and relevancy to TV shows.

The work presented in this appendix was realised during a visit to AT&T Labs Research in Middletown, NJ, U.S.A. in March-April 2011 where we worked with Dr. Andrea Basso in the group directed by Dr. Behzad Shahraray. This work was also presented at the International Workshop on Search and Mining User-generated Contents (SMUC) within CIKM 2011 [58].

## A.1 Introduction and motivations

The increasing popularity of Social Networks such as Twitter and Facebook opens up new possibilities of analysis and research as suggested in [117], where the characteristics of Twitter messages generated during the political TV debate between Obama and McCain in 2008 has been analyzed. The authors measured the volume of the data (messages) at each minute of the show, the density of the social graph and the most mentioned user and provided a match between the most significant term used in the

Twitter messages and the topic being discussed in real time. The corresponding sentiment analysis is presented in a second contribution [29]. In a more recent work [82] the authors include a sentiment analysis in their real-time per-topic peak detector on a generic Twitter dataset. Their focus is on the detection of the relevant Twitter messages in a hierarchical topic organization, based on the per-minute message rate and on the message content.

In this contribution we still concentrate on TV shows related messages but, instead of restricting to a particular event, we analyze 8 different programs on different broadcast networks with the aim of extracting the best messages in terms of their information quality. We can define the quality of a Twitter message based on two parameters: i) its relevancy to the topic and its information content and ii) being produced by a source, that according to a given criterion we can consider reliable and thus trusted. Our goal, in fact, is not to detect the topical relevancy of a Twitter message, nor to identify Twitter messages relative to a given topic, but to identify the best messages in terms of their reusability to extend and enrich a related Electronic Program Guide (EPG). The remainder of this appendix is organized as follows: in Section A.2 we present the overall framework and in section A.3 the datasets we employed later in our experiments and how we have built them. Finally, in Section A.4 we comment on the experiments we have performed and the key results we have obtained. In Section A.5 we present the conclusions and future work emerging from this study.

## A.2 Framework

The proposed framework is presented in Figure A.1. It consists of 3 main components:

1. a filtering component that has the task of selecting the relevant Twitter messages by source selection and message validation;
2. an analysis component that performs feature extraction of more than 100 features from the filtered Twitter stream, performs classification based on these features and computes their information quality;
3. synthesis where the Twitter messages with the highest content quality are aggregated together to extend a reference EPG related to that particular show.

### Filtering

The process of filtering relevant tweets from the totality of the Twitter messages is generally a quite complicated problem. Approaches presented in literature [105, 102, 107] make use of language understanding techniques to assess topic relevance and use articulated user profiling techniques to select Twitter users that can be considered trusted. Such methods target sentiment or in general trend estimations. In our approach we follow a different route because our focus is computation of content quality.

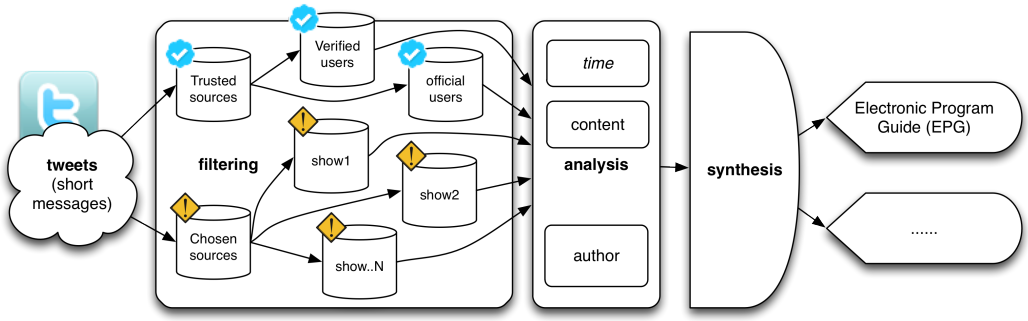



*Figure A.1.* The proposed Framework.

**Source Selection** We rely, for the selection of the relevant tweets, on authoritative sources defined as such by 3rd party entities. In particular for every show that we have selected, we considered all the Twitter users that are published in the official TV show website (*official sources* in Figure A.1) to be trusted sources. In addition we selected all the Twitter users that are members of the staff related to the show (i.e. authors, actors, directors, etc.) which are also *verified users* according to Twitter. Note that the set of Twitter users identified after the filtering process include both individual Twitter contributors as well as Twitter accounts that are the official spokespeople of the specific show (i.e. `@nbcnightlynews`). The union of *official users* and *verified sources* leads to the set of *trusted sources*[1].

**Message Validation** From the Twitter cloud, we validated the set of messages relevant to the specific TV show and generated the *trusted sources* by simple keyword matching. We will provide more technical details on the filtering step when discussing the datasets in section A.3.

## Analysis

The resulting aggregated message feed is then analyzed to study some specific characteristics of the Twitter users and study potential emerging patterns. Some of the questions that we are addressing in the analysis component are the following:

1. Is there a dominant writing style that characterizes the Twitter feed? In other words, is a specific Twitter author in general characterized by his own writing style or, on the contrary, every Twitter author somehow "adapts" his writing style to the rules specific to a short message framework as Twitter?

2. How different is the Twitter content and style of official sources (such as `@nightlynews`) from the ones coming from individual Twitters?

[1]We are using the terms *source* and *user* interchangeably.

3. Is there a writing style (official source w.r.t. individual contributors) that is the preferred one in terms of popularity and number of followers?

4. How can we measure the information content of a given Twitter feed?

   A detailed discussion of the analysis is presented in section A.4.

### Synthesis

This part of the framework component was left to future development and ideally contains the following tasks:

1. select the tweets with the highest information content

2. aggregate them in a writing style that is the most appropriate (i.e. the most popular one)

3. transform it in a form that is suitable for integration and enhancement of an Electronic Program Guide (EPG).

## A.3 Datasets

We crawled the Twitter messaging system for about 4 weeks (20 March 2011-28 April 2011), filtering through the official API[2] all the messages by TV shows as preliminary indicated in Section A.2. We considered 8 TV shows: 6 entertainment programs or series and 2 news programs. In the category of entertainment we analyzed: *Dancing with the Stars*, *Desperate Housewives*, *Dr. House*, *Grays Anatomy* and *Modern Family*. The news shows are *ABC 20/20* and *NBC nightly news*.

In Table A.1 we present some elementary statistics of our datasets (number of messages per TV show and number of "relevant" messages from trusted users) and the filtering components (users and keywords). The filtering components allowed us to select from the whole Twitter crowd only those messages that could be of interest for each TV show. The first filtering component regards the users: we stored all the messages produced by an *official user* of the TV show. An official user is the official Twitter account of the TV show, as indicates on the show's webpage, thus the most trusted source of information for the show itself. Other relevant users (not present in the table) are the verified users of Twitter. They are Twitter accounts which have been verified (offline) by Twitter as real, active and true users[3].

[2] `http://dev.twitter.com/doc`

[3] E.g., in Twitter there exist a lot of accounts referring to US president Barak Obama (`barak_obama`, `fansofobama`, `Barak0bama`, `BarakObama__`, `...`) but only one is the *verified* one, that means really maintained by the president (or his staff): `BarackObama`. A complete description of the verified users and the verification process can be found in `http://twitter.com/help/verified` (last check December 2013)

| TV show | tot mex | num. of mex. ( % trusted) | official Twitter account | filtering keywords |
|---|---|---|---|---|
| *) ABC 20/20 | 564 | 481 (85%) | @ABC2020 | *"ABC 2020"* |
| Dancing with the Stars | 320532 | 1620 (0.51%) | @ABC_DWTS | *dwts*, *DancingwiththeStars*, *dancing with the stars* |
| Desperate Housewives | 70556 | 103 (0.15%) | @DHousewivesSite, @DesperateABC | *desperatehousewives*, *desperate housewive* |
| Dr. House | 70952 | 284 (0.40%) | @HOUSEonFOX | *House*, *dr house* |
| Greys Anatomy | 59803 | 170 (0.28%) | @Greys_ABC, @GreysAnatomyFns, @mirandabaileymd | *GreysAnatomy*, *greys anatomy* |
| Modern Family | 61934 | 155 (0.25%) | @modernfamilytv, @ModernFamABC, @modfamquotes | *ModernFamily*, *modern family* |
| *) NBC nightly news | 5731 | 703 (12%) | @nbcnightlynews | *nbc nightly news*, *nbcnightlynews* |
| The Mentalist | 25599 | 165 (0.64%) | @Mentalist_CBS, @Mentalist_Fans | *Mentalist*, *TheMentalist*, *the mentalist* |

*Table A.1.* Datasets facts and figures plus filtering dimensions. *) are news channel.

We intersected all the verified and official accounts (trusted set of streams) with the other filtering dimension, represented by some *keywords* (Table A.1). We extracted from Twitter all the messages related to a TV show using these keywords, which were identified as the most relevant to each TV show after an extensive manual inspection. Since our goal is the quality of information (and not the quantity) we were conservative in the choice of the filtering keywords and did not allow more generic keywords to be used. Moreover we did not employ other specific topic-detection techniques [105, 102, 107] which are out-of-scope of our analysis. For this reason we obtained only a small amount of messages (less than 1%) from the trusted sources for shows other than news channels. For the news channels we obtained the opposite behavior, to the extreme case of *ABC 20/20* where almost all the messages are from the official source. Because of this, in the experimental part we will not report results relative to this news channel.

## A.4 Experiments

We performed two sets of experiments on our datasets to give an answer to the questions highlighted in Section A.2. With the first set of experiments we want to *characterize the streams* depending on their content and try to exploit the properties of the informative streams, while with the second set of experiments we want to find a way of *detecting the* (most) *informative streams* and *measure the information content*.

### Stream characterization

We divided the problem of characterizing the streams and their properties into two parts: features extraction and analysis based on the extracted features.

#### Feature extraction

We processed each message for each TV show to extract a variety of features, from the simple measurement of the frequency of symbols (i.e. colon, semicolon, parenthesis, brackets, ...) to the identification of emphasis in the message (through the detection of emoticons, abbreviations or "shoutings"[4]).

We mostly relied on two tools for the extraction of the features: a Part-Of-Speech (POS) tagger[5] and an *emphasis* detector. As anticipated, we considered as

- *emphasis* any of these expressions: emoticons, abbreviations and "shoutings".

The other features extracted are:

[4]"Shoutings" intended not as capital size word but word containing letters repeated more than twice (not present in any standard English dictionary).

[5]GATE: "General Architecture for Text Engineering", `http://gate.ac.uk/`, here the full list of tags detected: `http://tinyurl.com/gate-pos`

- *token*, from part of speech ( CC, CD, DT, EX, FW, IN, JJ, JJR, JJS, JJSS, LRB, LS, MD, NN, NNP, NNPS, NNS, NP, NPS, PDT, POS, PP, PRPR$, PRP, PRP$, RB, RBR, RBS, RP, STAART, SYM, TO, UH, VBD, VBG, VBN, VBP, VB, VBZ, WDT, WP$, WP, WRB)
- *symbol* (colon, comma, period, dollar, ddash, dquote, grave, lpar, rpar, pound, squote)
- *case* (upperInitial, allCaps, lowercase, mixedCaps)
- *kind* (punctuation, word, number, symbol, null, apostrophe)

It is to be noted that both *token* and *symbol* are sub-classes of *kind*. As result, we obtained a normalized per-message count of each feature. We added to the list of features also some language independent attributes to explore further the user behavior w.r.t. the messages:

- *retweets* (number of times a Twitter message was forwarded to other users by another user than the author of the message)
- *followers* (number of users who are following the author of the message)
- *messages* (the number of messages from one user)

**Stream analysis**

Since we did not have any a-priori information on the distribution of the features, we started our investigation by running an unsupervised clustering algorithm for each set of features in an independent way. Only later did we merge the selected groups of features to increase the relevancy of the results. We chose the Expectation Maximisation (EM) algorithm[6] with automatic number of cluster detection based on cross-validation as clustering algorithm.

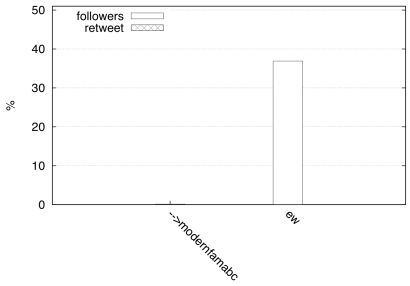
To better study the behaviour of the clustering algorithm given the features, we performed three different algorithm runs: i) considering each set of features as independent, ii) reducing the amount of components per feature and, only at a later stage, iii) merging sets of features together. The study of the independent set of features was done to understand the impact of each set on the user writing style or characteristics, to highlight the most relevant ones. To reduce the components of each feature set, we relied on previous work to decide which one to include or exclude, for example we consider only the most used and significant *tokens* from the POS tagger [41] (common and proper noun - NN,NNS,NNP,NNPS, pronoun - PRP, WP, verb, determiner -WDT, DT, WP$, PRP$) or ignored some rarely used *symbols* (grave, pound). We also tried to combine some sets of features according to their semantic:

[6]WEKA (`http://www.cs.waikato.ac.nz/ml/weka/`) implementation.

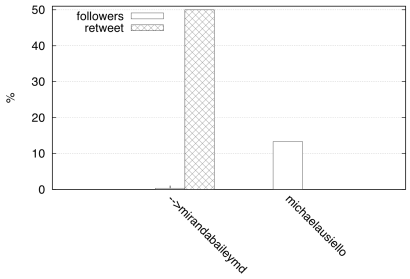
- *emphasis+*: combination of *emphasis* and *case* (all the possible emphasis that a user could give to the message)
- *language*: combination of *token* and *symbol* (the writing style of a user)
- *authoritativeness*: combination of *retweets* and *followers* (the social importance of the user in the community) [39]

What we were expecting was a kind of gentle separation between Twitter messages generated by the official user of the TV show and their related verified user (i.e. producer of the show, participant, conductor) and other Twitter users. However, the analysis of the set of features in an independent and, later, in a reduced way did not lead to a clear separation between groups of users. The official users were often spread into different clusters (depending on the TV show and the feature set) and sometimes mixed with other non official users and not directly related with the TV show. This could be observed mostly in the cluster obtained with the *token* features: possessive pronouns and endings and adjectives when all used may be indicators of an official user (to the contrary, the lack of one of the three could be an indicator of a non-official user), verbs in different tenses are present in all the official users while the lack of some tenses could be the indicator of a non-official user. We observed also that official users tend to use a more formal style (*case* features), i.e. to use words written with lowercase letters rather than with uppercase or a combination of lower and upper. Moreover, particularly from the usage of symbols, interjections and emphatic expressions (*kind*, *symbol*, *emphasis*) we observed an "adaptation" to the media (Twitter) for all the trusted (official + verified) users. For this reason, these features were not discriminative for official users, with the only exceptions of the news shows, where these features were rarely used. We did not further investigate the problem of the language adaptation to the media, which we left to future study.

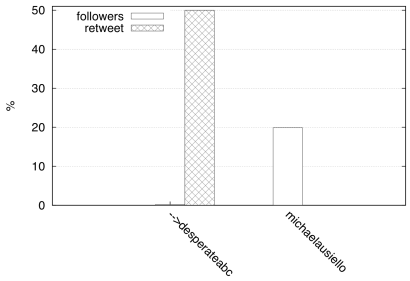
When combining the sets of features as explained above, we obtained even worse results, with the only exception of the *authoritativeness* features. The use of *retweets* and *followers* features alone allowed us to distinguish between official users, who have a high number of retweeted messages, and other verified but not institutional users, who to the contrary have a higher number of followers. This was somehow surprising, and was also verified when combining the two features together into the single set *authoritativeness*. This orthogonal behaviour is represented in Figure A.2, where the official users are in a separate cluster (represented is the centroid of the cluster, with the official users indicated with an arrow) due to their higher number of retweeted messages and other verified users are in another clusters because of their higher number of followers. When combining the normalized features together, we gave half weight to each of them, therefore they have a maximum value of 50% each. We were surprised because we were expecting the official users to have a higher number of both retweets and followers, while we discovered that other users are more followed than the official ones. This could be explained with the fact that generic users in Twitter tend to be



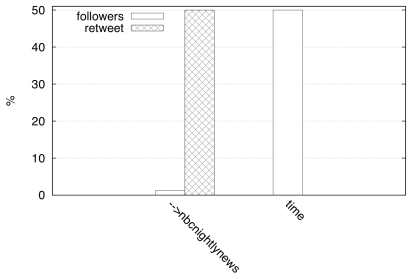

*(a)* Modern Family



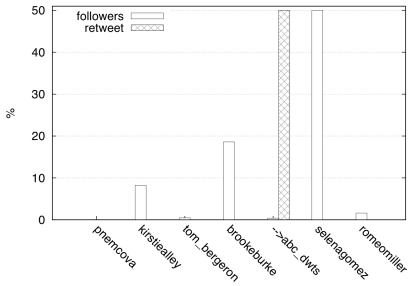

*(b)* Greys Anatomy



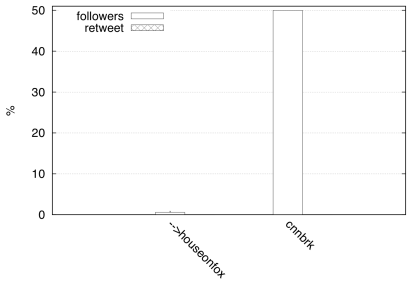

*(c)* Desperate Housewives




*(d)* NBC Nightly News




*(e)* Dancing with the stars




*(f)* Dr. House

*Figure A.2. Authoritativeness* features set for each TV show. Represented is the centroid of the cluster, with the official user indicated with an arrow (–> ). In combining the normalized features (*retweets* and *followers*) together, we gave half weight to each of them, therefore they have a maximum value of 50% each.

friends with the "coolest" or newest or most popular characters for each TV show (other than the official user), but they don't trust or don't think their messages are worth a retweet (lower number of retweets w.r.t. official user), thus these messages are less informative from a user point of view. It is also to be noted that the number of followers and retweets is computed among all the Twitter users and not just the subset of trusted (official+verified) we are considering in our analysis.

### Information content measurement

The second set of experiments we performed on the datasets aims at finding the most informative users among all the trusted ones. We followed a different approach than before and, instead of analyzing the features relative to each message, we considered the raw text of each message. We then concatenated all the messages of each trusted user per each show to build a single document that we later indexed with Lemur[7]. In doing this we obtained a per-show collection of documents, each representing a user.

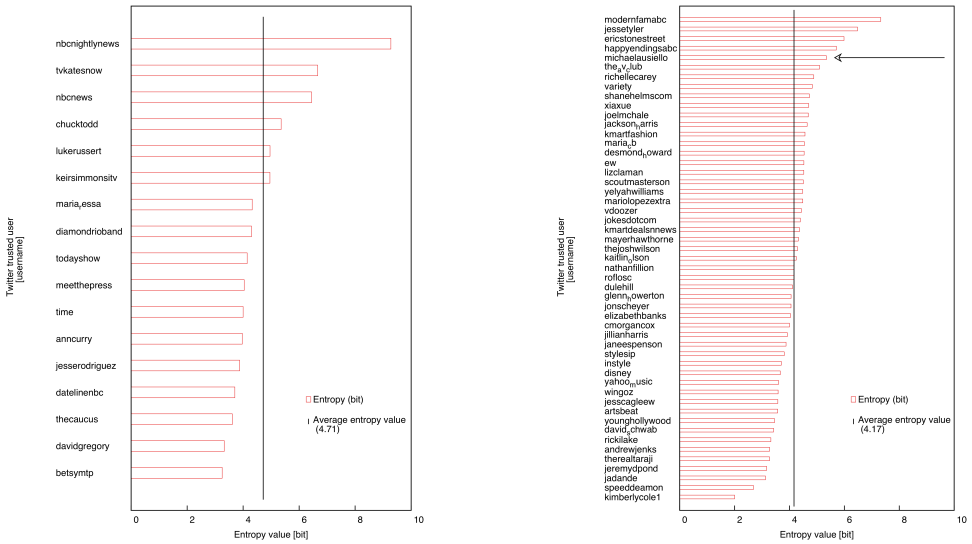
We used as a measure of entropy

$$E = -p(t|d)\log p(t|d)$$

where $p(t|d)$ is the term frequency in each document, to identify the most informative user. There are two reasons for doing this: the entropy captures the diversity in the content of a text and gives also its shorter representation [118]. The diversity in this context can be considered a measure of information richness: the topic of the messages is fixed (it is the TV show) and the diversity on a single topic can be seen as the great knowledge and expertise of the user on the single topic. Thus, the greater the entropy, the greater the diversity, so also the greater the expertise of a user and the informativeness of his messages.

This could be verified though the experiments: we discovered the most informative user is the official Twitter user of each TV program, the one having the highest entropy value. We represented as an example in Figure A.3 the entropy values for two TV shows, a news channel (NBC Nightly News, Figure A.3a) and a TV-series (Modern Family, Figure A.3b). The top users in each show (`nbcnightlynews` and `modernfamabc`, respectively) are exactly the official users we manually identified in the filtering block (Section A.2 and A.3) of our framework. In the same figure we also represented the average entropy value, which could be used as a separation border between the more informative (higher entropy level) and less informative streams (lower entropy level). If we look at the more informative users we can identify some similar or related to the official streams: `nbcnews` is clearly related with `nbcnightlynews` while `jessetyler` and `ericstonestreet` are actors playing in the Modern Family comedy. Immediately after these streams strongly related to TV shows, we found an interesting user: `michaelausiello` (highlighted with an arrow in Figure A.3b). This user is

[7] `http://www.lemurproject.org/`, a standard Information Retrieval Indexing software.




*(a)* NBC Nightly News *(b)* Modern Family

*Figure A.3.* Informative user discovery using entropy.

(according to his Twitter profile) "Founder and Editor-in-Chief [of] TVLine.com", an online website "to help TV enthusiasts cut through all the clutter and find a happy place where, more often than not, you will want to read most every post"[8]. In other words, he is an expert on TV shows and comedies. Besides his expertise, we considered this user interesting for two other reasons: we found him with high value of entropy also for other TV shows of the genre comedy or entertainment (and we did not know him before) and he was also discovered as centroid when running the clustering algorithm for the feature set *authoritativeness* (Figure A.2). This reinforces his position as expert and, having a high number of followers, makes him also a popular source. These are the properties an informative and valuable source should have, therefore it might be considered a new *official* but also optimal source to be later used in the synthesis block of our framework (Section A.2).

In our future work we are planning to implement in the synthesis block of our framework an unsupervisioned algorithm to automatically discover such interesting users within the trusted ones, eventually also including all the generic sources that were not yet considered because not verified. Once discovered, we can aggregate and use them, for example, to enrich the Electronic Program Guide of a TV content provider or in a recommender system for the end-user of the television set-up box.

[8] `http://www.TVLine.com/about-us/`

## A.5 Summary

In this study we presented a framework for the analysis of TV related Twitter messages (as a representative of a Social Network) from trusted sources and evaluated their informative value. We have extracted several features and characterized them in terms of quality and relevancy. From these features we were ablate identify an adaptation to the media in the writing style of the Social Media users, even from institutional or official sources. This makes the characterization of the trusted sources harder. We noticed, however, that the "social" component (retweets and followers features) plays an important role in helping to identify and characterize the most informative and valuable sources. Furthermore, from the analysis of the raw messages of each user (with a simple measure of entropy) we could easily detect the most expert and informative user given a show, but also discover new sources of information that could be used to complement the previous ones.

This work can be extended along different dimensions:

- cross validation of the findings by means of empirical evaluation (Mechanical Turk);
- introduction of other features in the analysis block, which include also sentiment detection and characterization of the sources [12];
- expansion of the analysis to a more heterogeneous collection of non-verified users;
- investigation on the time dimension (before, after and during a TV show);
- investigation of different and more advanced measures of entropy [66, 64] or others in general [43].

The latter one served as motivation for our studies of Author Identification for conversational documents (including Twitter, but not for SocialTV) in Chapter 5.

# Appendix B

# Sexual Predator Identification Approaches and Results

*I was never really told what to do. I think, looking back on it, that was a great precedent in my life, because he taught me to think that you could do things yourself without always checking up to see what the book said.*

Mavis Batey

In this appendix we provide the operative details and the results of the Sexual Predator Identification competition we organised in 2012 [56]. We already presented the competition in Chapter 2 (Section 2.4.1), while in Chapter 3 we introduced the collection we created as a testbed for the competition (Section 3.1.2). In the following Section B.1 we will describe the measures of performance employed in the competition and in Section B.2 we will give a general introduction of the approaches chosen by the participants, which we will comment in detail in Section B.3. We will summarise the achievements obtained within the framework of the competition in Section B.4.

We are going to briefly recall the problem of the Sexual Predator Identification competition, before providing the additional details. It is formulated as follow: given a collection of chat logs involving two (or more) persons, the participants in the competition had to:

1. Identify the sexual predators among all users in the different conversations (problem 1)

2. Identify the part (i.e. the lines) of the conversations which are the most distinctive of the predator behaviour (problem 2).

## B.1 Performance Measures

For the evaluation of the performance of the participants in the two problems, we referred to the standard Information Retrieval measure of Precision (P), Recall (R) and F (weighed harmonic mean between Precision and Recall):

$$\text{Precision (P)} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} \tag{B.1}$$

$$\text{Recall (R)} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} \tag{B.2}$$

$$\text{F} = \frac{1}{\alpha\frac{1}{\text{P}} + (1-\alpha)\frac{1}{\text{R}}} = \frac{(\beta^2+1)\text{PR}}{\beta^2\text{P}+\text{R}} \text{ where } \beta^2 = \frac{1-\alpha}{\alpha} \tag{B.3}$$

The retrieved "items" are in one case (problem 1, identify the predators) the ids of the authors considered perverted and in the second case (problem 2, identify the predators' lines) the line numbers considered indicative of a bad behaviour within a conversation. We also noticed that, while the standard F measure equally weighed P and R with $\beta$ equal to 1, this is not always desired. In our case, in fact, for the first problem, despite we observed that retrieving a lot of relevant authors is important (Recall), to facilitate the work of a police agent who would like to receive the largest number of suspects, what is more important is the fact that the retrieved authors are relevant (Precision). This to optimize the time the police agent dedicates to the "right" suspect, rather than "all" the possible suspects. For this reason we used a measure of F with the $\beta$ factor equal to 0.5, in order to emphasize Precision [81]. For the second problem, instead, we observed that retrieving a lot of relevant lines (Recall) is more important than finding only the relevant ones (Precision). Having a lot of relevant lines, in fact, augments the possibility of finding good evidence towards a suspect and for this reason we used a measure of F with the $\beta$ factor equal to 3, for emphasizing Recall [81].

It is also to be said that, while for the first problem the evaluation was quite straightforward having an a-priori indication of convicted perverted from the PJ website, for the second it was harder (and more discussed) to define the ground truth. We decided to adopt a TREC-like methodology for the evaluation and manually evaluated all submitted lines by at least one participant (this accounted for 91% of all the predators' lines). Given the particular nature of the task, that required a particular training for the evaluator in order to be able to distinguish between a predator chat and a regular chat, which could not be done in a distributed way (e.g. mechanical turk). Moreover given the limited time for the evaluation, we could not train other experts rather than us, thus relying on the evaluation of a single expert in our group. For this reason, evaluations contain a certain degree of subjectivity that we could not avoid. This is certainly a weak point in the 2012 competition.

## B.2 Overview of the Participants' Approaches

We received 16 submissions for the first problem (identifying the predators) and 14 for the second problem (identifying the distinctive chat lines of the predator behaviour) of the Sexual Predator Identification competition. Few users decided not to submit a notebook paper to explain their used methods, therefore we are presenting an analysis based on the 12 notebook papers received.

### Problem 1: identify predators

**Pre-filtering** For the first problem, where the participants had to return a list of potential predators, different pre-filtering techniques as well as classification methods have been applied. The collection given to the participants was by design very unbalanced (as most of them noticed) having few true positive authors (1% or less) in both training and testing datasets and containing a lot of false negatives that needed to be filtered out. A common approach to overcome this problem was the use of a two-stage classifier, where in the first stage the classifier had to distinguish between conversations involving a predator (true positive) and conversations without a predator (false negatives) [129, 87, 95, 50]. In addition to this, one of the most successful approaches [129] decides for the pre-filtering of all the conversations that manifested some particular patterns: presence of 1 participant only, those with less than 6 interventions per user or those that contained 3 long sequences of unrecognised characters. Similar attempts were done by other participants but with a rule-based approach and on different features for different approaches [92].

**Features** Apart from one case [97], where participants used machine learning approaches that work at character level (kernel with character 5-gram presence bit), in all the others submissions we can divide the used features into two main categories: "lexical" features and "behavioural" features. Lexical features are those that can be derived from the raw text of the conversation: examples of these features are unigram or bigram [129, 87, 34, 92], their weighting using TF-IDF or the cosine similarity and emoticons counting. Other examples are the name recognition of the participants in the conversation (self, other, group) [34] but also features obtained by the LIWC tool[1] that calculates the degree to which people use different categories of words across a wide array of texts [92, 127]. It is to be noted that, in general, lexical features have been used without any stemming or stopwords removal, to preserve each author's own style, including misspelling and grammatical errors.

Behavioural are all those features that capture the "actions" of a user within a conversation [50, 127]: the number of times a user starts a dialogue, the response time after a message of the partner in the conversation, the number of questions asked, the

[1] http://www.liwc.net/

frequency of turn-taking, intention (grooming, hooking, ...), etc. One of the most common approaches was the creation of a single set of features for each author, to be able to profile him and exploit his predator potential. Some participants decided to build up not just the Language Model (LM) of a single author, but also a LM as a combination of the LMs of the two participants in the chat [34]. Some other approaches were working, instead, at conversation or at line level, therefore participants that used these strategies had to aggregate the partial scores related to all the lines or conversations of an author to obtain a single set of features for each author [5, 71, 50, 95, 69].

**Classification approaches** In the classification step we could observe different proposed methods, but Support Vector Machines (SVM) were the most used [87, 92, 95, 129]. In general, they were used in most cases for the first (predator-vs-all), then also for the second step of the classification (predator-vs-victim). Sometimes participants found out that other solutions worked better than SVM, for example when they used a Neural Network classifier [129]. Other classifiers applied were based on Maximum-Entropy [34, 69], decision trees [71], k-NN [68, 97] and/or random forest [97] as well as Naïve Bayes [50, 5]. In combination with the classifier sometimes we observed a filtering approach based on a self-compiled dictionary of predatory terms.

To conclude, we shall notice that for this first problem we released a training set, which allowed for supervised algorithms to be easily used. The situation was different for the second problem, where no training data was available.

### Problem 2: identify predators' lines

For this second problem, no training data were available for the participants. This was intentionally done, mostly to test how participants approached the problem without an a-priori relevance.

The difficulty of the problem reduced the number of submissions (from 16 to 14) and obliged the participants to use different approaches, compared with the supervised ones of problem 1. The straightforward solution was to return, as relevant, all the conversations' lines of all the identified predators from the first problem [97]. One of the most used methods was the filtering of all the predator conversations through a dictionary of "perverted" terms or with a particular score (e.g. TF-IDF weighting) [95, 87, 92, 34]. Similar to this approach, another first computed the LMs of the part of the conversation considered predatory and then computed the differences between the actual conversation and the LMs [129]. To conclude, the last approach was simply to return those lines already labelled as predatory in the proposed algorithm by the default method for problem 1 (working at line level) [71, 50, 69].

## B.3 Evaluation Results of the Participants' Approaches and Discussion

As reported in Table A.1, participants received a training and a testing set, the first containing 142 users labelled as predators, the second containing 254 predators to be discovered. This was useful for the first problem (identify the predators), while for the second problem (identify the lines manifesting the predators bad behaviour) we did not release any training set. We wanted, in fact, to test how such a problem could be addressed without any evidence. We later evaluated manually all the 113,888 lines submitted by the participants and identified 6,478 that we considered expressions of a predator bad behaviour. In Table B.1 and Table B.2 we present the results for the first and second problem, with the measures of evaluation explained in Section B.1.

### Problem 1: identify predators

If we analyse in detail the results for the first problem, in particular the ranking in the case of the two different metrics F with $\beta = 1$ and F with $\beta = 0.5$, we will notice that only two positions swap ($1^{st}$ and $5^{th}$) in case we consider one or the other measure of F. This is due to the fact that we emphasised Precision with the F with $\beta = 0.5$. This choice did not encounter the favour of all the participants, in fact some manifested their disagreement and suggested giving more weight to Recall (thus, having an F measure with $\beta \geq 2$). In a real scenario, the proposed idea is to let the police agent decide who is a predator and "manually" filter the results automatically obtained. Another suggestion into this direction is the creation of a ranked list of suspects, which would help to prioritize the investigations.

Besides these issues, from an operational point of view, it is interesting to notice how important was the pre-filtering of unrelated conversations (at the cost of few true positive) [129] and the similar use of lexical features in all the first ranked approaches: bag-of-words with boolean weighting scheme [87, 129], unigrams with TF-IDF weighting scheme [92], unigram and bigram [34]. Participants also created a single profile for each author, by computing the features on an author-based file that collects all the posts/messages of that author [34, 92, 87]. Behavioural/conversational features were, on the other hand, used by all [34, 92, 87] except one [129] of the top-5 participants. This last one [129] also chose to use a Neural Network classifier instead of SVM (in both cases, two step classifiers) that were instead used by two others [92, 87], while others employed a Maximum-Entropy Classifier.

Despite the similar features used and the relatively closeness of the performance measures, the different classification strategies are a signal of still possible improvement possibilities in the problem.

### Problem 2: identify predators' lines

As mentioned before, problem 2 was more difficult than problem 1 and also presented more open-issues than problem 1. Despite the suggestion of giving more weight to Precision than to Recall, we should mention at least two issues that touched this part of the competition. The first one is a certain dependency from the first problem: identifying lines of the predator conversation requires at the beginning the correct identification of a good number of predators. This might disadvantage participants that performed poorly in the first part of the task. A solution to this problem might be having two stages for the competition that correspond to the two problems. The best result set of the first problem could be used as a starting point for the second task. It has to be noticed, however, that in the best-performers list (first-half of the ranking) we also find participants that were not in the top-5 of the first problem. A preliminary explanation for this is that few conversations of relatively few predators contribute to generate the ground truth for the predators' lines, therefore it is enough to identify such predators to obtain a good score for problem 2. This fact leads to a second issue for problem 2, the creation of the ground truth for the predators' lines. At the beginning of the competition, there was no ground truth for this second problem and we generated it on the basis of the received submissions. We could have generated the ground truth by analysing all the predators' conversations but by labelling only the submitted lines we spared 10% of all the conversations and approximatively 1 week of work time. The real issue was determined by the fact that one expert only labelled the lines of the conversation, leading to the exclusion of possibly relevant lines or the over-consideration of some others. We would have liked to have more experts (at least 2 or more) for labelling the relevant lines in all the predator conversations, but due to time and resource constraints that was not possible this year. For a future edition of the Sexual Predator Identification task, we should plan more time and resources for generating the ground truth and maybe we should consider the release of a training set for this part of the problem as well.

## B.4 Summary

We presented in this appendix the results of the first International Sexual Predator Identification Competition at PAN-2012 within CLEF 2012. Given a realistic and challenging collection containing chat logs involving two (or more) persons (introduced in Chapter 3), the 16 participants in the competition had to identify the predators among all the users in the different conversations and identify the part (the lines) of the predator conversations which were the most distinctive of the predator bad behaviour.

For the first problem we can conclude that lexical and behavioural features should be used when dealing with this kind of tasks. However, there is no single method to identify predators but different approaches could be used, from SVM to Maximum-

| Participant run | RETR. | REL. | P | R | $F_{\beta=1}$ | $F_{\beta=0.5}$ | Official run rank |
|---|---|---|---|---|---|---|---|
| villatorotello-run-2012-06-15-2157g | 204 | 200 | 0.9804 | 0.7874 | 0.8734 | 0.9346 | 1 |
| snider12-run-2012-06-16-0032 | 186 | 183 | 0.9839 | 0.7205 | 0.8318 | 0.9168 | 2 |
| *villatorotello-run-2012-06-15-2157c* | *211* | *200* | *0.9479* | *0.7874* | *0.8602* | *0.9107* | |
| parapar12-run-2012-06-15-0959j | 181 | 170 | 0.9392 | 0.6693 | 0.7816 | 0.8691 | 3 |
| morris12-run-2012-06-16-0752-main | 159 | 154 | 0.9686 | 0.6063 | 0.7458 | 0.8652 | 4 |
| eriksson12-run-2012-06-15-1949 | 265 | 227 | 0.8566 | 0.8937 | 0.8748 | 0.8638 | 5 |
| *parapar12-run-2012-06-15-0959g* | *171* | *162* | *0.9474* | *0.6378* | *0.7624* | *0.8635* | |
| *morris12-run-2012-06-17-0126* | *152* | *147* | *0.9671* | *0.5787* | *0.7241* | *0.8527* | |
| *parapar12-run-2012-06-15-0959i* | *173* | *161* | *0.9306* | *0.6339* | *0.7541* | *0.8510* | |
| *parapar12-run-2012-06-15-0959e* | *182* | *164* | *0.9011* | *0.6457* | *0.7523* | *0.8350* | |
| peersman12-run-2012-06-15-1559 | 170 | 152 | 0.8941 | 0.5984 | 0.7170 | 0.8137 | 6 |
| *parapar12-run-2012-06-15-0959d* | *175* | *151* | *0.8629* | *0.5945* | *0.7040* | *0.7914* | |
| *parapar12-run-2012-06-15-0959c* | *169* | *145* | *0.8580* | *0.5709* | *0.6856* | *0.7796* | |
| *villatorotello-run-2012-06-15-2157a* | *108* | *103* | *0.9537* | *0.4055* | *0.5691* | *0.7507* | |
| *parapar12-run-2012-06-15-0959b* | *205* | *160* | *0.7805* | *0.6299* | *0.6972* | *0.7449* | |
| grozea12-run-2012-06-14-1706b | 215 | 163 | 0.7581 | 0.6417 | 0.6951 | 0.7316 | 7 |
| *parapar12-run-2012-06-15-0959f* | *202* | *154* | *0.7624* | *0.6063* | *0.6754* | *0.7250* | |
| sitarz12-run-2012-0615-1515 | 218 | 159 | 0.7294 | 0.6260 | 0.6737 | 0.7060 | 8 |
| *parapar12-run-2012-06-15-0959h* | *223* | *161* | *0.7220* | *0.6339* | *0.6751* | *0.7024* | |
| *parapar12-run-2012-06-15-0959a* | *200* | *128* | *0.6400* | *0.5039* | *0.5639* | *0.6072* | |
| vartapetiance12-run-2012-06-15-1411 | 160 | 99 | 0.6188 | 0.3898 | 0.4783 | 0.5537 | 9 |
| *villatorotello-run-2012-06-15-2157f* | *269* | *143* | *0.5316* | *0.5630* | *0.5468* | *0.5376* | |
| *grozea12-run-2012-06-14-1706a* | *322* | *142* | *0.4410* | *0.5591* | *0.4931* | *0.4604* | |
| kontostathis-run-2012-06-16-0317e | 475 | 170 | 0.3579 | 0.6693 | 0.4664 | 0.3946 | 10 |
| *kontostathis-run-2012-06-16-0317d* | *688* | *172* | *0.2500* | *0.6772* | *0.3652* | *0.2861* | |
| kang12-run-2012-06-15-0904b | 930 | 203 | 0.2183 | 0.7992 | 0.3429 | 0.2554 | 11 |
| *kang12-run-2012-06-15-0904a* | *1049* | *202* | *0.1926* | *0.7953* | *0.3101* | *0.2270* | |
| kern12-run-2012-06-18-1827b | 1172 | 177 | 0.1510 | 0.6969 | 0.2482 | 0.1791 | 12 |
| *kern12-run-2012-06-18-1827a* | *1172* | *177* | *0.1510* | *0.6969* | *0.2482* | *0.1791* | |
| *villatorotello-run-2012-06-15-2157d* | *240* | *36* | *0.1500* | *0.1417* | *0.1457* | *0.1483* | |
| *kontostathis-run-2012-06-16-0317c* | *3696* | *206* | *0.0557* | *0.8110* | *0.1043* | *0.0685* | |
| *villatorotello-run-2012-06-15-2157b* | *204* | *12* | *0.0588* | *0.0472* | *0.0524* | *0.0561* | |
| *kontostathis-run-2012-06-16-0317a* | *5225* | *206* | *0.0394* | *0.8110* | *0.0752* | *0.0487* | |
| *kontostathis-run-2012-06-16-0317b* | *5625* | *221* | *0.0393* | *0.8701* | *0.0752* | *0.0486* | |
| *vilarino12-run-2012-06-14-2121a* | *9071* | *236* | *0.0260* | *0.9291* | *0.0506* | *0.0323* | |
| bogdanova12-run-2012-06-14-1117 | 2109 | 55 | 0.0261 | 0.2165 | 0.0466 | 0.0316 | 13 |
| prasath12-run-2012-06-15-2122 | 10289 | 207 | 0.0201 | 0.8150 | 0.0393 | 0.0250 | 14 |
| vilarino12-run-2012-06-14-2121b | 5225 | 98 | 0.0188 | 0.3858 | 0.0358 | 0.0232 | 15 |
| *villatorotello-run-2012-06-15-2157e* | *305* | *6* | *0.0197* | *0.0236* | *0.0215* | *0.0204* | |
| gomezhidalgo12-2012-06-15-1900 | 150 | 1 | 0.0067 | 0.0039 | 0.0050 | 0.0059 | 16 |

*Table B.1.* Results for problem 1): identify predators. The table reports the evaluation of all the runs submitted ordered by value of F score with $\beta = 0.5$. Runs with ranking number are the ones used for official evaluation. RET. = Retrieved documents, REL. = Relevant document retrieved. P = Precision. R = Recall.

| Participant run | RETR. | REL. | P | R | $F_{\beta=1}$ | $F_{\beta=3}$ | Official run rank |
|---|---|---|---|---|---|---|---|
| grozea12-run-2012-06-14-1706b | 63290 | 5790 | 0.0915 | 0.8938 | 0.1660 | 0.4762 | 1 |
| kontostathis-run-2012-06-16-0317e | 19535 | 3249 | 0.1663 | 0.5015 | 0.2498 | 0.4174 | 2 |
| peersman12-run-2012-06-15-1559 | 4717 | 1688 | 0.3579 | 0.2606 | 0.3016 | 0.2679 | 3 |
| sitarz12-run-2012-0615-1515 | 4558 | 1486 | 0.3260 | 0.2294 | 0.2693 | 0.2364 | 4 |
| morris12-run-2012-06-16-0752-main | 2685 | 1211 | 0.4510 | 0.1869 | 0.2643 | 0.1986 | 5 |
| kern12-run-2012-06-18-1827b | 15533 | 1357 | 0.0874 | 0.2095 | 0.1233 | 0.1838 | 6 |
| eriksson12-run-2012-06-15-1949 | 10416 | 1122 | 0.1077 | 0.1732 | 0.1328 | 0.1633 | 7 |
| prasath12-run-2012-06-15-2122 | 77255 | 1044 | 0.0135 | 0.1612 | 0.0249 | 0.0770 | 8 |
| parapar12-run-2012-06-15-0959j | 2037 | 105 | 0.0515 | 0.0162 | 0.0247 | 0.0174 | 9 |
| vartapetiance12-run-2012-06-15-1411 | 607 | 91 | 0.1499 | 0.0140 | 0.0257 | 0.0154 | 10 |
| vilarino12-run-2012-06-14-2121b | 6787 | 48 | 0.0071 | 0.0074 | 0.0072 | 0.0074 | 11 |
| bogdanova12-run-2012-06-14-1117 | 49 | 4 | 0.0816 | 0.0006 | 0.0012 | 0.0007 | 12 |
| villatorotello-run-2012-06-15-2157g | 50 | 1 | 0.0200 | 0.0002 | 0.0003 | 0.0002 | 13 |
| gomezhidalgo12-2012-06-15-1900 | 400 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 14 |

*Table B.2.* Results for problem 2): identify predators' lines. The table reports the evaluation of all the runs submitted ordered by value of F score with $\beta = 3$. RET. = Retrieved documents, REL. = Relevant document retrieved. P = Precision. R = Recall.

Entropy algorithm. Having a pre-filtering step to prune irrelevant conversations seems an important addition to the systems. For the second problem the most effective methods appeared to be those based on filtering on a dictionary or LM basis, partly due to the lack of ground truth for this specific problem (if we exclude the one based on 5-gram characters presence bit). The identification of a common set of features and a group of effective strategies to identify predators is an achievement for this first part of the task.

During the competition some issues were raised about the measurement of performances for the two problems, whether we should emphasise Precision or Recall and about the degree of subjectivity in the creation of the ground truth for problem 2. This is an achievement, too: with this competition we wanted to give researchers a unique place for comparing their methods but also for discussing and debating future directions on this research area.

# Appendix C

# Tabular results

*They say you got to stay hungry hey*
*baby I'm just about starving tonight*
*I'm dying for some action I'm sick of*
*sitting 'round here trying to write*
*This book*
*I need a love reaction come on baby*
*give me just one look*

Bruce Springsteen

In this appendix we are reporting the values for the results of the experiments performed. The values in the following tables match those represented in the figures of Section 5.5.2 of Chapter 5. For collections of newspaper articles (Associated Press, La Stampa, Glasgow Herald) we were only able to compute values for method x0, due to the absence of conversations in these collections. For collections of conversational documents (Freenode, Krijn, Twitter) we are reporting values for all the three methods (x0, x1, x2) and the $p$-values obtained performing the Wilcoxon (paired) statistical test. We tested at $p < 0.05$ and expressed the alternative hypothesis *"are results of method z statistically better than results of method y?"* with the expression "z $\gg$ y".

| Voc. % | method x0 |
|---|---|
| 5 | 0.9889 |
| 10 | 0.9833 |
| 15 | 0.9833 |
| 20 | 0.9833 |
| 25 | 0.9833 |
| 30 | 0.9833 |
| 35 | 0.9833 |
| 40 | 0.9833 |
| 45 | 0.9833 |
| 50 | 0.9833 |
| 55 | 0.9833 |
| 60 | 0.9917 |
| 65 | 0.9833 |
| 70 | 0.9833 |
| 75 | 0.9917 |
| 80 | 0.9917 |
| 85 | 0.9917 |
| 90 | 0.9833 |
| 95 | 0.9750 |
| 100 | 0.9917 |

*(a)* KLD terms scoring, KLD classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.7194 |
| 10 | 0.7973 |
| 15 | 0.7822 |
| 20 | 0.7815 |
| 25 | 0.7814 |
| 30 | 0.7878 |
| 35 | 0.7798 |
| 40 | 0.7795 |
| 45 | 0.7797 |
| 50 | 0.7797 |
| 55 | 0.7797 |
| 60 | 0.7797 |
| 65 | 0.7800 |
| 70 | 0.7800 |
| 75 | 0.7731 |
| 80 | 0.7731 |
| 85 | 0.7601 |
| 90 | 0.7514 |
| 95 | 0.7664 |
| 100 | 1.0000 |

*(b)* KLD terms scoring, $\chi^2$ classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.2403 |
| 10 | 0.2340 |
| 15 | 0.2293 |
| 20 | 0.2315 |
| 25 | 0.2413 |
| 30 | 0.2434 |
| 35 | 0.2643 |
| 40 | 0.2656 |
| 45 | 0.2676 |
| 50 | 0.2702 |
| 55 | 0.2684 |
| 60 | 0.2684 |
| 65 | 0.2678 |
| 70 | 0.2668 |
| 75 | 0.2662 |
| 80 | 0.2652 |
| 85 | 0.2614 |
| 90 | 0.2583 |
| 95 | 0.2561 |
| 100 | 0.2515 |

*(c)* KLD terms scoring, Delta classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.9694 |
| 10 | 1.0000 |
| 15 | 1.0000 |
| 20 | 1.0000 |
| 25 | 1.0000 |
| 30 | 1.0000 |
| 35 | 1.0000 |
| 40 | 1.0000 |
| 45 | 1.0000 |
| 50 | 1.0000 |
| 55 | 1.0000 |
| 60 | 1.0000 |
| 65 | 1.0000 |
| 70 | 1.0000 |
| 75 | 1.0000 |
| 80 | 1.0000 |
| 85 | 1.0000 |
| 90 | 1.0000 |
| 95 | 0.7831 |
| 100 | 0.9917 |

*(d)* TF-IDF terms scoring, KLD classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.5474 |
| 10 | 0.4278 |
| 15 | 0.4187 |
| 20 | 0.4133 |
| 25 | 0.4173 |
| 30 | 0.4134 |
| 35 | 0.4083 |
| 40 | 0.4140 |
| 45 | 0.4076 |
| 50 | 0.4040 |
| 55 | 0.4005 |
| 60 | 0.3971 |
| 65 | 0.3926 |
| 70 | 0.3824 |
| 75 | 0.3727 |
| 80 | 0.3583 |
| 85 | 0.3415 |
| 90 | 0.3281 |
| 95 | 0.3707 |
| 100 | 1.0000 |

*(e)* TF-IDF terms scoring, $\chi^2$ classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.3073 |
| 10 | 0.2480 |
| 15 | 0.2320 |
| 20 | 0.2356 |
| 25 | 0.2642 |
| 30 | 0.2742 |
| 35 | 0.2686 |
| 40 | 0.2726 |
| 45 | 0.2745 |
| 50 | 0.2743 |
| 55 | 0.2695 |
| 60 | 0.2664 |
| 65 | 0.2650 |
| 70 | 0.2636 |
| 75 | 0.2611 |
| 80 | 0.2611 |
| 85 | 0.2586 |
| 90 | 0.2564 |
| 95 | 0.2529 |
| 100 | 0.2515 |

*(f)* TF-IDF terms scoring, Delta classifier

*Table C.1.* Tabular Results: 20 users, Associated Press

| Voc. % | method x0 |
|---|---|
| 5 | 1.0000 |
| 10 | 1.0000 |
| 15 | 1.0000 |
| 20 | 1.0000 |
| 25 | 1.0000 |
| 30 | 0.9917 |
| 35 | 0.9917 |
| 40 | 1.0000 |
| 45 | 1.0000 |
| 50 | 1.0000 |
| 55 | 1.0000 |
| 60 | 1.0000 |
| 65 | 1.0000 |
| 70 | 1.0000 |
| 75 | 1.0000 |
| 80 | 1.0000 |
| 85 | 1.0000 |
| 90 | 1.0000 |
| 95 | 1.0000 |
| 100 | 1.0000 |

*(a)* KLD terms scoring, KLD classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.8238 |
| 10 | 0.7738 |
| 15 | 0.7835 |
| 20 | 0.7752 |
| 25 | 0.7566 |
| 30 | 0.7637 |
| 35 | 0.7831 |
| 40 | 0.7826 |
| 45 | 0.7834 |
| 50 | 0.7701 |
| 55 | 0.7706 |
| 60 | 0.7790 |
| 65 | 0.7873 |
| 70 | 0.7948 |
| 75 | 0.7948 |
| 80 | 0.7948 |
| 85 | 0.7945 |
| 90 | 0.7820 |
| 95 | 0.7709 |
| 100 | 1.0000 |

*(b)* KLD terms scoring, $\chi^2$ classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.2283 |
| 10 | 0.2149 |
| 15 | 0.2062 |
| 20 | 0.2076 |
| 25 | 0.2096 |
| 30 | 0.2025 |
| 35 | 0.2164 |
| 40 | 0.2206 |
| 45 | 0.2204 |
| 50 | 0.2209 |
| 55 | 0.2189 |
| 60 | 0.2263 |
| 65 | 0.2258 |
| 70 | 0.2250 |
| 75 | 0.2245 |
| 80 | 0.2242 |
| 85 | 0.2235 |
| 90 | 0.2232 |
| 95 | 0.2216 |
| 100 | 0.2197 |

*(c)* KLD terms scoring, Delta classifier

| Voc. % | method x0 |
|---|---|
| 5 | 1.0000 |
| 10 | 1.0000 |
| 15 | 1.0000 |
| 20 | 1.0000 |
| 25 | 1.0000 |
| 30 | 1.0000 |
| 35 | 1.0000 |
| 40 | 1.0000 |
| 45 | 1.0000 |
| 50 | 1.0000 |
| 55 | 1.0000 |
| 60 | 1.0000 |
| 65 | 1.0000 |
| 70 | 1.0000 |
| 75 | 1.0000 |
| 80 | 1.0000 |
| 85 | 1.0000 |
| 90 | 1.0000 |
| 95 | 0.9292 |
| 100 | 1.0000 |

*(d)* TF-IDF terms scoring, KLD classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.8273 |
| 10 | 0.6187 |
| 15 | 0.5490 |
| 20 | 0.5798 |
| 25 | 0.5935 |
| 30 | 0.5924 |
| 35 | 0.5709 |
| 40 | 0.5651 |
| 45 | 0.5532 |
| 50 | 0.5488 |
| 55 | 0.5554 |
| 60 | 0.5279 |
| 65 | 0.4932 |
| 70 | 0.4702 |
| 75 | 0.4206 |
| 80 | 0.4114 |
| 85 | 0.3970 |
| 90 | 0.3602 |
| 95 | 0.3722 |
| 100 | 1.0000 |

*(e)* TF-IDF terms scoring, $\chi^2$ classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.2918 |
| 10 | 0.2323 |
| 15 | 0.2085 |
| 20 | 0.2123 |
| 25 | 0.2202 |
| 30 | 0.2148 |
| 35 | 0.2167 |
| 40 | 0.2215 |
| 45 | 0.2211 |
| 50 | 0.2183 |
| 55 | 0.2267 |
| 60 | 0.2257 |
| 65 | 0.2250 |
| 70 | 0.2246 |
| 75 | 0.2246 |
| 80 | 0.2237 |
| 85 | 0.2231 |
| 90 | 0.2220 |
| 95 | 0.2213 |
| 100 | 0.2197 |

*(f)* TF-IDF terms scoring, Delta classifier

*Table C.2.* Tabular Results: 20 users, La Stampa

| Voc. % | method x0 |
|---|---|
| 5 | 0.9875 |
| 10 | 0.9889 |
| 15 | 0.9889 |
| 20 | 0.9889 |
| 25 | 0.9889 |
| 30 | 0.9917 |
| 35 | 1.0000 |
| 40 | 1.0000 |
| 45 | 1.0000 |
| 50 | 1.0000 |
| 55 | 1.0000 |
| 60 | 1.0000 |
| 65 | 1.0000 |
| 70 | 1.0000 |
| 75 | 1.0000 |
| 80 | 1.0000 |
| 85 | 1.0000 |
| 90 | 1.0000 |
| 95 | 1.0000 |
| 100 | 1.0000 |

*(a)* KLD terms scoring, KLD classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.7356 |
| 10 | 0.7361 |
| 15 | 0.7277 |
| 20 | 0.7325 |
| 25 | 0.7325 |
| 30 | 0.7325 |
| 35 | 0.7325 |
| 40 | 0.7333 |
| 45 | 0.7431 |
| 50 | 0.7431 |
| 55 | 0.7431 |
| 60 | 0.7417 |
| 65 | 0.7431 |
| 70 | 0.7431 |
| 75 | 0.7450 |
| 80 | 0.7460 |
| 85 | 0.7391 |
| 90 | 0.7516 |
| 95 | 0.7408 |
| 100 | 0.9833 |

*(b)* KLD terms scoring, $\chi^2$ classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.2153 |
| 10 | 0.2084 |
| 15 | 0.2101 |
| 20 | 0.2127 |
| 25 | 0.2096 |
| 30 | 0.2198 |
| 35 | 0.2247 |
| 40 | 0.2245 |
| 45 | 0.2304 |
| 50 | 0.2327 |
| 55 | 0.2333 |
| 60 | 0.2340 |
| 65 | 0.2334 |
| 70 | 0.2300 |
| 75 | 0.2294 |
| 80 | 0.2284 |
| 85 | 0.2282 |
| 90 | 0.2266 |
| 95 | 0.2242 |
| 100 | 0.2198 |

*(c)* TF-IDF terms scoring, Delta classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.9917 |
| 10 | 0.9667 |
| 15 | 0.9667 |
| 20 | 0.9750 |
| 25 | 0.9917 |
| 30 | 0.9917 |
| 35 | 1.0000 |
| 40 | 1.0000 |
| 45 | 1.0000 |
| 50 | 1.0000 |
| 55 | 1.0000 |
| 60 | 1.0000 |
| 65 | 1.0000 |
| 70 | 1.0000 |
| 75 | 1.0000 |
| 80 | 1.0000 |
| 85 | 1.0000 |
| 90 | 0.9917 |
| 95 | 0.9607 |
| 100 | 1.0000 |

*(d)* TF-IDF terms scoring, KLD classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.5263 |
| 10 | 0.3699 |
| 15 | 0.3113 |
| 20 | 0.3013 |
| 25 | 0.3150 |
| 30 | 0.3148 |
| 35 | 0.3264 |
| 40 | 0.3170 |
| 45 | 0.3209 |
| 50 | 0.3230 |
| 55 | 0.3195 |
| 60 | 0.3097 |
| 65 | 0.2953 |
| 70 | 0.2869 |
| 75 | 0.2748 |
| 80 | 0.2550 |
| 85 | 0.2442 |
| 90 | 0.2226 |
| 95 | 0.2450 |
| 100 | 0.9833 |

*(e)* TF-IDF terms scoring, $\chi^2$ classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.2950 |
| 10 | 0.2113 |
| 15 | 0.2049 |
| 20 | 0.2113 |
| 25 | 0.2197 |
| 30 | 0.2238 |
| 35 | 0.2278 |
| 40 | 0.2324 |
| 45 | 0.2334 |
| 50 | 0.2348 |
| 55 | 0.2334 |
| 60 | 0.2298 |
| 65 | 0.2275 |
| 70 | 0.2266 |
| 75 | 0.2264 |
| 80 | 0.2280 |
| 85 | 0.2273 |
| 90 | 0.2257 |
| 95 | 0.2239 |
| 100 | 0.2198 |

*(f)* TF-IDF terms scoring, Delta classifier

*Table C.3.* Tabular Results: 20 users, Glasgow Herald

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.8112 | 0.9431 | 0.9264 | 0.000 | 0.006 | 0.771 | 1.000 | 0.994 |
| 10 | 0.8583 | 0.9482 | 0.9417 | 0.007 | 0.011 | 0.669 | 0.994 | 0.990 |
| 15 | 0.8549 | 0.9639 | 0.9167 | 0.003 | 0.058 | 0.943 | 0.997 | 0.948 |
| 20 | 0.8711 | 0.9542 | 0.9292 | 0.010 | 0.091 | 0.810 | 0.992 | 0.917 |
| 25 | 0.8542 | 0.9833 | 0.9347 | 0.001 | 0.012 | 0.978 | 0.999 | 0.990 |
| 30 | 0.8362 | 0.9833 | 0.9583 | 0.001 | 0.001 | 0.932 | 1.000 | 0.999 |
| 35 | 0.8362 | 0.9833 | 0.9667 | 0.001 | 0.001 | 0.885 | 1.000 | 0.999 |
| 40 | 0.8396 | 0.9833 | 0.9556 | 0.001 | 0.002 | 0.981 | 1.000 | 0.999 |
| 45 | 0.8562 | 0.9917 | 0.9617 | 0.001 | 0.003 | 0.972 | 0.999 | 0.997 |
| 50 | 0.8872 | 0.9833 | 0.9458 | 0.006 | 0.023 | 0.962 | 0.996 | 0.982 |
| 55 | 0.8608 | 0.9833 | 0.9708 | 0.002 | 0.003 | 0.862 | 0.999 | 0.998 |
| 60 | 0.8686 | 0.9833 | 0.9722 | 0.003 | 0.003 | 0.862 | 0.998 | 0.998 |
| 65 | 0.8888 | 0.9667 | 0.9639 | 0.014 | 0.010 | 0.715 | 0.988 | 0.992 |
| 70 | 0.9082 | 0.9833 | 0.9639 | 0.014 | 0.037 | 0.963 | 0.990 | 0.975 |
| 75 | 0.9194 | 0.9833 | 0.9722 | 0.024 | 0.053 | 0.862 | 0.983 | 0.961 |
| 80 | 0.9083 | 0.9917 | 0.9639 | 0.010 | 0.040 | 0.972 | 0.993 | 0.971 |
| 85 | 0.9000 | 0.9833 | 0.9639 | 0.010 | 0.025 | 0.963 | 0.993 | 0.981 |
| 90 | 0.8985 | 0.9750 | 0.9617 | 0.012 | 0.026 | 0.963 | 0.991 | 0.980 |
| 95 | 0.9057 | 0.9750 | 0.9533 | 0.018 | 0.055 | 0.972 | 0.988 | 0.957 |
| 100 | 0.5675 | 0.7051 | 0.6792 | 0.000 | 0.000 | 0.990 | 1.000 | 1.000 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.8303 | 0.9464 | 0.8599 | 0.005 | 0.092 | 0.988 | 0.995 | 0.916 |
| 10 | 0.8131 | 0.9228 | 0.8686 | 0.012 | 0.050 | 0.909 | 0.989 | 0.954 |
| 15 | 0.8533 | 0.9403 | 0.9283 | 0.003 | 0.003 | 0.738 | 0.997 | 0.997 |
| 20 | 0.8554 | 0.9417 | 0.9333 | 0.007 | 0.001 | 0.725 | 0.994 | 0.999 |
| 25 | 0.8176 | 0.9139 | 0.9056 | 0.002 | 0.000 | 0.807 | 0.999 | 1.000 |
| 30 | 0.7838 | 0.9111 | 0.8806 | 0.000 | 0.001 | 0.986 | 1.000 | 1.000 |
| 35 | 0.7512 | 0.9000 | 0.8861 | 0.000 | 0.000 | 0.861 | 1.000 | 1.000 |
| 40 | 0.7435 | 0.8625 | 0.8583 | 0.000 | 0.000 | 0.707 | 1.000 | 1.000 |
| 45 | 0.7236 | 0.8528 | 0.8458 | 0.000 | 0.000 | 0.913 | 1.000 | 1.000 |
| 50 | 0.7281 | 0.8389 | 0.8339 | 0.001 | 0.000 | 0.819 | 1.000 | 1.000 |
| 55 | 0.7052 | 0.8417 | 0.8186 | 0.000 | 0.000 | 0.915 | 1.000 | 1.000 |
| 60 | 0.7117 | 0.8444 | 0.8006 | 0.000 | 0.004 | 0.993 | 1.000 | 0.996 |
| 65 | 0.7236 | 0.8528 | 0.8269 | 0.000 | 0.000 | 0.964 | 1.000 | 1.000 |
| 70 | 0.7497 | 0.8889 | 0.8478 | 0.001 | 0.002 | 0.967 | 0.999 | 0.998 |
| 75 | 0.7668 | 0.8944 | 0.8917 | 0.000 | 0.000 | 0.601 | 1.000 | 1.000 |
| 80 | 0.8246 | 0.9319 | 0.9417 | 0.003 | 0.002 | 0.331 | 0.998 | 0.998 |
| 85 | 0.8281 | 0.9639 | 0.9708 | 0.000 | 0.000 | 0.395 | 1.000 | 1.000 |
| 90 | 0.8647 | 0.9242 | 0.9319 | 0.070 | 0.053 | 0.365 | 0.937 | 0.954 |
| 95 | 0.7274 | 0.8962 | 0.7826 | 0.001 | 0.112 | 0.995 | 0.999 | 0.892 |
| 100 | 0.5675 | 0.7051 | 0.6792 | 0.000 | 0.000 | 0.990 | 1.000 | 1.000 |

*(b)* TF-IDF terms scoring

*Table C.4.* Tabular Results: 20 users, Freenode, KLD classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.3760 | 0.5715 | 0.5128 | 0.000 | 0.001 | 0.890 | 1.000 | 0.999 |
| 10 | 0.3554 | 0.5192 | 0.5160 | 0.000 | 0.000 | 0.567 | 1.000 | 1.000 |
| 15 | 0.3533 | 0.5101 | 0.4691 | 0.000 | 0.002 | 0.914 | 1.000 | 0.999 |
| 20 | 0.3609 | 0.4994 | 0.4818 | 0.002 | 0.003 | 0.738 | 0.999 | 0.997 |
| 25 | 0.3534 | 0.5046 | 0.4879 | 0.000 | 0.001 | 0.614 | 1.000 | 0.999 |
| 30 | 0.3663 | 0.4913 | 0.4971 | 0.001 | 0.003 | 0.373 | 0.999 | 0.997 |
| 35 | 0.3661 | 0.4928 | 0.4699 | 0.000 | 0.004 | 0.639 | 1.000 | 0.997 |
| 40 | 0.3638 | 0.5004 | 0.4911 | 0.000 | 0.001 | 0.723 | 1.000 | 0.999 |
| 45 | 0.3633 | 0.4896 | 0.4884 | 0.000 | 0.001 | 0.653 | 1.000 | 0.999 |
| 50 | 0.3678 | 0.4952 | 0.4891 | 0.000 | 0.004 | 0.630 | 1.000 | 0.997 |
| 55 | 0.3701 | 0.4943 | 0.4810 | 0.001 | 0.004 | 0.693 | 0.999 | 0.996 |
| 60 | 0.3648 | 0.4982 | 0.4928 | 0.000 | 0.001 | 0.574 | 1.000 | 0.999 |
| 65 | 0.3781 | 0.5094 | 0.5026 | 0.000 | 0.001 | 0.675 | 1.000 | 0.999 |
| 70 | 0.3826 | 0.4864 | 0.4858 | 0.000 | 0.004 | 0.596 | 1.000 | 0.996 |
| 75 | 0.3903 | 0.4861 | 0.4997 | 0.003 | 0.007 | 0.356 | 0.997 | 0.993 |
| 80 | 0.3898 | 0.4844 | 0.4908 | 0.003 | 0.009 | 0.569 | 0.998 | 0.991 |
| 85 | 0.3914 | 0.4846 | 0.4945 | 0.009 | 0.023 | 0.570 | 0.992 | 0.978 |
| 90 | 0.3809 | 0.4757 | 0.5069 | 0.010 | 0.013 | 0.231 | 0.991 | 0.987 |
| 95 | 0.3676 | 0.4771 | 0.5039 | 0.006 | 0.009 | 0.337 | 0.994 | 0.992 |
| 100 | 0.9056 | 0.9444 | 0.9542 | 0.060 | 0.023 | 0.285 | 0.952 | 0.983 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.4721 | 0.6235 | 0.5718 | 0.000 | 0.000 | 0.976 | 1.000 | 1.000 |
| 10 | 0.3935 | 0.5191 | 0.5140 | 0.000 | 0.000 | 0.894 | 1.000 | 1.000 |
| 15 | 0.3523 | 0.5196 | 0.4987 | 0.000 | 0.000 | 0.978 | 1.000 | 1.000 |
| 20 | 0.3318 | 0.4725 | 0.4663 | 0.000 | 0.000 | 0.947 | 1.000 | 1.000 |
| 25 | 0.3187 | 0.4289 | 0.4082 | 0.000 | 0.001 | 0.998 | 1.000 | 1.000 |
| 30 | 0.3081 | 0.4353 | 0.4017 | 0.000 | 0.000 | 0.999 | 1.000 | 1.000 |
| 35 | 0.3074 | 0.4361 | 0.3779 | 0.000 | 0.003 | 1.000 | 1.000 | 0.997 |
| 40 | 0.2880 | 0.4176 | 0.3799 | 0.000 | 0.001 | 0.993 | 1.000 | 1.000 |
| 45 | 0.2703 | 0.4020 | 0.3665 | 0.000 | 0.000 | 0.978 | 1.000 | 1.000 |
| 50 | 0.2781 | 0.3828 | 0.3597 | 0.000 | 0.002 | 0.953 | 1.000 | 0.999 |
| 55 | 0.2608 | 0.3669 | 0.3506 | 0.000 | 0.000 | 0.915 | 1.000 | 1.000 |
| 60 | 0.2685 | 0.3779 | 0.3525 | 0.000 | 0.000 | 0.998 | 1.000 | 1.000 |
| 65 | 0.2610 | 0.3635 | 0.3463 | 0.000 | 0.000 | 0.997 | 1.000 | 1.000 |
| 70 | 0.2679 | 0.3617 | 0.3448 | 0.000 | 0.002 | 0.995 | 1.000 | 0.998 |
| 75 | 0.2661 | 0.3594 | 0.3437 | 0.000 | 0.000 | 0.996 | 1.000 | 1.000 |
| 80 | 0.2658 | 0.3580 | 0.3238 | 0.000 | 0.007 | 0.998 | 1.000 | 0.994 |
| 85 | 0.2682 | 0.4215 | 0.3801 | 0.000 | 0.001 | 0.999 | 1.000 | 0.999 |
| 90 | 0.2626 | 0.4476 | 0.3557 | 0.000 | 0.001 | 1.000 | 1.000 | 0.999 |
| 95 | 0.3128 | 0.3931 | 0.3665 | 0.099 | 0.191 | 0.776 | 0.903 | 0.813 |
| 100 | 0.9056 | 0.9444 | 0.9542 | 0.060 | 0.023 | 0.285 | 0.952 | 0.983 |

*(b)* TF-IDF terms scoring

*Table C.5.* Tabular Results: 20 users, Freenode, $\chi^2$ classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.2491 | 0.2987 | 0.2846 | 0.001 | 0.005 | 0.732 | 1.000 | 0.995 |
| 10 | 0.2305 | 0.2048 | 0.2043 | 0.253 | 0.334 | 0.664 | 0.751 | 0.671 |
| 15 | 0.2324 | 0.2099 | 0.1996 | 0.604 | 0.370 | 0.539 | 0.401 | 0.636 |
| 20 | 0.2371 | 0.2118 | 0.1982 | 0.418 | 0.730 | 0.798 | 0.587 | 0.276 |
| 25 | 0.2121 | 0.2192 | 0.2289 | 0.435 | 0.110 | 0.171 | 0.570 | 0.893 |
| 30 | 0.2111 | 0.2264 | 0.2258 | 0.356 | 0.216 | 0.352 | 0.650 | 0.788 |
| 35 | 0.2083 | 0.2268 | 0.2222 | 0.452 | 0.401 | 0.477 | 0.554 | 0.605 |
| 40 | 0.2086 | 0.2252 | 0.2215 | 0.552 | 0.428 | 0.430 | 0.454 | 0.578 |
| 45 | 0.2079 | 0.2237 | 0.2188 | 0.449 | 0.370 | 0.528 | 0.557 | 0.636 |
| 50 | 0.2045 | 0.2235 | 0.2188 | 0.334 | 0.248 | 0.581 | 0.671 | 0.756 |
| 55 | 0.2047 | 0.2313 | 0.2184 | 0.283 | 0.348 | 0.703 | 0.721 | 0.658 |
| 60 | 0.2044 | 0.2395 | 0.2183 | 0.174 | 0.295 | 0.846 | 0.830 | 0.709 |
| 65 | 0.2043 | 0.2395 | 0.2182 | 0.161 | 0.311 | 0.847 | 0.842 | 0.694 |
| 70 | 0.2043 | 0.2395 | 0.2182 | 0.161 | 0.311 | 0.847 | 0.842 | 0.694 |
| 75 | 0.2043 | 0.2395 | 0.2182 | 0.161 | 0.311 | 0.847 | 0.842 | 0.694 |
| 80 | 0.2045 | 0.2395 | 0.2182 | 0.184 | 0.356 | 0.847 | 0.820 | 0.649 |
| 85 | 0.2042 | 0.2395 | 0.2182 | 0.142 | 0.291 | 0.847 | 0.860 | 0.714 |
| 90 | 0.2043 | 0.2395 | 0.2181 | 0.144 | 0.293 | 0.847 | 0.859 | 0.712 |
| 95 | 0.2043 | 0.2395 | 0.2181 | 0.144 | 0.293 | 0.847 | 0.859 | 0.712 |
| 100 | 0.2043 | 0.2395 | 0.2181 | 0.144 | 0.293 | 0.847 | 0.859 | 0.712 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.3005 | 0.3980 | 0.3599 | 0.000 | 0.000 | 0.993 | 1.000 | 1.000 |
| 10 | 0.2562 | 0.3272 | 0.3104 | 0.000 | 0.000 | 0.906 | 1.000 | 1.000 |
| 15 | 0.2404 | 0.2747 | 0.2636 | 0.002 | 0.006 | 0.776 | 0.998 | 0.995 |
| 20 | 0.2123 | 0.2500 | 0.2512 | 0.004 | 0.015 | 0.532 | 0.996 | 0.986 |
| 25 | 0.2049 | 0.2210 | 0.2241 | 0.223 | 0.166 | 0.693 | 0.781 | 0.838 |
| 30 | 0.2050 | 0.2327 | 0.2212 | 0.380 | 0.373 | 0.811 | 0.625 | 0.632 |
| 35 | 0.2054 | 0.2344 | 0.2215 | 0.442 | 0.442 | 0.603 | 0.564 | 0.565 |
| 40 | 0.2050 | 0.2235 | 0.2186 | 0.440 | 0.386 | 0.547 | 0.566 | 0.620 |
| 45 | 0.2048 | 0.2220 | 0.2187 | 0.454 | 0.341 | 0.516 | 0.552 | 0.665 |
| 50 | 0.2056 | 0.2304 | 0.2187 | 0.400 | 0.402 | 0.688 | 0.605 | 0.604 |
| 55 | 0.2052 | 0.2311 | 0.2183 | 0.328 | 0.384 | 0.723 | 0.677 | 0.622 |
| 60 | 0.2047 | 0.2395 | 0.2183 | 0.215 | 0.350 | 0.846 | 0.790 | 0.655 |
| 65 | 0.2045 | 0.2395 | 0.2183 | 0.190 | 0.320 | 0.846 | 0.814 | 0.685 |
| 70 | 0.2043 | 0.2395 | 0.2182 | 0.167 | 0.311 | 0.817 | 0.837 | 0.694 |
| 75 | 0.2042 | 0.2381 | 0.2181 | 0.160 | 0.299 | 0.847 | 0.843 | 0.707 |
| 80 | 0.2041 | 0.2395 | 0.2181 | 0.132 | 0.273 | 0.847 | 0.871 | 0.732 |
| 85 | 0.2042 | 0.2381 | 0.2181 | 0.153 | 0.299 | 0.847 | 0.850 | 0.707 |
| 90 | 0.2043 | 0.2381 | 0.2181 | 0.150 | 0.293 | 0.847 | 0.853 | 0.712 |
| 95 | 0.2043 | 0.2395 | 0.2181 | 0.144 | 0.293 | 0.847 | 0.859 | 0.712 |
| 100 | 0.2043 | 0.2395 | 0.2181 | 0.144 | 0.293 | 0.847 | 0.859 | 0.712 |

*(b)* TF-IDF terms scoring

*Table C.6.* Tabular Results: 20 users, Freenode, Delta classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.8894 | 0.8722 | 0.9081 | 0.873 | 0.479 | 0.084 | 0.135 | 0.542 |
| 10 | 0.8756 | 0.9090 | 0.9464 | 0.202 | 0.019 | 0.117 | 0.811 | 0.984 |
| 15 | 0.8715 | 0.9256 | 0.9372 | 0.156 | 0.084 | 0.383 | 0.857 | 0.924 |
| 20 | 0.8806 | 0.9422 | 0.9292 | 0.107 | 0.173 | 0.618 | 0.902 | 0.838 |
| 25 | 0.9047 | 0.9361 | 0.9361 | 0.230 | 0.272 | 0.548 | 0.789 | 0.748 |
| 30 | 0.8881 | 0.9361 | 0.9389 | 0.132 | 0.116 | 0.529 | 0.879 | 0.894 |
| 35 | 0.8992 | 0.9506 | 0.9403 | 0.076 | 0.150 | 0.633 | 0.934 | 0.864 |
| 40 | 0.8815 | 0.9597 | 0.9381 | 0.017 | 0.090 | 0.856 | 0.986 | 0.919 |
| 45 | 0.9236 | 0.9672 | 0.9403 | 0.074 | 0.397 | 0.954 | 0.942 | 0.642 |
| 50 | 0.8978 | 0.9792 | 0.9464 | 0.021 | 0.155 | 0.955 | 0.982 | 0.863 |
| 55 | 0.8931 | 0.9792 | 0.9561 | 0.015 | 0.087 | 0.933 | 0.987 | 0.925 |
| 60 | 0.8778 | 0.9867 | 0.9631 | 0.007 | 0.015 | 0.963 | 0.994 | 0.988 |
| 65 | 0.8889 | 0.9783 | 0.9783 | 0.016 | 0.010 | 0.681 | 0.987 | 0.992 |
| 70 | 0.8861 | 0.9875 | 0.9783 | 0.009 | 0.011 | 0.963 | 0.993 | 0.991 |
| 75 | 0.8694 | 0.9792 | 0.9746 | 0.001 | 0.001 | 0.789 | 0.999 | 1.000 |
| 80 | 0.8556 | 0.9875 | 0.9653 | 0.000 | 0.000 | 0.969 | 1.000 | 1.000 |
| 85 | 0.8556 | 0.9917 | 0.9644 | 0.000 | 0.001 | 0.969 | 1.000 | 1.000 |
| 90 | 0.8472 | 0.9792 | 0.9640 | 0.000 | 0.000 | 0.865 | 1.000 | 1.000 |
| 95 | 0.8589 | 0.9792 | 0.9533 | 0.003 | 0.006 | 0.932 | 0.997 | 0.995 |
| 100 | 0.8694 | 0.9750 | 0.9556 | 0.000 | 0.001 | 0.972 | 1.000 | 0.999 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.9556 | 0.9833 | 0.9833 | 0.091 | 0.091 | 1.000 | 0.940 | 0.940 |
| 10 | 0.9556 | 0.9833 | 0.9833 | 0.120 | 0.091 | 0.681 | 0.912 | 0.940 |
| 15 | 0.9472 | 0.9917 | 0.9833 | 0.035 | 0.052 | 0.977 | 0.977 | 0.965 |
| 20 | 0.9389 | 0.9917 | 0.9708 | 0.021 | 0.100 | 0.963 | 0.986 | 0.921 |
| 25 | 0.9472 | 0.9611 | 0.9597 | 0.410 | 0.396 | 0.707 | 0.647 | 0.669 |
| 30 | 0.9472 | 0.9611 | 0.9569 | 0.410 | 0.396 | 0.769 | 0.647 | 0.669 |
| 35 | 0.9306 | 0.9500 | 0.9486 | 0.465 | 0.279 | 0.605 | 0.604 | 0.768 |
| 40 | 0.9278 | 0.9569 | 0.9492 | 0.333 | 0.259 | 0.819 | 0.717 | 0.785 |
| 45 | 0.9194 | 0.9556 | 0.9611 | 0.217 | 0.118 | 0.087 | 0.813 | 0.905 |
| 50 | 0.9278 | 0.9486 | 0.9597 | 0.403 | 0.164 | 0.135 | 0.643 | 0.865 |
| 55 | 0.9111 | 0.9472 | 0.9708 | 0.201 | 0.008 | 0.099 | 0.827 | 0.995 |
| 60 | 0.9111 | 0.9639 | 0.9722 | 0.021 | 0.007 | 0.500 | 0.986 | 0.996 |
| 65 | 0.9194 | 0.9722 | 0.9833 | 0.049 | 0.005 | 0.293 | 0.962 | 0.997 |
| 70 | 0.9194 | 0.9833 | 0.9833 | 0.005 | 0.005 | 0.681 | 0.997 | 0.997 |
| 75 | 0.9194 | 0.9917 | 0.9750 | 0.003 | 0.017 | 0.970 | 0.998 | 0.988 |
| 80 | 0.9194 | 0.9722 | 0.9833 | 0.069 | 0.012 | 0.293 | 0.944 | 0.991 |
| 85 | 0.8917 | 0.9722 | 0.9556 | 0.004 | 0.016 | 0.885 | 0.997 | 0.987 |
| 90 | 0.8968 | 0.9450 | 0.9375 | 0.050 | 0.091 | 0.986 | 0.960 | 0.925 |
| 95 | 0.8154 | 0.9394 | 0.9390 | 0.001 | 0.001 | 0.605 | 1.000 | 0.999 |
| 100 | 0.8694 | 0.9750 | 0.9556 | 0.000 | 0.001 | 0.972 | 1.000 | 0.999 |

*(b)* TF-IDF terms scoring

*Table C.7.* Tabular Results: 20 users, Krijn, KLD classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.5249 | 0.6173 | 0.5875 | 0.015 | 0.136 | 0.684 | 0.985 | 0.867 |
| 10 | 0.3999 | 0.5503 | 0.5942 | 0.002 | 0.000 | 0.165 | 0.998 | 1.000 |
| 15 | 0.3659 | 0.5080 | 0.5456 | 0.006 | 0.002 | 0.199 | 0.994 | 0.998 |
| 20 | 0.3753 | 0.4936 | 0.5068 | 0.017 | 0.016 | 0.475 | 0.984 | 0.985 |
| 25 | 0.3491 | 0.5029 | 0.4990 | 0.008 | 0.004 | 0.506 | 0.992 | 0.996 |
| 30 | 0.3816 | 0.4900 | 0.4979 | 0.055 | 0.039 | 0.280 | 0.946 | 0.962 |
| 35 | 0.4038 | 0.4897 | 0.5099 | 0.081 | 0.057 | 0.334 | 0.920 | 0.945 |
| 40 | 0.4268 | 0.4879 | 0.5031 | 0.182 | 0.138 | 0.479 | 0.821 | 0.865 |
| 45 | 0.4278 | 0.5019 | 0.4916 | 0.115 | 0.173 | 0.693 | 0.887 | 0.829 |
| 50 | 0.4214 | 0.4928 | 0.4905 | 0.090 | 0.192 | 0.600 | 0.911 | 0.811 |
| 55 | 0.4268 | 0.4759 | 0.4777 | 0.240 | 0.319 | 0.405 | 0.763 | 0.685 |
| 60 | 0.4370 | 0.4728 | 0.4827 | 0.365 | 0.309 | 0.422 | 0.638 | 0.694 |
| 65 | 0.4330 | 0.4878 | 0.4721 | 0.234 | 0.261 | 0.649 | 0.770 | 0.742 |
| 70 | 0.4281 | 0.4894 | 0.4860 | 0.225 | 0.232 | 0.463 | 0.779 | 0.771 |
| 75 | 0.4313 | 0.4682 | 0.4798 | 0.294 | 0.216 | 0.256 | 0.710 | 0.787 |
| 80 | 0.4312 | 0.4648 | 0.4751 | 0.275 | 0.152 | 0.144 | 0.728 | 0.851 |
| 85 | 0.4176 | 0.4618 | 0.4765 | 0.281 | 0.134 | 0.195 | 0.723 | 0.869 |
| 90 | 0.4187 | 0.4672 | 0.4585 | 0.238 | 0.182 | 0.634 | 0.765 | 0.821 |
| 95 | 0.4040 | 0.4687 | 0.4652 | 0.147 | 0.198 | 0.804 | 0.855 | 0.805 |
| 100 | 0.7368 | 0.7499 | 0.7518 | 0.357 | 0.294 | 0.415 | 0.653 | 0.717 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.7228 | 0.8422 | 0.8065 | 0.005 | 0.023 | 0.933 | 0.996 | 0.979 |
| 10 | 0.5942 | 0.6197 | 0.6295 | 0.228 | 0.110 | 0.382 | 0.777 | 0.893 |
| 15 | 0.5162 | 0.5557 | 0.5224 | 0.141 | 0.230 | 0.909 | 0.862 | 0.774 |
| 20 | 0.4581 | 0.4821 | 0.4618 | 0.431 | 0.260 | 0.388 | 0.575 | 0.744 |
| 25 | 0.4445 | 0.4565 | 0.4273 | 0.635 | 0.652 | 0.393 | 0.370 | 0.353 |
| 30 | 0.4159 | 0.4259 | 0.4084 | 0.846 | 0.798 | 0.272 | 0.157 | 0.205 |
| 35 | 0.4069 | 0.4129 | 0.3942 | 0.917 | 0.941 | 0.353 | 0.085 | 0.061 |
| 40 | 0.4101 | 0.3955 | 0.3850 | 0.970 | 0.935 | 0.074 | 0.031 | 0.067 |
| 45 | 0.3998 | 0.3832 | 0.3790 | 0.971 | 0.931 | 0.062 | 0.030 | 0.070 |
| 50 | 0.3992 | 0.3740 | 0.3690 | 0.987 | 0.989 | 0.308 | 0.013 | 0.011 |
| 55 | 0.3813 | 0.3618 | 0.3463 | 0.987 | 0.992 | 0.296 | 0.013 | 0.008 |
| 60 | 0.3789 | 0.3514 | 0.3331 | 0.995 | 0.994 | 0.177 | 0.006 | 0.006 |
| 65 | 0.3633 | 0.3373 | 0.3494 | 0.995 | 0.952 | 0.010 | 0.005 | 0.050 |
| 70 | 0.3581 | 0.3387 | 0.3480 | 0.986 | 0.942 | 0.033 | 0.014 | 0.060 |
| 75 | 0.3436 | 0.3270 | 0.3329 | 0.972 | 0.889 | 0.026 | 0.028 | 0.113 |
| 80 | 0.3284 | 0.3261 | 0.3309 | 0.884 | 0.824 | 0.061 | 0.118 | 0.178 |
| 85 | 0.3498 | 0.3139 | 0.3158 | 0.968 | 0.979 | 0.335 | 0.033 | 0.022 |
| 90 | 0.2945 | 0.3154 | 0.2987 | 0.740 | 0.852 | 0.581 | 0.264 | 0.150 |
| 95 | 0.3279 | 0.2984 | 0.2933 | 0.881 | 0.599 | 0.070 | 0.121 | 0.405 |
| 100 | 0.7368 | 0.7499 | 0.7518 | 0.357 | 0.294 | 0.415 | 0.653 | 0.717 |

*(b)* TF-IDF terms scoring

*Table C.8.* Tabular Results: 20 users, Krijn, $\chi^2$ classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.3406 | 0.3877 | 0.3562 | 0.012 | 0.042 | 0.998 | 0.988 | 0.959 |
| 10 | 0.2187 | 0.2878 | 0.2948 | 0.007 | 0.016 | 0.868 | 0.993 | 0.985 |
| 15 | 0.1927 | 0.2269 | 0.2448 | 0.049 | 0.034 | 0.410 | 0.952 | 0.967 |
| 20 | 0.1897 | 0.2202 | 0.2182 | 0.107 | 0.200 | 0.816 | 0.895 | 0.803 |
| 25 | 0.1905 | 0.2347 | 0.2190 | 0.044 | 0.138 | 0.909 | 0.957 | 0.865 |
| 30 | 0.1905 | 0.2090 | 0.2280 | 0.448 | 0.150 | 0.240 | 0.556 | 0.852 |
| 35 | 0.1911 | 0.2231 | 0.2366 | 0.302 | 0.168 | 0.771 | 0.701 | 0.835 |
| 40 | 0.1920 | 0.2156 | 0.2289 | 0.324 | 0.254 | 0.652 | 0.680 | 0.749 |
| 45 | 0.1944 | 0.2151 | 0.2300 | 0.302 | 0.226 | 0.652 | 0.702 | 0.777 |
| 50 | 0.1917 | 0.2154 | 0.2207 | 0.189 | 0.193 | 0.757 | 0.814 | 0.810 |
| 55 | 0.1918 | 0.2143 | 0.2149 | 0.177 | 0.168 | 0.761 | 0.826 | 0.835 |
| 60 | 0.1889 | 0.2128 | 0.2147 | 0.151 | 0.153 | 0.715 | 0.852 | 0.849 |
| 65 | 0.1889 | 0.2132 | 0.2127 | 0.133 | 0.163 | 0.767 | 0.869 | 0.840 |
| 70 | 0.1895 | 0.2036 | 0.2126 | 0.232 | 0.170 | 0.484 | 0.771 | 0.832 |
| 75 | 0.1881 | 0.2035 | 0.2122 | 0.153 | 0.170 | 0.618 | 0.850 | 0.833 |
| 80 | 0.1878 | 0.2040 | 0.2105 | 0.143 | 0.246 | 0.898 | 0.859 | 0.758 |
| 85 | 0.1877 | 0.2039 | 0.2105 | 0.130 | 0.228 | 0.884 | 0.872 | 0.775 |
| 90 | 0.1875 | 0.2039 | 0.2101 | 0.119 | 0.247 | 0.946 | 0.882 | 0.756 |
| 95 | 0.1874 | 0.2038 | 0.2099 | 0.119 | 0.255 | 0.954 | 0.882 | 0.749 |
| 100 | 0.1873 | 0.2030 | 0.2097 | 0.165 | 0.237 | 0.862 | 0.838 | 0.766 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.4837 | 0.6700 | 0.6231 | 0.000 | 0.000 | 0.995 | 1.000 | 1.000 |
| 10 | 0.3489 | 0.4880 | 0.4336 | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 |
| 15 | 0.2890 | 0.4256 | 0.3931 | 0.000 | 0.000 | 0.998 | 1.000 | 1.000 |
| 20 | 0.2573 | 0.3973 | 0.3739 | 0.000 | 0.000 | 0.964 | 1.000 | 1.000 |
| 25 | 0.2355 | 0.3534 | 0.3029 | 0.000 | 0.002 | 0.992 | 1.000 | 0.998 |
| 30 | 0.2129 | 0.2971 | 0.2836 | 0.001 | 0.000 | 0.976 | 0.999 | 1.000 |
| 35 | 0.1952 | 0.2472 | 0.2310 | 0.008 | 0.057 | 0.982 | 0.992 | 0.944 |
| 40 | 0.1946 | 0.2344 | 0.2250 | 0.112 | 0.140 | 0.796 | 0.890 | 0.863 |
| 45 | 0.1923 | 0.2185 | 0.2164 | 0.229 | 0.240 | 0.939 | 0.774 | 0.763 |
| 50 | 0.1924 | 0.2158 | 0.2130 | 0.265 | 0.222 | 0.915 | 0.738 | 0.781 |
| 55 | 0.1915 | 0.2144 | 0.2131 | 0.170 | 0.132 | 0.596 | 0.833 | 0.871 |
| 60 | 0.1887 | 0.2135 | 0.2122 | 0.166 | 0.181 | 0.700 | 0.836 | 0.821 |
| 65 | 0.1892 | 0.2042 | 0.2128 | 0.185 | 0.154 | 0.521 | 0.817 | 0.849 |
| 70 | 0.1886 | 0.2036 | 0.2122 | 0.210 | 0.194 | 0.644 | 0.793 | 0.809 |
| 75 | 0.1877 | 0.2034 | 0.2121 | 0.158 | 0.161 | 0.674 | 0.845 | 0.841 |
| 80 | 0.1876 | 0.2047 | 0.2103 | 0.087 | 0.237 | 0.914 | 0.914 | 0.766 |
| 85 | 0.1875 | 0.2045 | 0.2117 | 0.091 | 0.181 | 0.796 | 0.910 | 0.822 |
| 90 | 0.1874 | 0.2043 | 0.2114 | 0.099 | 0.191 | 0.839 | 0.903 | 0.812 |
| 95 | 0.1873 | 0.2032 | 0.2097 | 0.150 | 0.251 | 0.893 | 0.852 | 0.753 |
| 100 | 0.1873 | 0.2030 | 0.2097 | 0.165 | 0.237 | 0.862 | 0.838 | 0.766 |

*(b)* TF-IDF terms scoring

*Table C.9.* Tabular Results: 20 users, Krijn, Delta classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.5909 | 0.5455 | 0.5859 | 0.963 | 0.977 | 0.185 | 0.185 | 0.500 |
| 10 | 0.6714 | 0.5500 | 0.6524 | 0.990 | 0.709 | 0.027 | 0.017 | 0.392 |
| 15 | 0.7162 | 0.5946 | 0.6622 | 0.985 | 0.903 | 0.093 | 0.023 | 0.140 |
| 20 | 0.7658 | 0.6081 | 0.7523 | 0.996 | 0.661 | 0.012 | 0.006 | 0.445 |
| 25 | 0.7387 | 0.6622 | 0.7523 | 0.932 | 0.500 | 0.068 | 0.101 | 0.707 |
| 30 | 0.7207 | 0.6757 | 0.7523 | 0.801 | 0.248 | 0.099 | 0.234 | 0.830 |
| 35 | 0.6802 | 0.6450 | 0.7027 | 0.693 | 0.204 | 0.176 | 0.340 | 0.865 |
| 40 | 0.6554 | 0.6554 | 0.6824 | 0.429 | 0.195 | 0.315 | 0.617 | 0.849 |
| 45 | 0.6428 | 0.6090 | 0.6757 | 0.640 | 0.145 | 0.047 | 0.399 | 0.898 |
| 50 | 0.6315 | 0.6054 | 0.6450 | 0.765 | 0.363 | 0.086 | 0.268 | 0.701 |
| 55 | 0.6279 | 0.5633 | 0.6257 | 0.950 | 0.611 | 0.040 | 0.062 | 0.444 |
| 60 | 0.5977 | 0.5440 | 0.6115 | 0.899 | 0.264 | 0.028 | 0.120 | 0.800 |
| 65 | 0.5791 | 0.5174 | 0.6228 | 0.946 | 0.152 | 0.004 | 0.063 | 0.871 |
| 70 | 0.5987 | 0.5277 | 0.5942 | 0.941 | 0.625 | 0.041 | 0.070 | 0.458 |
| 75 | 0.5773 | 0.5066 | 0.6052 | 0.949 | 0.104 | 0.011 | 0.058 | 0.929 |
| 80 | 0.6034 | 0.5305 | 0.5947 | 0.924 | 0.584 | 0.104 | 0.086 | 0.472 |
| 85 | 0.5804 | 0.5158 | 0.5804 | 0.925 | 0.472 | 0.055 | 0.086 | 0.584 |
| 90 | 0.5931 | 0.5563 | 0.5753 | 0.880 | 0.777 | 0.306 | 0.142 | 0.277 |
| 95 | 0.6312 | 0.6125 | 0.6201 | 0.777 | 0.709 | 0.335 | 0.276 | 0.392 |
| 100 | 0.7572 | 0.7399 | 0.7575 | 0.658 | 0.500 | 0.356 | 0.446 | 0.977 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.6452 | 0.5484 | 0.6129 | 0.978 | 0.977 | 0.173 | 0.074 | 0.500 |
| 10 | 0.7020 | 0.6402 | 0.7030 | 0.972 | 0.500 | 0.087 | 0.087 | 0.977 |
| 15 | 0.7108 | 0.6507 | 0.7559 | 0.972 | 0.179 | 0.042 | 0.087 | 0.901 |
| 20 | 0.7184 | 0.6612 | 0.7333 | 0.970 | 0.392 | 0.085 | 0.173 | 0.709 |
| 25 | 0.7342 | 0.6795 | 0.7216 | 0.972 | 0.644 | 0.179 | 0.087 | 0.500 |
| 30 | 0.7432 | 0.6885 | 0.7036 | 0.933 | 0.909 | 0.292 | 0.110 | 0.211 |
| 35 | 0.7149 | 0.6318 | 0.7081 | 0.940 | 0.466 | 0.017 | 0.076 | 0.600 |
| 40 | 0.6998 | 0.6854 | 0.7239 | 0.500 | 0.049 | 0.181 | 0.568 | 0.979 |
| 45 | 0.6993 | 0.6899 | 0.6869 | 0.738 | 0.815 | 0.500 | 0.336 | 0.500 |
| 50 | 0.6813 | 0.6477 | 0.7050 | 0.963 | 0.179 | 0.023 | 0.053 | 0.901 |
| 55 | 0.6694 | 0.5950 | 0.6453 | 0.992 | 0.856 | 0.042 | 0.011 | 0.179 |
| 60 | 0.6413 | 0.6050 | 0.6385 | 0.872 | 0.571 | 0.086 | 0.154 | 0.476 |
| 65 | 0.6363 | 0.6034 | 0.6219 | 0.855 | 0.819 | 0.310 | 0.180 | 0.292 |
| 70 | 0.6055 | 0.5866 | 0.6055 | 0.751 | 0.681 | 0.342 | 0.342 | 0.681 |
| 75 | 0.6069 | 0.5811 | 0.6082 | 0.848 | 0.500 | 0.116 | 0.178 | 0.707 |
| 80 | 0.6019 | 0.5655 | 0.6345 | 0.938 | 0.068 | 0.016 | 0.075 | 0.956 |
| 85 | 0.6048 | 0.5809 | 0.6223 | 0.979 | 0.181 | 0.029 | 0.049 | 0.899 |
| 90 | 0.6380 | 0.6183 | 0.6414 | 0.882 | 0.376 | 0.242 | 0.155 | 0.700 |
| 95 | 0.7036 | 0.6541 | 0.6974 | 0.971 | 0.554 | 0.045 | 0.046 | 0.554 |
| 100 | 0.7572 | 0.7399 | 0.7575 | 0.658 | 0.500 | 0.356 | 0.446 | 0.977 |

*(b)* TF-IDF terms scoring

*Table C.10.* Tabular Results: 20 users, Twitter, KLD classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.3275 | 0.3126 | 0.3459 | 0.333 | 0.068 | 0.337 | 0.696 | 0.944 |
| 10 | 0.3949 | 0.3105 | 0.4053 | 0.685 | 0.092 | 0.084 | 0.335 | 0.919 |
| 15 | 0.4360 | 0.3606 | 0.4167 | 0.758 | 0.331 | 0.191 | 0.258 | 0.684 |
| 20 | 0.4674 | 0.3572 | 0.4894 | 0.719 | 0.067 | 0.089 | 0.294 | 0.940 |
| 25 | 0.4392 | 0.4062 | 0.4720 | 0.466 | 0.059 | 0.216 | 0.556 | 0.947 |
| 30 | 0.4334 | 0.4115 | 0.4713 | 0.301 | 0.020 | 0.128 | 0.715 | 0.982 |
| 35 | 0.3937 | 0.3641 | 0.4067 | 0.689 | 0.142 | 0.152 | 0.330 | 0.871 |
| 40 | 0.3543 | 0.3680 | 0.3682 | 0.278 | 0.041 | 0.377 | 0.736 | 0.967 |
| 45 | 0.3355 | 0.3351 | 0.3520 | 0.174 | 0.205 | 0.490 | 0.840 | 0.817 |
| 50 | 0.3163 | 0.3323 | 0.3089 | 0.210 | 0.575 | 0.890 | 0.802 | 0.444 |
| 55 | 0.2744 | 0.2756 | 0.2902 | 0.416 | 0.115 | 0.339 | 0.596 | 0.893 |
| 60 | 0.2511 | 0.2504 | 0.2690 | 0.550 | 0.150 | 0.144 | 0.475 | 0.864 |
| 65 | 0.2467 | 0.2327 | 0.2480 | 0.878 | 0.453 | 0.086 | 0.133 | 0.578 |
| 70 | 0.2381 | 0.2169 | 0.2233 | 0.736 | 0.444 | 0.238 | 0.288 | 0.583 |
| 75 | 0.2201 | 0.2074 | 0.2141 | 0.603 | 0.147 | 0.174 | 0.414 | 0.866 |
| 80 | 0.2067 | 0.1822 | 0.2051 | 0.889 | 0.314 | 0.018 | 0.122 | 0.706 |
| 85 | 0.1925 | 0.1711 | 0.1956 | 0.828 | 0.231 | 0.033 | 0.184 | 0.790 |
| 90 | 0.1904 | 0.1712 | 0.1892 | 0.790 | 0.608 | 0.228 | 0.223 | 0.422 |
| 95 | 0.1900 | 0.1808 | 0.1851 | 0.586 | 0.637 | 0.378 | 0.431 | 0.390 |
| 100 | 0.4761 | 0.4125 | 0.3908 | 0.985 | 0.999 | 0.682 | 0.017 | 0.001 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.4161 | 0.3204 | 0.4177 | 0.896 | 0.360 | 0.046 | 0.117 | 0.683 |
| 10 | 0.6389 | 0.4960 | 0.5717 | 0.987 | 0.925 | 0.130 | 0.016 | 0.102 |
| 15 | 0.6912 | 0.5819 | 0.7304 | 0.974 | 0.273 | 0.020 | 0.038 | 0.781 |
| 20 | 0.7095 | 0.5519 | 0.7071 | 0.995 | 0.557 | 0.009 | 0.007 | 0.500 |
| 25 | 0.7432 | 0.5942 | 0.6644 | 0.993 | 0.914 | 0.053 | 0.010 | 0.116 |
| 30 | 0.7297 | 0.6071 | 0.6351 | 0.980 | 0.986 | 0.202 | 0.026 | 0.020 |
| 35 | 0.7194 | 0.5467 | 0.6757 | 0.993 | 0.798 | 0.018 | 0.008 | 0.238 |
| 40 | 0.6856 | 0.6142 | 0.6712 | 0.882 | 0.738 | 0.171 | 0.143 | 0.335 |
| 45 | 0.6698 | 0.5949 | 0.6077 | 0.949 | 0.987 | 0.378 | 0.063 | 0.018 |
| 50 | 0.6392 | 0.5404 | 0.5851 | 0.994 | 0.968 | 0.205 | 0.008 | 0.042 |
| 55 | 0.5842 | 0.4543 | 0.4874 | 0.999 | 0.999 | 0.238 | 0.001 | 0.002 |
| 60 | 0.5279 | 0.4260 | 0.4446 | 0.995 | 0.997 | 0.163 | 0.006 | 0.004 |
| 65 | 0.5248 | 0.3755 | 0.4005 | 0.999 | 1.000 | 0.164 | 0.001 | 0.001 |
| 70 | 0.4604 | 0.3347 | 0.3750 | 0.995 | 0.998 | 0.177 | 0.006 | 0.003 |
| 75 | 0.4608 | 0.3105 | 0.3415 | 0.999 | 0.999 | 0.051 | 0.001 | 0.002 |
| 80 | 0.4262 | 0.3217 | 0.3303 | 0.989 | 0.993 | 0.124 | 0.012 | 0.009 |
| 85 | 0.4127 | 0.3169 | 0.3397 | 0.990 | 0.972 | 0.059 | 0.012 | 0.037 |
| 90 | 0.4217 | 0.3149 | 0.3012 | 0.982 | 0.992 | 0.477 | 0.020 | 0.009 |
| 95 | 0.4209 | 0.2832 | 0.3236 | 0.998 | 0.999 | 0.070 | 0.002 | 0.001 |
| 100 | 0.4761 | 0.4125 | 0.3908 | 0.985 | 0.999 | 0.682 | 0.017 | 0.001 |

*(b)* TF-IDF terms scoring

*Table C.11.* Tabular Results: 20 users, Twitter, $\chi^2$ classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.3124 | 0.2592 | 0.3332 | 0.594 | 0.561 | 0.363 | 0.453 | 0.480 |
| 10 | 0.3706 | 0.2801 | 0.3726 | 0.594 | 0.175 | 0.500 | 0.425 | 0.847 |
| 15 | 0.4102 | 0.3242 | 0.3982 | 0.628 | 0.312 | 0.500 | 0.388 | 0.712 |
| 20 | 0.4450 | 0.3304 | 0.4867 | 0.729 | 0.029 | 0.155 | 0.285 | 0.979 |
| 25 | 0.4297 | 0.3811 | 0.4715 | 0.349 | 0.022 | 0.296 | 0.670 | 0.984 |
| 30 | 0.4059 | 0.3905 | 0.4712 | 0.245 | 0.003 | 0.129 | 0.769 | 0.998 |
| 35 | 0.3755 | 0.3440 | 0.4072 | 0.594 | 0.131 | 0.139 | 0.424 | 0.883 |
| 40 | 0.3375 | 0.3406 | 0.3583 | 0.253 | 0.122 | 0.453 | 0.759 | 0.889 |
| 45 | 0.3043 | 0.3028 | 0.3420 | 0.118 | 0.053 | 0.306 | 0.891 | 0.951 |
| 50 | 0.2832 | 0.2785 | 0.2711 | 0.233 | 0.548 | 0.896 | 0.778 | 0.468 |
| 55 | 0.2545 | 0.2411 | 0.2422 | 0.460 | 0.644 | 0.762 | 0.556 | 0.372 |
| 60 | 0.2271 | 0.2254 | 0.2324 | 0.681 | 0.833 | 0.853 | 0.341 | 0.181 |
| 65 | 0.2179 | 0.2096 | 0.2080 | 0.972 | 0.968 | 0.472 | 0.031 | 0.036 |
| 70 | 0.2104 | 0.2082 | 0.1898 | 0.810 | 0.941 | 0.925 | 0.200 | 0.065 |
| 75 | 0.1878 | 0.1950 | 0.1816 | 0.838 | 0.828 | 0.524 | 0.169 | 0.184 |
| 80 | 0.1795 | 0.1891 | 0.1734 | 0.857 | 0.831 | 0.605 | 0.152 | 0.183 |
| 85 | 0.1642 | 0.1652 | 0.1622 | 0.675 | 0.681 | 0.791 | 0.345 | 0.341 |
| 90 | 0.1642 | 0.1646 | 0.1587 | 0.810 | 0.961 | 0.911 | 0.207 | 0.045 |
| 95 | 0.1464 | 0.1483 | 0.1428 | 0.778 | 0.930 | 0.821 | 0.234 | 0.077 |
| 100 | 0.1591 | 0.1706 | 0.1570 | 0.109 | 0.826 | 0.974 | 0.900 | 0.189 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.3281 | 0.2297 | 0.3069 | 0.828 | 0.305 | 0.046 | 0.191 | 0.730 |
| 10 | 0.4079 | 0.3492 | 0.4106 | 0.709 | 0.118 | 0.074 | 0.319 | 0.904 |
| 15 | 0.4066 | 0.3507 | 0.4696 | 0.472 | 0.032 | 0.012 | 0.556 | 0.973 |
| 20 | 0.4080 | 0.3650 | 0.4523 | 0.656 | 0.133 | 0.012 | 0.378 | 0.885 |
| 25 | 0.4315 | 0.3967 | 0.4526 | 0.378 | 0.025 | 0.011 | 0.656 | 0.980 |
| 30 | 0.4438 | 0.4046 | 0.4366 | 0.538 | 0.181 | 0.029 | 0.481 | 0.835 |
| 35 | 0.4272 | 0.3540 | 0.4177 | 0.692 | 0.211 | 0.015 | 0.324 | 0.804 |
| 40 | 0.3726 | 0.3771 | 0.4120 | 0.046 | 0.015 | 0.254 | 0.958 | 0.987 |
| 45 | 0.3374 | 0.3489 | 0.3526 | 0.166 | 0.041 | 0.417 | 0.846 | 0.966 |
| 50 | 0.3039 | 0.2839 | 0.3332 | 0.543 | 0.211 | 0.090 | 0.474 | 0.803 |
| 55 | 0.2719 | 0.2544 | 0.2618 | 0.790 | 0.904 | 0.606 | 0.224 | 0.105 |
| 60 | 0.2442 | 0.2393 | 0.2378 | 0.644 | 0.704 | 0.583 | 0.377 | 0.319 |
| 65 | 0.2081 | 0.2273 | 0.1965 | 0.516 | 0.889 | 0.909 | 0.500 | 0.122 |
| 70 | 0.1907 | 0.2091 | 0.1820 | 0.515 | 0.871 | 0.905 | 0.500 | 0.143 |
| 75 | 0.1850 | 0.1983 | 0.1780 | 0.794 | 0.926 | 0.793 | 0.216 | 0.082 |
| 80 | 0.1802 | 0.1887 | 0.1740 | 0.918 | 0.947 | 0.583 | 0.088 | 0.059 |
| 85 | 0.1683 | 0.1675 | 0.1621 | 0.769 | 0.947 | 0.963 | 0.246 | 0.059 |
| 90 | 0.1639 | 0.1540 | 0.1486 | 0.676 | 0.888 | 0.963 | 0.339 | 0.122 |
| 95 | 0.1465 | 0.1489 | 0.1423 | 0.723 | 0.947 | 0.899 | 0.293 | 0.059 |
| 100 | 0.1591 | 0.1706 | 0.1570 | 0.109 | 0.826 | 0.974 | 0.900 | 0.189 |

*(b)* TF-IDF terms scoring

*Table C.12.* Tabular Results: 20 users, Twitter, Delta classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.4138 |
| 10 | 0.4242 |
| 15 | 0.4227 |
| 20 | 0.4160 |
| 25 | 0.5035 |
| 30 | 0.5377 |
| 35 | 0.5407 |
| 40 | 0.5463 |
| 45 | 0.5502 |
| 50 | 0.5533 |
| 55 | 0.5578 |
| 60 | 0.5607 |
| 65 | 0.5643 |
| 70 | 0.5705 |
| 75 | 0.5763 |
| 80 | 0.5826 |
| 85 | 0.5850 |
| 90 | 0.5855 |
| 95 | 0.5826 |
| 100 | 0.6735 |

*(a)* KLD terms scoring, KLD classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.2911 |
| 10 | 0.2705 |
| 15 | 0.2346 |
| 20 | 0.2209 |
| 25 | 0.2576 |
| 30 | 0.2660 |
| 35 | 0.2589 |
| 40 | 0.2544 |
| 45 | 0.2506 |
| 50 | 0.2459 |
| 55 | 0.2418 |
| 60 | 0.2363 |
| 65 | 0.2370 |
| 70 | 0.2347 |
| 75 | 0.2312 |
| 80 | 0.2219 |
| 85 | 0.2137 |
| 90 | 0.1891 |
| 95 | 0.1446 |
| 100 | 0.4879 |

*(b)* KLD terms scoring, $\chi^2$ classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.0613 |
| 10 | 0.0406 |
| 15 | 0.0333 |
| 20 | 0.0281 |
| 25 | 0.0280 |
| 30 | 0.0281 |
| 35 | 0.0260 |
| 40 | 0.0229 |
| 45 | 0.0216 |
| 50 | 0.0198 |
| 55 | 0.0176 |
| 60 | 0.0162 |
| 65 | 0.0154 |
| 70 | 0.0144 |
| 75 | 0.0138 |
| 80 | 0.0136 |
| 85 | 0.0139 |
| 90 | 0.0139 |
| 95 | 0.0138 |
| 100 | 0.0124 |

*(c)* KLD terms scoring, Delta classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.4490 |
| 10 | 0.4382 |
| 15 | 0.4558 |
| 20 | 0.4714 |
| 25 | 0.4790 |
| 30 | 0.5436 |
| 35 | 0.5515 |
| 40 | 0.5575 |
| 45 | 0.5662 |
| 50 | 0.5773 |
| 55 | 0.5851 |
| 60 | 0.5937 |
| 65 | 0.6023 |
| 70 | 0.6114 |
| 75 | 0.6217 |
| 80 | 0.6350 |
| 85 | 0.6468 |
| 90 | 0.6519 |
| 95 | 0.6599 |
| 100 | 0.6735 |

*(d)* TF-IDF terms scoring, KLD classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.4298 |
| 10 | 0.4645 |
| 15 | 0.3880 |
| 20 | 0.4034 |
| 25 | 0.4408 |
| 30 | 0.5038 |
| 35 | 0.5064 |
| 40 | 0.5054 |
| 45 | 0.5149 |
| 50 | 0.5233 |
| 55 | 0.5302 |
| 60 | 0.5409 |
| 65 | 0.5456 |
| 70 | 0.5539 |
| 75 | 0.5656 |
| 80 | 0.5704 |
| 85 | 0.5714 |
| 90 | 0.5697 |
| 95 | 0.5379 |
| 100 | 0.4879 |

*(e)* TF-IDF terms scoring, $\chi^2$ classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.0859 |
| 10 | 0.0558 |
| 15 | 0.0398 |
| 20 | 0.0320 |
| 25 | 0.0293 |
| 30 | 0.0275 |
| 35 | 0.0245 |
| 40 | 0.0230 |
| 45 | 0.0206 |
| 50 | 0.0192 |
| 55 | 0.0171 |
| 60 | 0.0161 |
| 65 | 0.0148 |
| 70 | 0.0143 |
| 75 | 0.0138 |
| 80 | 0.0136 |
| 85 | 0.0139 |
| 90 | 0.0140 |
| 95 | 0.0132 |
| 100 | 0.0124 |

*(f)* TF-IDF terms scoring, Delta classifier

*Table C.13.* Tabular Results: Hundreds users, Associated Press

| Voc. % | method x0 |
|---|---|
| 5 | 0.4744 |
| 10 | 0.4792 |
| 15 | 0.4701 |
| 20 | 0.4556 |
| 25 | 0.4448 |
| 30 | 0.4325 |
| 35 | 0.4263 |
| 40 | 0.4114 |
| 45 | 0.4092 |
| 50 | 0.4051 |
| 55 | 0.3997 |
| 60 | 0.4038 |
| 65 | 0.4072 |
| 70 | 0.4081 |
| 75 | 0.4158 |
| 80 | 0.4241 |
| 85 | 0.4242 |
| 90 | 0.4302 |
| 95 | 0.4319 |
| 100 | 0.5790 |

*(a)* KLD terms scoring, KLD classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.3356 |
| 10 | 0.3137 |
| 15 | 0.3005 |
| 20 | 0.2892 |
| 25 | 0.2810 |
| 30 | 0.2742 |
| 35 | 0.2613 |
| 40 | 0.2479 |
| 45 | 0.2379 |
| 50 | 0.2335 |
| 55 | 0.2240 |
| 60 | 0.2218 |
| 65 | 0.2214 |
| 70 | 0.2126 |
| 75 | 0.1997 |
| 80 | 0.1859 |
| 85 | 0.1605 |
| 90 | 0.1185 |
| 95 | 0.0848 |
| 100 | 0.3006 |

*(b)* KLD terms scoring, $\chi^2$ classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.1187 |
| 10 | 0.0904 |
| 15 | 0.0800 |
| 20 | 0.0668 |
| 25 | 0.0529 |
| 30 | 0.0452 |
| 35 | 0.0367 |
| 40 | 0.0298 |
| 45 | 0.0249 |
| 50 | 0.0202 |
| 55 | 0.0176 |
| 60 | 0.0179 |
| 65 | 0.0152 |
| 70 | 0.0128 |
| 75 | 0.0096 |
| 80 | 0.0079 |
| 85 | 0.0076 |
| 90 | 0.0075 |
| 95 | 0.0075 |
| 100 | 0.0075 |

*(c)* KLD terms scoring, Delta classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.3483 |
| 10 | 0.3363 |
| 15 | 0.3122 |
| 20 | 0.3047 |
| 25 | 0.2931 |
| 30 | 0.2828 |
| 35 | 0.2800 |
| 40 | 0.2777 |
| 45 | 0.2757 |
| 50 | 0.2789 |
| 55 | 0.2825 |
| 60 | 0.2890 |
| 65 | 0.2949 |
| 70 | 0.3038 |
| 75 | 0.3103 |
| 80 | 0.3190 |
| 85 | 0.3317 |
| 90 | 0.3419 |
| 95 | 0.3435 |
| 100 | 0.3925 |

*(d)* TF-IDF terms scoring, KLD classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.3751 |
| 10 | 0.3603 |
| 15 | 0.3425 |
| 20 | 0.3236 |
| 25 | 0.3050 |
| 30 | 0.2965 |
| 35 | 0.2810 |
| 40 | 0.2749 |
| 45 | 0.2646 |
| 50 | 0.2641 |
| 55 | 0.2596 |
| 60 | 0.2592 |
| 65 | 0.2475 |
| 70 | 0.2456 |
| 75 | 0.2407 |
| 80 | 0.2320 |
| 85 | 0.2202 |
| 90 | 0.2036 |
| 95 | 0.1893 |
| 100 | 0.2038 |

*(e)* TF-IDF terms scoring, $\chi^2$ classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.0955 |
| 10 | 0.0782 |
| 15 | 0.0643 |
| 20 | 0.0518 |
| 25 | 0.0401 |
| 30 | 0.0316 |
| 35 | 0.0258 |
| 40 | 0.0204 |
| 45 | 0.0163 |
| 50 | 0.0136 |
| 55 | 0.0128 |
| 60 | 0.0105 |
| 65 | 0.0094 |
| 70 | 0.0081 |
| 75 | 0.0062 |
| 80 | 0.0054 |
| 85 | 0.0052 |
| 90 | 0.0051 |
| 95 | 0.0051 |
| 100 | 0.0051 |

*(f)* TF-IDF terms scoring, Delta classifier

*Table C.14.* Tabular Results: Hundreds users, La Stampa

| Voc. % | method x0 |
|---|---|
| 5 | 0.4912 |
| 10 | 0.5066 |
| 15 | 0.5101 |
| 20 | 0.5091 |
| 25 | 0.5053 |
| 30 | 0.5017 |
| 35 | 0.5052 |
| 40 | 0.5038 |
| 45 | 0.4983 |
| 50 | 0.4985 |
| 55 | 0.4975 |
| 60 | 0.5048 |
| 65 | 0.5065 |
| 70 | 0.5006 |
| 75 | 0.5088 |
| 80 | 0.5156 |
| 85 | 0.5286 |
| 90 | 0.5350 |
| 95 | 0.5314 |
| 100 | 0.5976 |

*(a)* KLD terms scoring, KLD classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.2943 |
| 10 | 0.2701 |
| 15 | 0.2634 |
| 20 | 0.2530 |
| 25 | 0.2473 |
| 30 | 0.2293 |
| 35 | 0.2314 |
| 40 | 0.2236 |
| 45 | 0.2151 |
| 50 | 0.2058 |
| 55 | 0.2026 |
| 60 | 0.1986 |
| 65 | 0.1911 |
| 70 | 0.1881 |
| 75 | 0.1903 |
| 80 | 0.1867 |
| 85 | 0.1732 |
| 90 | 0.1577 |
| 95 | 0.1246 |
| 100 | 0.2845 |

*(b)* KLD terms scoring, $\chi^2$ classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.0914 |
| 10 | 0.0625 |
| 15 | 0.0483 |
| 20 | 0.0365 |
| 25 | 0.0288 |
| 30 | 0.0224 |
| 35 | 0.0199 |
| 40 | 0.0154 |
| 45 | 0.0145 |
| 50 | 0.0100 |
| 55 | 0.0100 |
| 60 | 0.0105 |
| 65 | 0.0092 |
| 70 | 0.0084 |
| 75 | 0.0079 |
| 80 | 0.0078 |
| 85 | 0.0078 |
| 90 | 0.0078 |
| 95 | 0.0078 |
| 100 | 0.0078 |

*(c)* TF-IDF terms scoring, Delta classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.5045 |
| 10 | 0.5096 |
| 15 | 0.5082 |
| 20 | 0.5033 |
| 25 | 0.4972 |
| 30 | 0.4880 |
| 35 | 0.4878 |
| 40 | 0.4957 |
| 45 | 0.4972 |
| 50 | 0.5035 |
| 55 | 0.5024 |
| 60 | 0.5062 |
| 65 | 0.5130 |
| 70 | 0.5198 |
| 75 | 0.5308 |
| 80 | 0.5407 |
| 85 | 0.5450 |
| 90 | 0.5531 |
| 95 | 0.5683 |
| 100 | 0.5976 |

*(d)* TF-IDF terms scoring, KLD classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.4179 |
| 10 | 0.4035 |
| 15 | 0.3885 |
| 20 | 0.3725 |
| 25 | 0.3423 |
| 30 | 0.3257 |
| 35 | 0.3126 |
| 40 | 0.3080 |
| 45 | 0.3168 |
| 50 | 0.3209 |
| 55 | 0.3083 |
| 60 | 0.3224 |
| 65 | 0.3139 |
| 70 | 0.3039 |
| 75 | 0.2950 |
| 80 | 0.3020 |
| 85 | 0.2897 |
| 90 | 0.2780 |
| 95 | 0.2367 |
| 100 | 0.2845 |

*(e)* TF-IDF terms scoring, $\chi^2$ classifier

| Voc. % | method x0 |
|---|---|
| 5 | 0.1094 |
| 10 | 0.0839 |
| 15 | 0.0666 |
| 20 | 0.0454 |
| 25 | 0.0316 |
| 30 | 0.0244 |
| 35 | 0.0196 |
| 40 | 0.0160 |
| 45 | 0.0125 |
| 50 | 0.0110 |
| 55 | 0.0110 |
| 60 | 0.0102 |
| 65 | 0.0088 |
| 70 | 0.0085 |
| 75 | 0.0079 |
| 80 | 0.0078 |
| 85 | 0.0078 |
| 90 | 0.0078 |
| 95 | 0.0078 |
| 100 | 0.0078 |

*(f)* TF-IDF terms scoring, Delta classifier

*Table C.15.* Tabular Results: Hundreds users, Glasgow Herald

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.1055 | 0.1234 | 0.1321 | 0.006 | 0.001 | 0.542 | 0.994 | 0.999 |
| 10 | 0.1219 | 0.1609 | 0.1745 | 0.000 | 0.000 | 0.433 | 1.000 | 1.000 |
| 15 | 0.1513 | 0.1786 | 0.1880 | 0.006 | 0.001 | 0.430 | 0.994 | 0.999 |
| 20 | 0.1734 | 0.1983 | 0.2111 | 0.027 | 0.001 | 0.658 | 0.973 | 0.999 |
| 25 | 0.1892 | 0.2161 | 0.2176 | 0.010 | 0.013 | 0.902 | 0.990 | 0.987 |
| 30 | 0.2057 | 0.2271 | 0.2144 | 0.087 | 0.078 | 0.995 | 0.913 | 0.922 |
| 35 | 0.2107 | 0.2210 | 0.2023 | 0.376 | 0.896 | 1.000 | 0.624 | 0.104 |
| 40 | 0.2165 | 0.2170 | 0.2070 | 0.877 | 0.974 | 1.000 | 0.123 | 0.026 |
| 45 | 0.2137 | 0.2231 | 0.2156 | 0.789 | 0.964 | 1.000 | 0.211 | 0.036 |
| 50 | 0.2254 | 0.2236 | 0.2146 | 0.987 | 0.992 | 0.999 | 0.013 | 0.008 |
| 55 | 0.2263 | 0.2174 | 0.2090 | 1.000 | 0.998 | 0.994 | 0.000 | 0.002 |
| 60 | 0.2306 | 0.2172 | 0.2103 | 0.992 | 0.992 | 0.983 | 0.008 | 0.008 |
| 65 | 0.2289 | 0.2139 | 0.2097 | 0.999 | 0.996 | 0.982 | 0.001 | 0.004 |
| 70 | 0.2269 | 0.2167 | 0.2088 | 0.991 | 0.986 | 0.987 | 0.009 | 0.014 |
| 75 | 0.2249 | 0.2213 | 0.2133 | 0.954 | 0.879 | 0.993 | 0.046 | 0.121 |
| 80 | 0.2305 | 0.2157 | 0.2061 | 0.883 | 0.905 | 0.990 | 0.117 | 0.095 |
| 85 | 0.2322 | 0.2163 | 0.2076 | 0.994 | 0.902 | 0.909 | 0.007 | 0.098 |
| 90 | 0.2271 | 0.2131 | 0.2079 | 0.845 | 0.806 | 0.767 | 0.155 | 0.194 |
| 95 | 0.2190 | 0.2091 | 0.1988 | 0.964 | 0.979 | 0.869 | 0.036 | 0.021 |
| 100 | 0.2391 | 0.2054 | 0.1893 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0814 | 0.1023 | 0.1282 | 0.015 | 0.000 | 0.060 | 0.985 | 1.000 |
| 10 | 0.1059 | 0.1287 | 0.1539 | 0.010 | 0.000 | 0.028 | 0.990 | 1.000 |
| 15 | 0.1225 | 0.1540 | 0.1795 | 0.022 | 0.000 | 0.030 | 0.978 | 1.000 |
| 20 | 0.1495 | 0.1734 | 0.1762 | 0.057 | 0.031 | 0.901 | 0.943 | 0.969 |
| 25 | 0.1703 | 0.1721 | 0.1894 | 0.589 | 0.098 | 0.365 | 0.412 | 0.902 |
| 30 | 0.1767 | 0.1806 | 0.1858 | 0.973 | 0.660 | 0.855 | 0.027 | 0.341 |
| 35 | 0.1793 | 0.1754 | 0.1696 | 0.992 | 0.972 | 1.000 | 0.008 | 0.028 |
| 40 | 0.1834 | 0.1732 | 0.1676 | 0.995 | 0.999 | 1.000 | 0.005 | 0.001 |
| 45 | 0.1892 | 0.1763 | 0.1653 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 50 | 0.1948 | 0.1759 | 0.1724 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 55 | 0.1949 | 0.1743 | 0.1718 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 60 | 0.1925 | 0.1714 | 0.1639 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 65 | 0.1959 | 0.1607 | 0.1630 | 1.000 | 1.000 | 0.997 | 0.000 | 0.000 |
| 70 | 0.1992 | 0.1638 | 0.1580 | 1.000 | 1.000 | 0.999 | 0.000 | 0.000 |
| 75 | 0.2107 | 0.1714 | 0.1573 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 80 | 0.2122 | 0.1723 | 0.1603 | 1.000 | 1.000 | 0.999 | 0.000 | 0.000 |
| 85 | 0.2216 | 0.1758 | 0.1594 | 1.000 | 1.000 | 0.997 | 0.000 | 0.000 |
| 90 | 0.2273 | 0.1888 | 0.1686 | 1.000 | 1.000 | 0.997 | 0.000 | 0.000 |
| 95 | 0.2254 | 0.1953 | 0.1791 | 1.000 | 1.000 | 0.999 | 0.000 | 0.000 |
| 100 | 0.2391 | 0.2054 | 0.1893 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |

*(b)* TF-IDF terms scoring

*Table C.16.* Tabular Results: Hundreds users, Freenode, KLD classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0645 | 0.0884 | 0.0883 | 0.004 | 0.009 | 0.919 | 0.996 | 0.992 |
| 10 | 0.0723 | 0.1036 | 0.1058 | 0.004 | 0.017 | 0.999 | 0.996 | 0.983 |
| 15 | 0.0898 | 0.1162 | 0.1193 | 0.095 | 0.085 | 0.999 | 0.905 | 0.915 |
| 20 | 0.1035 | 0.1255 | 0.1345 | 0.077 | 0.119 | 1.000 | 0.923 | 0.881 |
| 25 | 0.1152 | 0.1302 | 0.1328 | 0.592 | 0.389 | 1.000 | 0.409 | 0.612 |
| 30 | 0.1233 | 0.1392 | 0.1323 | 0.790 | 0.587 | 1.000 | 0.210 | 0.413 |
| 35 | 0.1227 | 0.1285 | 0.1250 | 0.888 | 0.903 | 1.000 | 0.112 | 0.097 |
| 40 | 0.1198 | 0.1219 | 0.1247 | 0.989 | 0.986 | 1.000 | 0.011 | 0.014 |
| 45 | 0.1247 | 0.1255 | 0.1155 | 0.998 | 1.000 | 1.000 | 0.002 | 0.000 |
| 50 | 0.1287 | 0.1260 | 0.1178 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 55 | 0.1274 | 0.1163 | 0.1098 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 60 | 0.1293 | 0.1063 | 0.1057 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 65 | 0.1201 | 0.1101 | 0.1068 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 70 | 0.1099 | 0.1039 | 0.1055 | 1.000 | 0.999 | 1.000 | 0.000 | 0.001 |
| 75 | 0.0997 | 0.0954 | 0.0986 | 1.000 | 0.993 | 1.000 | 0.000 | 0.007 |
| 80 | 0.0953 | 0.0920 | 0.0936 | 0.997 | 0.994 | 1.000 | 0.003 | 0.007 |
| 85 | 0.0962 | 0.0856 | 0.0901 | 1.000 | 1.000 | 0.998 | 0.000 | 0.001 |
| 90 | 0.0928 | 0.0812 | 0.0818 | 0.993 | 0.979 | 0.978 | 0.007 | 0.021 |
| 95 | 0.0901 | 0.0780 | 0.0744 | 0.995 | 0.982 | 0.925 | 0.005 | 0.018 |
| 100 | 0.1980 | 0.2051 | 0.2070 | 0.581 | 0.361 | 0.668 | 0.419 | 0.639 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0819 | 0.1109 | 0.1320 | 0.000 | 0.000 | 0.046 | 1.000 | 1.000 |
| 10 | 0.1057 | 0.1394 | 0.1683 | 0.000 | 0.000 | 0.003 | 1.000 | 1.000 |
| 15 | 0.1295 | 0.1664 | 0.1945 | 0.000 | 0.000 | 0.002 | 1.000 | 1.000 |
| 20 | 0.1543 | 0.1948 | 0.2006 | 0.000 | 0.000 | 0.609 | 1.000 | 1.000 |
| 25 | 0.1746 | 0.1954 | 0.2161 | 0.052 | 0.000 | 0.030 | 0.948 | 1.000 |
| 30 | 0.1786 | 0.2053 | 0.2336 | 0.029 | 0.000 | 0.058 | 0.971 | 1.000 |
| 35 | 0.1833 | 0.2091 | 0.2171 | 0.090 | 0.006 | 0.682 | 0.910 | 0.994 |
| 40 | 0.1851 | 0.2134 | 0.2168 | 0.077 | 0.016 | 0.731 | 0.924 | 0.984 |
| 45 | 0.1907 | 0.2085 | 0.2099 | 0.345 | 0.176 | 0.904 | 0.655 | 0.824 |
| 50 | 0.1935 | 0.2104 | 0.2052 | 0.221 | 0.382 | 0.955 | 0.779 | 0.618 |
| 55 | 0.1970 | 0.2085 | 0.2041 | 0.764 | 0.414 | 0.721 | 0.236 | 0.586 |
| 60 | 0.1960 | 0.2039 | 0.2011 | 0.892 | 0.683 | 0.882 | 0.108 | 0.318 |
| 65 | 0.1950 | 0.1983 | 0.2033 | 0.981 | 0.818 | 0.607 | 0.019 | 0.182 |
| 70 | 0.1925 | 0.1929 | 0.2046 | 0.999 | 0.621 | 0.479 | 0.001 | 0.379 |
| 75 | 0.1973 | 0.2019 | 0.1942 | 0.973 | 0.644 | 0.610 | 0.027 | 0.356 |
| 80 | 0.2006 | 0.1951 | 0.2064 | 0.905 | 0.284 | 0.247 | 0.095 | 0.716 |
| 85 | 0.1964 | 0.2030 | 0.2142 | 0.772 | 0.072 | 0.049 | 0.228 | 0.928 |
| 90 | 0.1988 | 0.2108 | 0.2115 | 0.517 | 0.099 | 0.597 | 0.483 | 0.901 |
| 95 | 0.2000 | 0.2134 | 0.2202 | 0.565 | 0.044 | 0.240 | 0.435 | 0.956 |
| 100 | 0.1980 | 0.2051 | 0.2070 | 0.581 | 0.361 | 0.668 | 0.419 | 0.639 |

*(b)* TF-IDF terms scoring

*Table C.17.* Tabular Results: Hundreds users, Freenode, $\chi^2$ classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0616 | 0.0833 | 0.0892 | 0.205 | 0.041 | 0.994 | 0.795 | 0.959 |
| 10 | 0.0697 | 0.0889 | 0.0984 | 0.641 | 0.470 | 1.000 | 0.360 | 0.530 |
| 15 | 0.0798 | 0.0914 | 0.1037 | 0.991 | 0.759 | 1.000 | 0.009 | 0.241 |
| 20 | 0.0909 | 0.1094 | 0.1195 | 0.998 | 0.896 | 1.000 | 0.002 | 0.104 |
| 25 | 0.0967 | 0.1156 | 0.1139 | 0.998 | 0.990 | 1.000 | 0.002 | 0.010 |
| 30 | 0.1054 | 0.1153 | 0.1149 | 1.000 | 0.998 | 1.000 | 0.000 | 0.002 |
| 35 | 0.1046 | 0.1037 | 0.1013 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 40 | 0.1005 | 0.0936 | 0.0919 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 45 | 0.0915 | 0.0842 | 0.0811 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 50 | 0.0824 | 0.0810 | 0.0781 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 55 | 0.0770 | 0.0752 | 0.0725 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 60 | 0.0714 | 0.0677 | 0.0674 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 65 | 0.0718 | 0.0657 | 0.0633 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 70 | 0.0684 | 0.0624 | 0.0604 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 75 | 0.0643 | 0.0599 | 0.0576 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 80 | 0.0590 | 0.0572 | 0.0551 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 85 | 0.0552 | 0.0545 | 0.0533 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 90 | 0.0539 | 0.0564 | 0.0519 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 95 | 0.0561 | 0.0548 | 0.0509 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 100 | 0.0515 | 0.0539 | 0.0507 | 0.991 | 0.997 | 0.966 | 0.009 | 0.003 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0551 | 0.0775 | 0.0864 | 0.002 | 0.003 | 0.930 | 0.998 | 0.997 |
| 10 | 0.0661 | 0.0928 | 0.1059 | 0.386 | 0.044 | 0.991 | 0.614 | 0.956 |
| 15 | 0.0685 | 0.1012 | 0.1167 | 0.482 | 0.181 | 0.999 | 0.518 | 0.819 |
| 20 | 0.0911 | 0.1169 | 0.1217 | 0.886 | 0.706 | 1.000 | 0.114 | 0.294 |
| 25 | 0.1101 | 0.1238 | 0.1209 | 0.998 | 0.994 | 1.000 | 0.002 | 0.006 |
| 30 | 0.1103 | 0.1214 | 0.1174 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 35 | 0.1099 | 0.1115 | 0.1025 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 40 | 0.1028 | 0.0972 | 0.0946 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 45 | 0.0960 | 0.0841 | 0.0810 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 50 | 0.0914 | 0.0853 | 0.0782 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 55 | 0.0844 | 0.0786 | 0.0757 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 60 | 0.0777 | 0.0719 | 0.0699 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 65 | 0.0729 | 0.0669 | 0.0666 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 70 | 0.0684 | 0.0641 | 0.0633 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 75 | 0.0676 | 0.0587 | 0.0578 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 80 | 0.0623 | 0.0559 | 0.0554 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 85 | 0.0569 | 0.0575 | 0.0542 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 90 | 0.0557 | 0.0566 | 0.0549 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 95 | 0.0554 | 0.0565 | 0.0526 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 100 | 0.0515 | 0.0539 | 0.0507 | 0.991 | 0.997 | 0.966 | 0.009 | 0.003 |

*(b)* TF-IDF terms scoring

*Table C.18.* Tabular Results: Hundreds users, Freenode, Delta classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0500 | 0.0803 | 0.0934 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 |
| 10 | 0.0798 | 0.1085 | 0.1135 | 0.000 | 0.000 | 0.339 | 1.000 | 1.000 |
| 15 | 0.0949 | 0.1154 | 0.1185 | 0.000 | 0.000 | 0.791 | 1.000 | 1.000 |
| 20 | 0.1055 | 0.1213 | 0.1213 | 0.000 | 0.000 | 0.981 | 1.000 | 1.000 |
| 25 | 0.1096 | 0.1228 | 0.1207 | 0.000 | 0.002 | 1.000 | 1.000 | 0.998 |
| 30 | 0.1150 | 0.1249 | 0.1189 | 0.052 | 0.143 | 1.000 | 0.948 | 0.857 |
| 35 | 0.1212 | 0.1219 | 0.1208 | 0.889 | 0.772 | 1.000 | 0.111 | 0.228 |
| 40 | 0.1212 | 0.1219 | 0.1188 | 0.989 | 0.994 | 1.000 | 0.011 | 0.006 |
| 45 | 0.1189 | 0.1185 | 0.1146 | 1.000 | 0.999 | 1.000 | 0.000 | 0.001 |
| 50 | 0.1192 | 0.1168 | 0.1112 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 55 | 0.1212 | 0.1132 | 0.1123 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 60 | 0.1202 | 0.1123 | 0.1092 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 65 | 0.1196 | 0.1122 | 0.1105 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 70 | 0.1224 | 0.1131 | 0.1111 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 75 | 0.1242 | 0.1140 | 0.1098 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 80 | 0.1267 | 0.1133 | 0.1089 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 85 | 0.1293 | 0.1146 | 0.1115 | 1.000 | 0.983 | 1.000 | 0.000 | 0.017 |
| 90 | 0.1283 | 0.1152 | 0.1127 | 1.000 | 0.958 | 0.999 | 0.000 | 0.042 |
| 95 | 0.1284 | 0.1189 | 0.1081 | 1.000 | 0.998 | 1.000 | 0.001 | 0.002 |
| 100 | 0.1686 | 0.1444 | 0.1359 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0519 | 0.0774 | 0.0915 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 |
| 10 | 0.0799 | 0.0992 | 0.1106 | 0.000 | 0.000 | 0.013 | 1.000 | 1.000 |
| 15 | 0.0949 | 0.1111 | 0.1213 | 0.000 | 0.000 | 0.053 | 1.000 | 1.000 |
| 20 | 0.1061 | 0.1206 | 0.1250 | 0.000 | 0.000 | 0.827 | 1.000 | 1.000 |
| 25 | 0.1116 | 0.1202 | 0.1223 | 0.029 | 0.000 | 0.994 | 0.971 | 1.000 |
| 30 | 0.1151 | 0.1197 | 0.1203 | 0.110 | 0.003 | 1.000 | 0.890 | 0.997 |
| 35 | 0.1204 | 0.1193 | 0.1164 | 0.976 | 0.748 | 1.000 | 0.024 | 0.252 |
| 40 | 0.1225 | 0.1166 | 0.1135 | 1.000 | 0.973 | 1.000 | 0.000 | 0.027 |
| 45 | 0.1215 | 0.1132 | 0.1112 | 1.000 | 0.998 | 1.000 | 0.000 | 0.002 |
| 50 | 0.1244 | 0.1094 | 0.1075 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 55 | 0.1222 | 0.1083 | 0.1054 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 60 | 0.1245 | 0.1082 | 0.1006 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 65 | 0.1226 | 0.1074 | 0.1012 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 70 | 0.1240 | 0.1128 | 0.1013 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 75 | 0.1262 | 0.1111 | 0.1029 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 80 | 0.1321 | 0.1141 | 0.1059 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 85 | 0.1365 | 0.1191 | 0.1078 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 90 | 0.1406 | 0.1247 | 0.1156 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 95 | 0.1509 | 0.1328 | 0.1244 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 100 | 0.1686 | 0.0970 | 0.1359 | 1.000 | 1.000 | 0.000 | 0.000 | 0.000 |

*(b)* TF-IDF terms scoring

*Table C.19.* Tabular Results: Hundreds users, Krijn, KLD classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0418 | 0.0644 | 0.0787 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 |
| 10 | 0.0641 | 0.0841 | 0.0880 | 0.000 | 0.000 | 0.905 | 1.000 | 1.000 |
| 15 | 0.0760 | 0.0920 | 0.0930 | 0.000 | 0.000 | 0.997 | 1.000 | 1.000 |
| 20 | 0.0818 | 0.0945 | 0.0932 | 0.002 | 0.000 | 1.000 | 0.998 | 1.000 |
| 25 | 0.0854 | 0.0943 | 0.0899 | 0.023 | 0.047 | 1.000 | 0.977 | 0.953 |
| 30 | 0.0898 | 0.0896 | 0.0896 | 0.669 | 0.383 | 1.000 | 0.331 | 0.617 |
| 35 | 0.0929 | 0.0870 | 0.0885 | 0.992 | 0.915 | 1.000 | 0.008 | 0.085 |
| 40 | 0.0920 | 0.0813 | 0.0816 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 45 | 0.0926 | 0.0787 | 0.0791 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 50 | 0.0892 | 0.0777 | 0.0752 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 55 | 0.0862 | 0.0750 | 0.0739 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 60 | 0.0853 | 0.0703 | 0.0706 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 65 | 0.0827 | 0.0715 | 0.0686 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 70 | 0.0823 | 0.0712 | 0.0649 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 75 | 0.0800 | 0.0677 | 0.0611 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 80 | 0.0757 | 0.0651 | 0.0574 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 85 | 0.0741 | 0.0602 | 0.0531 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 90 | 0.0681 | 0.0558 | 0.0506 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 95 | 0.0659 | 0.0574 | 0.0516 | 0.999 | 1.000 | 1.000 | 0.001 | 0.000 |
| 100 | 0.1305 | 0.1222 | 0.1163 | 1.000 | 0.996 | 1.000 | 0.000 | 0.004 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0467 | 0.0717 | 0.0876 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 |
| 10 | 0.0701 | 0.0949 | 0.1046 | 0.000 | 0.000 | 0.035 | 1.000 | 1.000 |
| 15 | 0.0846 | 0.1048 | 0.1203 | 0.000 | 0.000 | 0.001 | 1.000 | 1.000 |
| 20 | 0.0954 | 0.1181 | 0.1223 | 0.000 | 0.000 | 0.678 | 1.000 | 1.000 |
| 25 | 0.1010 | 0.1170 | 0.1185 | 0.000 | 0.000 | 0.986 | 1.000 | 1.000 |
| 30 | 0.1043 | 0.1156 | 0.1154 | 0.000 | 0.000 | 0.998 | 1.000 | 1.000 |
| 35 | 0.1115 | 0.1147 | 0.1151 | 0.435 | 0.002 | 0.991 | 0.565 | 0.998 |
| 40 | 0.1124 | 0.1162 | 0.1123 | 0.286 | 0.057 | 0.998 | 0.714 | 0.943 |
| 45 | 0.1134 | 0.1135 | 0.1089 | 0.646 | 0.416 | 1.000 | 0.353 | 0.584 |
| 50 | 0.1150 | 0.1132 | 0.1090 | 0.924 | 0.459 | 0.998 | 0.076 | 0.541 |
| 55 | 0.1100 | 0.1070 | 0.1067 | 0.992 | 0.705 | 1.000 | 0.008 | 0.295 |
| 60 | 0.1131 | 0.1056 | 0.1076 | 1.000 | 0.942 | 1.000 | 0.000 | 0.058 |
| 65 | 0.1116 | 0.1066 | 0.1041 | 1.000 | 0.877 | 0.995 | 0.000 | 0.123 |
| 70 | 0.1149 | 0.1063 | 0.1005 | 1.000 | 0.913 | 1.000 | 0.000 | 0.087 |
| 75 | 0.1116 | 0.1073 | 0.1028 | 0.962 | 0.424 | 0.989 | 0.038 | 0.576 |
| 80 | 0.1110 | 0.1064 | 0.0985 | 0.991 | 0.818 | 0.994 | 0.009 | 0.182 |
| 85 | 0.1128 | 0.1066 | 0.0992 | 0.997 | 0.873 | 0.980 | 0.003 | 0.127 |
| 90 | 0.1108 | 0.1051 | 0.1020 | 0.987 | 0.874 | 0.986 | 0.013 | 0.126 |
| 95 | 0.1159 | 0.1104 | 0.1031 | 1.000 | 0.992 | 0.996 | 0.000 | 0.008 |
| 100 | 0.1305 | 0.0864 | 0.1163 | 1.000 | 0.996 | 0.000 | 0.000 | 0.004 |

*(b)* TF-IDF terms scoring

*Table C.20.* Tabular Results: Hundreds users, Krijn, $\chi^2$ classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0331 | 0.0541 | 0.0612 | 0.000 | 0.000 | 0.911 | 1.000 | 1.000 |
| 10 | 0.0573 | 0.0703 | 0.0709 | 0.076 | 0.005 | 1.000 | 0.924 | 0.995 |
| 15 | 0.0680 | 0.0748 | 0.0744 | 0.566 | 0.229 | 1.000 | 0.434 | 0.771 |
| 20 | 0.0709 | 0.0803 | 0.0732 | 0.932 | 0.829 | 1.000 | 0.068 | 0.171 |
| 25 | 0.0693 | 0.0692 | 0.0632 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 30 | 0.0648 | 0.0659 | 0.0596 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 35 | 0.0648 | 0.0603 | 0.0549 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 40 | 0.0610 | 0.0547 | 0.0494 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 45 | 0.0561 | 0.0459 | 0.0439 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 50 | 0.0489 | 0.0402 | 0.0381 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 55 | 0.0462 | 0.0371 | 0.0359 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 60 | 0.0437 | 0.0346 | 0.0328 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 65 | 0.0414 | 0.0320 | 0.0308 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 70 | 0.0400 | 0.0319 | 0.0301 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 75 | 0.0389 | 0.0319 | 0.0282 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 80 | 0.0369 | 0.0304 | 0.0280 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 85 | 0.0356 | 0.0295 | 0.0273 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 90 | 0.0350 | 0.0285 | 0.0265 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 95 | 0.0349 | 0.0288 | 0.0269 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 100 | 0.0370 | 0.0288 | 0.0275 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0336 | 0.0513 | 0.0616 | 0.000 | 0.000 | 0.781 | 1.000 | 1.000 |
| 10 | 0.0539 | 0.0684 | 0.0699 | 0.041 | 0.001 | 1.000 | 0.959 | 0.999 |
| 15 | 0.0653 | 0.0703 | 0.0716 | 0.641 | 0.132 | 1.000 | 0.359 | 0.869 |
| 20 | 0.0709 | 0.0771 | 0.0704 | 0.986 | 0.933 | 1.000 | 0.014 | 0.067 |
| 25 | 0.0668 | 0.0705 | 0.0650 | 0.999 | 0.997 | 1.000 | 0.001 | 0.003 |
| 30 | 0.0631 | 0.0666 | 0.0604 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 35 | 0.0626 | 0.0614 | 0.0560 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 40 | 0.0605 | 0.0527 | 0.0503 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 45 | 0.0541 | 0.0455 | 0.0440 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 50 | 0.0509 | 0.0406 | 0.0398 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 55 | 0.0461 | 0.0377 | 0.0361 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 60 | 0.0434 | 0.0354 | 0.0336 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 65 | 0.0411 | 0.0327 | 0.0313 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 70 | 0.0392 | 0.0316 | 0.0301 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 75 | 0.0378 | 0.0314 | 0.0284 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 80 | 0.0376 | 0.0302 | 0.0274 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 85 | 0.0364 | 0.0294 | 0.0266 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 90 | 0.0364 | 0.0290 | 0.0272 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 95 | 0.0356 | 0.0284 | 0.0269 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 100 | 0.0370 | 0.0197 | 0.0275 | 1.000 | 1.000 | 0.000 | 0.000 | 0.000 |

*(b)* TF-IDF terms scoring

*Table C.21.* Tabular Results: Hundreds users, Krijn, Delta classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0294 | 0.0400 | 0.0569 | 0.001 | 0.000 | 0.000 | 0.999 | 1.000 |
| 10 | 0.0509 | 0.0591 | 0.0831 | 0.020 | 0.000 | 0.000 | 0.980 | 1.000 |
| 15 | 0.0689 | 0.0821 | 0.0931 | 0.007 | 0.000 | 0.013 | 0.993 | 1.000 |
| 20 | 0.0818 | 0.0810 | 0.0982 | 0.611 | 0.001 | 0.003 | 0.389 | 0.999 |
| 25 | 0.0793 | 0.0702 | 0.0851 | 0.990 | 0.058 | 0.002 | 0.010 | 0.942 |
| 30 | 0.0703 | 0.0570 | 0.0720 | 1.000 | 0.124 | 0.008 | 0.000 | 0.876 |
| 35 | 0.0628 | 0.0516 | 0.0622 | 1.000 | 0.742 | 0.295 | 0.000 | 0.258 |
| 40 | 0.0527 | 0.0461 | 0.0558 | 1.000 | 0.804 | 0.397 | 0.000 | 0.196 |
| 45 | 0.0499 | 0.0386 | 0.0523 | 1.000 | 0.982 | 0.394 | 0.000 | 0.018 |
| 50 | 0.0465 | 0.0366 | 0.0487 | 1.000 | 0.999 | 0.167 | 0.000 | 0.001 |
| 55 | 0.0450 | 0.0360 | 0.0478 | 1.000 | 1.000 | 0.630 | 0.000 | 0.000 |
| 60 | 0.0428 | 0.0338 | 0.0462 | 1.000 | 1.000 | 0.891 | 0.000 | 0.000 |
| 65 | 0.0397 | 0.0348 | 0.0438 | 1.000 | 1.000 | 0.994 | 0.000 | 0.000 |
| 70 | 0.0386 | 0.0332 | 0.0416 | 1.000 | 1.000 | 0.975 | 0.000 | 0.000 |
| 75 | 0.0372 | 0.0336 | 0.0416 | 1.000 | 1.000 | 0.940 | 0.000 | 0.000 |
| 80 | 0.0391 | 0.0320 | 0.0427 | 1.000 | 1.000 | 0.694 | 0.000 | 0.000 |
| 85 | 0.0387 | 0.0323 | 0.0419 | 1.000 | 1.000 | 0.623 | 0.000 | 0.001 |
| 90 | 0.0377 | 0.0319 | 0.0401 | 1.000 | 0.995 | 0.581 | 0.000 | 0.005 |
| 95 | 0.0403 | 0.0367 | 0.0420 | 1.000 | 0.805 | 0.059 | 0.000 | 0.195 |
| 100 | 0.0838 | 0.0863 | 0.0999 | 0.991 | 0.035 | 0.017 | 0.009 | 0.965 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0346 | 0.0347 | 0.0576 | 0.346 | 0.000 | 0.000 | 0.654 | 1.000 |
| 10 | 0.0511 | 0.0573 | 0.0794 | 0.029 | 0.000 | 0.000 | 0.971 | 1.000 |
| 15 | 0.0655 | 0.0760 | 0.0901 | 0.008 | 0.000 | 0.005 | 0.992 | 1.000 |
| 20 | 0.0762 | 0.0757 | 0.0906 | 0.475 | 0.002 | 0.009 | 0.525 | 0.998 |
| 25 | 0.0777 | 0.0680 | 0.0838 | 0.969 | 0.027 | 0.005 | 0.031 | 0.973 |
| 30 | 0.0685 | 0.0637 | 0.0719 | 0.792 | 0.043 | 0.155 | 0.208 | 0.957 |
| 35 | 0.0650 | 0.0576 | 0.0651 | 0.995 | 0.623 | 0.352 | 0.005 | 0.377 |
| 40 | 0.0611 | 0.0542 | 0.0580 | 0.997 | 0.894 | 0.512 | 0.003 | 0.106 |
| 45 | 0.0626 | 0.0527 | 0.0614 | 1.000 | 0.992 | 0.569 | 0.000 | 0.009 |
| 50 | 0.0608 | 0.0518 | 0.0566 | 1.000 | 0.999 | 0.821 | 0.000 | 0.001 |
| 55 | 0.0589 | 0.0554 | 0.0558 | 1.000 | 0.999 | 0.813 | 0.000 | 0.001 |
| 60 | 0.0575 | 0.0564 | 0.0615 | 1.000 | 0.957 | 0.561 | 0.000 | 0.043 |
| 65 | 0.0551 | 0.0559 | 0.0622 | 0.994 | 0.627 | 0.649 | 0.006 | 0.373 |
| 70 | 0.0553 | 0.0565 | 0.0678 | 0.995 | 0.498 | 0.530 | 0.005 | 0.502 |
| 75 | 0.0594 | 0.0582 | 0.0700 | 0.994 | 0.448 | 0.214 | 0.007 | 0.552 |
| 80 | 0.0624 | 0.0601 | 0.0710 | 0.993 | 0.304 | 0.067 | 0.007 | 0.696 |
| 85 | 0.0617 | 0.0626 | 0.0777 | 0.963 | 0.029 | 0.018 | 0.037 | 0.971 |
| 90 | 0.0686 | 0.0664 | 0.0796 | 0.807 | 0.014 | 0.015 | 0.193 | 0.986 |
| 95 | 0.0746 | 0.0733 | 0.0884 | 0.904 | 0.001 | 0.000 | 0.096 | 0.999 |
| 100 | 0.0838 | 0.0863 | 0.0999 | 0.991 | 0.035 | 0.017 | 0.009 | 0.965 |

*(b)* TF-IDF terms scoring

*Table C.22.* Tabular Results: Hundreds users, Twitter, KLD classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0283 | 0.0348 | 0.0482 | 0.011 | 0.000 | 0.000 | 0.990 | 1.000 |
| 10 | 0.0450 | 0.0482 | 0.0629 | 0.090 | 0.000 | 0.000 | 0.910 | 1.000 |
| 15 | 0.0588 | 0.0671 | 0.0759 | 0.030 | 0.000 | 0.011 | 0.970 | 1.000 |
| 20 | 0.0694 | 0.0684 | 0.0839 | 0.564 | 0.001 | 0.001 | 0.436 | 0.999 |
| 25 | 0.0729 | 0.0670 | 0.0789 | 0.908 | 0.040 | 0.001 | 0.092 | 0.960 |
| 30 | 0.0691 | 0.0620 | 0.0757 | 0.962 | 0.012 | 0.001 | 0.038 | 0.988 |
| 35 | 0.0653 | 0.0600 | 0.0747 | 0.962 | 0.023 | 0.021 | 0.038 | 0.977 |
| 40 | 0.0622 | 0.0562 | 0.0703 | 0.983 | 0.023 | 0.040 | 0.017 | 0.977 |
| 45 | 0.0600 | 0.0545 | 0.0672 | 0.999 | 0.213 | 0.017 | 0.001 | 0.786 |
| 50 | 0.0590 | 0.0510 | 0.0634 | 1.000 | 0.403 | 0.015 | 0.000 | 0.597 |
| 55 | 0.0578 | 0.0496 | 0.0582 | 1.000 | 0.879 | 0.076 | 0.000 | 0.121 |
| 60 | 0.0550 | 0.0479 | 0.0487 | 1.000 | 0.953 | 0.316 | 0.000 | 0.047 |
| 65 | 0.0519 | 0.0466 | 0.0467 | 1.000 | 0.918 | 0.527 | 0.000 | 0.082 |
| 70 | 0.0493 | 0.0420 | 0.0411 | 1.000 | 0.989 | 0.521 | 0.000 | 0.011 |
| 75 | 0.0476 | 0.0413 | 0.0370 | 1.000 | 0.999 | 0.697 | 0.000 | 0.001 |
| 80 | 0.0437 | 0.0369 | 0.0323 | 1.000 | 0.999 | 0.883 | 0.000 | 0.001 |
| 85 | 0.0381 | 0.0332 | 0.0286 | 1.000 | 0.999 | 0.838 | 0.000 | 0.001 |
| 90 | 0.0328 | 0.0277 | 0.0269 | 1.000 | 1.000 | 0.859 | 0.000 | 0.000 |
| 95 | 0.0320 | 0.0264 | 0.0215 | 1.000 | 1.000 | 0.895 | 0.000 | 0.000 |
| 100 | 0.0999 | 0.1043 | 0.1215 | 0.767 | 0.000 | 0.000 | 0.233 | 1.000 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0347 | 0.0372 | 0.0566 | 0.114 | 0.000 | 0.000 | 0.886 | 1.000 |
| 10 | 0.0472 | 0.0603 | 0.0790 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 |
| 15 | 0.0597 | 0.0771 | 0.0955 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 |
| 20 | 0.0743 | 0.0818 | 0.1051 | 0.026 | 0.000 | 0.000 | 0.974 | 1.000 |
| 25 | 0.0766 | 0.0808 | 0.1030 | 0.090 | 0.000 | 0.000 | 0.910 | 1.000 |
| 30 | 0.0753 | 0.0824 | 0.0961 | 0.050 | 0.000 | 0.000 | 0.950 | 1.000 |
| 35 | 0.0731 | 0.0772 | 0.0939 | 0.196 | 0.000 | 0.000 | 0.804 | 1.000 |
| 40 | 0.0704 | 0.0767 | 0.0876 | 0.089 | 0.000 | 0.001 | 0.911 | 1.000 |
| 45 | 0.0743 | 0.0742 | 0.0905 | 0.645 | 0.000 | 0.000 | 0.355 | 1.000 |
| 50 | 0.0774 | 0.0721 | 0.0834 | 0.894 | 0.027 | 0.003 | 0.106 | 0.973 |
| 55 | 0.0745 | 0.0734 | 0.0790 | 0.938 | 0.136 | 0.003 | 0.062 | 0.865 |
| 60 | 0.0709 | 0.0743 | 0.0794 | 0.905 | 0.003 | 0.001 | 0.095 | 0.997 |
| 65 | 0.0715 | 0.0706 | 0.0780 | 0.865 | 0.003 | 0.002 | 0.136 | 0.997 |
| 70 | 0.0744 | 0.0738 | 0.0764 | 0.875 | 0.011 | 0.005 | 0.125 | 0.989 |
| 75 | 0.0741 | 0.0728 | 0.0759 | 0.794 | 0.007 | 0.007 | 0.206 | 0.993 |
| 80 | 0.0740 | 0.0733 | 0.0812 | 0.904 | 0.032 | 0.017 | 0.096 | 0.968 |
| 85 | 0.0750 | 0.0778 | 0.0889 | 0.742 | 0.003 | 0.003 | 0.258 | 0.997 |
| 90 | 0.0793 | 0.0809 | 0.0918 | 0.724 | 0.002 | 0.008 | 0.276 | 0.998 |
| 95 | 0.0838 | 0.0837 | 0.0978 | 0.736 | 0.001 | 0.000 | 0.264 | 0.999 |
| 100 | 0.0999 | 0.1043 | 0.1215 | 0.767 | 0.000 | 0.000 | 0.233 | 1.000 |

*(b)* TF-IDF terms scoring

*Table C.23.* Tabular Results: Hundreds users, Twitter, $\chi^2$ classifier

| Voc. % | method | | | *p*-value | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0254 | 0.0330 | 0.0494 | 0.010 | 0.000 | 0.000 | 0.991 | 1.000 |
| 10 | 0.0404 | 0.0487 | 0.0630 | 0.054 | 0.000 | 0.000 | 0.947 | 1.000 |
| 15 | 0.0587 | 0.0688 | 0.0719 | 0.082 | 0.004 | 0.469 | 0.918 | 0.996 |
| 20 | 0.0679 | 0.0603 | 0.0684 | 0.996 | 0.597 | 0.510 | 0.004 | 0.404 |
| 25 | 0.0629 | 0.0456 | 0.0505 | 1.000 | 0.999 | 0.778 | 0.000 | 0.001 |
| 30 | 0.0465 | 0.0341 | 0.0359 | 1.000 | 1.000 | 0.986 | 0.000 | 0.000 |
| 35 | 0.0387 | 0.0294 | 0.0263 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 40 | 0.0290 | 0.0220 | 0.0212 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 45 | 0.0242 | 0.0153 | 0.0170 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 50 | 0.0213 | 0.0121 | 0.0120 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 55 | 0.0174 | 0.0115 | 0.0111 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 60 | 0.0145 | 0.0096 | 0.0102 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 65 | 0.0123 | 0.0082 | 0.0088 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 70 | 0.0119 | 0.0081 | 0.0087 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 75 | 0.0098 | 0.0067 | 0.0073 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 80 | 0.0102 | 0.0066 | 0.0068 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 85 | 0.0084 | 0.0065 | 0.0060 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 90 | 0.0076 | 0.0059 | 0.0053 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 95 | 0.0063 | 0.0053 | 0.0047 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 100 | 0.0061 | 0.0042 | 0.0048 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |

*(a)* KLD terms scoring

| Voc. % | method | | | *p*-values | | | | |
|---|---|---|---|---|---|---|---|---|
| | x0 | x1 | x2 | x1 ≫ x0 | x2 ≫ x0 | x2 ≫ x1 | x0 ≫ x1 | x0 ≫ x2 |
| 5 | 0.0258 | 0.0291 | 0.0467 | 0.111 | 0.000 | 0.000 | 0.890 | 1.000 |
| 10 | 0.0356 | 0.0420 | 0.0550 | 0.100 | 0.000 | 0.005 | 0.900 | 1.000 |
| 15 | 0.0479 | 0.0581 | 0.0631 | 0.055 | 0.000 | 0.283 | 0.945 | 1.000 |
| 20 | 0.0613 | 0.0548 | 0.0620 | 0.997 | 0.396 | 0.335 | 0.003 | 0.604 |
| 25 | 0.0562 | 0.0411 | 0.0503 | 1.000 | 0.984 | 0.418 | 0.000 | 0.016 |
| 30 | 0.0436 | 0.0360 | 0.0368 | 1.000 | 0.999 | 0.983 | 0.000 | 0.001 |
| 35 | 0.0355 | 0.0272 | 0.0263 | 1.000 | 1.000 | 0.999 | 0.000 | 0.000 |
| 40 | 0.0273 | 0.0205 | 0.0211 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 45 | 0.0249 | 0.0153 | 0.0162 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 50 | 0.0211 | 0.0123 | 0.0124 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 55 | 0.0167 | 0.0116 | 0.0105 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 60 | 0.0137 | 0.0094 | 0.0106 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 65 | 0.0118 | 0.0082 | 0.0086 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 70 | 0.0117 | 0.0088 | 0.0086 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 75 | 0.0096 | 0.0070 | 0.0074 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 80 | 0.0096 | 0.0066 | 0.0068 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 85 | 0.0085 | 0.0064 | 0.0060 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 90 | 0.0074 | 0.0057 | 0.0058 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 95 | 0.0065 | 0.0052 | 0.0053 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 100 | 0.0061 | 0.0042 | 0.0048 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |

*(b)* TF-IDF terms scoring

*Table C.24.* Tabular Results: Hundreds users, Twitter, Delta classifier

# Appendix D

# Running times

*Sing us a song you're the piano man*
*Sing us a song tonight*
*Well we're all in the mood for a melody*
*And you've got us feeling alright*

Billy Joel

In this appendix we are concatenating samples of the log files generated during the experimental evaluation. We are reporting one example of log file for collection type (Associated Press -AP- for newspaper articles, Krijn and Twitter for conversational documents) and sorting algorithm (KLD). For simplicity we merged together lines of the same experimental settings with proper spacing. Each line describes:

- the stopwords strategy employed (0:TF, 3:NIDF, 5:Indri) or the sorting method employed (TF-IDF or KLD) with the percentage of vocabulary (5 to 100),
- the collection (AP, Krijn,Ttwitter) for which we are testing,
- the method under investigation (x0 for newspaper articles; x0, x1, x2 for conversational documents),
- the timestamp.

Sample of running times for 20 users, stopwords strategy.

```
Running stopwords_0 for ap method x0 Thu Dec 4 19:03:34 CET 2014
Running stopwords_3 for ap method x0 Thu Dec 4 19:04:00 CET 2014
Running stopwords_5 for ap method x0 Thu Dec 4 19:04:27 CET 2014

Running stopwords_3 for krijn method x0 Thu Dec 4 00:09:06 CET 2014
Running stopwords_3 for krijn method x1 Thu Dec 4 00:09:19 CET 2014
Running stopwords_3 for krijn method x2 Thu Dec 4 00:09:28 CET 2014
Running stopwords_5 for krijn method x0 Thu Dec 4 00:09:37 CET 2014
Running stopwords_5 for krijn method x1 Thu Dec 4 00:09:51 CET 2014
Running stopwords_5 for krijn method x2 Thu Dec 4 00:10:00 CET 2014

Running stopwords_0 for twitter method x0 Thu Dec 4 00:10:09 CET 2014
Running stopwords_0 for twitter method x1 Thu Dec 4 00:10:18 CET 2014
Running stopwords_0 for twitter method x2 Thu Dec 4 00:10:25 CET 2014
Running stopwords_3 for twitter method x0 Thu Dec 4 00:10:32 CET 2014
Running stopwords_3 for twitter method x1 Thu Dec 4 00:10:41 CET 2014
Running stopwords_3 for twitter method x2 Thu Dec 4 00:10:47 CET 2014
Running stopwords_5 for twitter method x0 Thu Dec 4 00:10:54 CET 2014
Running stopwords_5 for twitter method x1 Thu Dec 4 00:11:03 CET 2014
Running stopwords_5 for twitter method x2 Thu Dec 4 00:11:10 CET 2014
```

Sample of running times for hundreds users, stopwords strategy.

```
Running stopwords_0 for ap method x0 Sun Aug 10 01:42:56 CEST 2014
Running stopwords_3 for ap method x0 Sun Aug 10 01:58:38 CEST 2014
Running stopwords_5 for ap method x0 Sun Aug 10 02:13:47 CEST 2014

Running stopwords_0 for krijn method x0 Tue Jun 17 01:15:19 CEST 2014
Running stopwords_0 for krijn method x1 Tue Jun 17 01:17:23 CEST 2014
Running stopwords_0 for krijn method x2 Tue Jun 17 01:19:43 CEST 2014
Running stopwords_3 for krijn method x0 Tue Jun 17 01:22:31 CEST 2014
Running stopwords_3 for krijn method x1 Tue Jun 17 01:24:33 CEST 2014
Running stopwords_3 for krijn method x2 Tue Jun 17 01:26:55 CEST 2014
Running stopwords_5 for krijn method x0 Tue Jun 17 01:29:27 CEST 2014
Running stopwords_5 for krijn method x1 Tue Jun 17 01:31:34 CEST 2014
Running stopwords_5 for krijn method x2 Tue Jun 17 01:34:05 CEST 2014

Running stopwords_0 for twitter method x0 Sat Aug 9 21:48:52 CEST 2014
Running stopwords_0 for twitter method x1 Sat Aug 9 21:56:15 CEST 2014
Running stopwords_0 for twitter method x2 Sat Aug 9 22:05:03 CEST 2014
Running stopwords_3 for twitter method x0 Sat Aug 9 22:14:27 CEST 2014
Running stopwords_3 for twitter method x1 Sat Aug 9 22:21:51 CEST 2014
Running stopwords_3 for twitter method x2 Sat Aug 9 22:30:32 CEST 2014
—
Running stopwords_5 for twitter method x0 Mon Jul 14 04:33:22 CEST 2014
Running stopwords_5 for twitter method x1 Mon Jul 14 04:41:29 CEST 2014
Running stopwords_5 for twitter method x2 Mon Jul 14 05:15:18 CEST 2014
```

Sample of running times for 20 users, vocabulary selection strategy.

```
Running kld_5  for ap method x0 Sun Dec 7 00:37:47 CET 2014
Running kld_10 for ap method x0 Sun Dec 7 00:38:17 CET 2014
Running kld_15 for ap method x0 Sun Dec 7 00:38:47 CET 2014
Running kld_20 for ap method x0 Sun Dec 7 00:39:16 CET 2014
Running kld_25 for ap method x0 Sun Dec 7 00:39:48 CET 2014
Running kld_30 for ap method x0 Sun Dec 7 00:40:21 CET 2014
Running kld_35 for ap method x0 Sun Dec 7 00:40:49 CET 2014
Running kld_40 for ap method x0 Sun Dec 7 00:41:20 CET 2014
Running kld_45 for ap method x0 Sun Dec 7 00:41:51 CET 2014
Running kld_50 for ap method x0 Sun Dec 7 00:42:23 CET 2014
Running kld_55 for ap method x0 Sun Dec 7 00:42:52 CET 2014
Running kld_60 for ap method x0 Sun Dec 7 00:43:23 CET 2014
Running kld_65 for ap method x0 Sun Dec 7 00:43:51 CET 2014
Running kld_70 for ap method x0 Sun Dec 7 00:44:23 CET 2014
Running kld_75 for ap method x0 Sun Dec 7 00:44:53 CET 2014
Running kld_80 for ap method x0 Sun Dec 7 00:45:25 CET 2014
Running kld_85 for ap method x0 Sun Dec 7 00:45:57 CET 2014
Running kld_90 for ap method x0 Sun Dec 7 00:46:26 CET 2014
Running kld_95 for ap method x0 Sun Dec 7 00:46:58 CET 2014
Running kld_100 for ap method x0 Sun Dec 7 00:47:30 CET 2014


Running kld_5  for krijn method x0 Sat Dec 6 22:42:27 CET 2014
Running kld_5  for krijn method x1 Sat Dec 6 22:42:42 CET 2014
Running kld_5  for krijn method x2 Sat Dec 6 22:42:52 CET 2014
Running kld_10 for krijn method x0 Sat Dec 6 22:43:02 CET 2014
Running kld_10 for krijn method x1 Sat Dec 6 22:43:18 CET 2014
Running kld_10 for krijn method x2 Sat Dec 6 22:43:28 CET 2014
Running kld_15 for krijn method x0 Sat Dec 6 22:43:39 CET 2014
Running kld_15 for krijn method x1 Sat Dec 6 22:43:55 CET 2014
Running kld_15 for krijn method x2 Sat Dec 6 22:44:05 CET 2014
Running kld_20 for krijn method x0 Sat Dec 6 22:44:16 CET 2014
Running kld_20 for krijn method x1 Sat Dec 6 22:44:31 CET 2014
Running kld_20 for krijn method x2 Sat Dec 6 22:44:42 CET 2014
Running kld_25 for krijn method x0 Sat Dec 6 22:44:53 CET 2014
Running kld_25 for krijn method x1 Sat Dec 6 22:45:08 CET 2014
Running kld_25 for krijn method x2 Sat Dec 6 22:45:18 CET 2014
Running kld_30 for krijn method x0 Sat Dec 6 22:45:29 CET 2014
Running kld_30 for krijn method x1 Sat Dec 6 22:45:45 CET 2014
Running kld_30 for krijn method x2 Sat Dec 6 22:45:55 CET 2014
Running kld_35 for krijn method x0 Sat Dec 6 22:46:06 CET 2014
Running kld_35 for krijn method x1 Sat Dec 6 22:46:22 CET 2014
Running kld_35 for krijn method x2 Sat Dec 6 22:46:32 CET 2014
Running kld_40 for krijn method x0 Sat Dec 6 22:46:42 CET 2014
Running kld_40 for krijn method x1 Sat Dec 6 22:46:58 CET 2014
Running kld_40 for krijn method x2 Sat Dec 6 22:47:08 CET 2014
Running kld_45 for krijn method x0 Sat Dec 6 22:47:19 CET 2014
Running kld_45 for krijn method x1 Sat Dec 6 22:47:35 CET 2014
Running kld_45 for krijn method x2 Sat Dec 6 22:47:45 CET 2014
```

```
Running kld_50 for krijn method x0 Sat Dec 6 22:47:56 CET 2014
Running kld_50 for krijn method x1 Sat Dec 6 22:48:11 CET 2014
Running kld_50 for krijn method x2 Sat Dec 6 22:48:22 CET 2014
Running kld_55 for krijn method x0 Sat Dec 6 22:48:33 CET 2014
Running kld_55 for krijn method x1 Sat Dec 6 22:48:49 CET 2014
Running kld_55 for krijn method x2 Sat Dec 6 22:48:59 CET 2014
Running kld_60 for krijn method x0 Sat Dec 6 22:49:10 CET 2014
Running kld_60 for krijn method x1 Sat Dec 6 22:49:26 CET 2014
Running kld_60 for krijn method x2 Sat Dec 6 22:49:36 CET 2014
Running kld_65 for krijn method x0 Sat Dec 6 22:49:46 CET 2014
Running kld_65 for krijn method x1 Sat Dec 6 22:50:01 CET 2014
Running kld_65 for krijn method x2 Sat Dec 6 22:50:11 CET 2014
Running kld_70 for krijn method x0 Sat Dec 6 22:50:23 CET 2014
Running kld_70 for krijn method x1 Sat Dec 6 22:50:39 CET 2014
Running kld_70 for krijn method x2 Sat Dec 6 22:50:49 CET 2014
Running kld_75 for krijn method x0 Sat Dec 6 22:51:00 CET 2014
Running kld_75 for krijn method x1 Sat Dec 6 22:51:16 CET 2014
Running kld_75 for krijn method x2 Sat Dec 6 22:51:26 CET 2014
Running kld_80 for krijn method x0 Sat Dec 6 22:51:37 CET 2014
Running kld_80 for krijn method x1 Sat Dec 6 22:51:53 CET 2014
Running kld_80 for krijn method x2 Sat Dec 6 22:52:04 CET 2014
Running kld_85 for krijn method x0 Sat Dec 6 22:52:15 CET 2014
Running kld_85 for krijn method x1 Sat Dec 6 22:52:30 CET 2014
Running kld_85 for krijn method x2 Sat Dec 6 22:52:41 CET 2014
Running kld_90 for krijn method x0 Sat Dec 6 22:52:52 CET 2014
Running kld_90 for krijn method x1 Sat Dec 6 22:53:08 CET 2014
Running kld_90 for krijn method x2 Sat Dec 6 22:53:19 CET 2014
Running kld_95 for krijn method x0 Sat Dec 6 22:53:30 CET 2014
Running kld_95 for krijn method x1 Sat Dec 6 22:53:46 CET 2014
Running kld_95 for krijn method x2 Sat Dec 6 22:53:57 CET 2014
Running kld_100 for krijn method x0 Sat Dec 6 22:54:07 CET 2014
Running kld_100 for krijn method x1 Sat Dec 6 22:54:23 CET 2014
Running kld_100 for krijn method x2 Sat Dec 6 22:54:34 CET 2014


Running kld_5  for twitter method x0 Sat Dec 6 22:54:45 CET 2014
Running kld_5  for twitter method x1 Sat Dec 6 22:54:55 CET 2014
Running kld_5  for twitter method x2 Sat Dec 6 22:55:02 CET 2014
Running kld_10 for twitter method x0 Sat Dec 6 22:55:09 CET 2014
Running kld_10 for twitter method x1 Sat Dec 6 22:55:19 CET 2014
Running kld_10 for twitter method x2 Sat Dec 6 22:55:27 CET 2014
Running kld_15 for twitter method x0 Sat Dec 6 22:55:34 CET 2014
Running kld_15 for twitter method x1 Sat Dec 6 22:55:45 CET 2014
Running kld_15 for twitter method x2 Sat Dec 6 22:55:53 CET 2014
Running kld_20 for twitter method x0 Sat Dec 6 22:56:01 CET 2014
Running kld_20 for twitter method x1 Sat Dec 6 22:56:11 CET 2014
Running kld_20 for twitter method x2 Sat Dec 6 22:56:19 CET 2014
Running kld_25 for twitter method x0 Sat Dec 6 22:56:27 CET 2014
Running kld_25 for twitter method x1 Sat Dec 6 22:56:39 CET 2014
Running kld_25 for twitter method x2 Sat Dec 6 22:56:46 CET 2014
Running kld_30 for twitter method x0 Sat Dec 6 22:56:54 CET 2014
```

```
Running kld_30 for twitter method x1 Sat Dec 6 22:57:06 CET 2014
Running kld_30 for twitter method x2 Sat Dec 6 22:57:14 CET 2014
Running kld_35 for twitter method x0 Sat Dec 6 22:57:23 CET 2014
Running kld_35 for twitter method x1 Sat Dec 6 22:57:35 CET 2014
Running kld_35 for twitter method x2 Sat Dec 6 22:57:43 CET 2014
Running kld_40 for twitter method x0 Sat Dec 6 22:57:52 CET 2014
Running kld_40 for twitter method x1 Sat Dec 6 22:58:02 CET 2014
Running kld_40 for twitter method x2 Sat Dec 6 22:58:11 CET 2014
Running kld_45 for twitter method x0 Sat Dec 6 22:58:20 CET 2014
Running kld_45 for twitter method x1 Sat Dec 6 22:58:31 CET 2014
Running kld_45 for twitter method x2 Sat Dec 6 22:58:39 CET 2014
Running kld_50 for twitter method x0 Sat Dec 6 22:58:47 CET 2014
Running kld_50 for twitter method x1 Sat Dec 6 22:58:59 CET 2014
Running kld_50 for twitter method x2 Sat Dec 6 22:59:07 CET 2014
Running kld_55 for twitter method x0 Sat Dec 6 22:59:16 CET 2014
Running kld_55 for twitter method x1 Sat Dec 6 22:59:28 CET 2014
Running kld_55 for twitter method x2 Sat Dec 6 22:59:37 CET 2014
Running kld_60 for twitter method x0 Sat Dec 6 22:59:46 CET 2014
Running kld_60 for twitter method x1 Sat Dec 6 22:59:58 CET 2014
Running kld_60 for twitter method x2 Sat Dec 6 23:00:06 CET 2014
Running kld_65 for twitter method x0 Sat Dec 6 23:00:15 CET 2014
Running kld_65 for twitter method x1 Sat Dec 6 23:00:26 CET 2014
Running kld_65 for twitter method x2 Sat Dec 6 23:00:35 CET 2014
Running kld_70 for twitter method x0 Sat Dec 6 23:00:44 CET 2014
Running kld_70 for twitter method x1 Sat Dec 6 23:00:56 CET 2014
Running kld_70 for twitter method x2 Sat Dec 6 23:01:05 CET 2014
Running kld_75 for twitter method x0 Sat Dec 6 23:01:14 CET 2014
Running kld_75 for twitter method x1 Sat Dec 6 23:01:26 CET 2014
Running kld_75 for twitter method x2 Sat Dec 6 23:01:34 CET 2014
Running kld_80 for twitter method x0 Sat Dec 6 23:01:43 CET 2014
Running kld_80 for twitter method x1 Sat Dec 6 23:01:55 CET 2014
Running kld_80 for twitter method x2 Sat Dec 6 23:02:03 CET 2014
Running kld_85 for twitter method x0 Sat Dec 6 23:02:12 CET 2014
Running kld_85 for twitter method x1 Sat Dec 6 23:02:24 CET 2014
Running kld_85 for twitter method x2 Sat Dec 6 23:02:32 CET 2014
Running kld_90 for twitter method x0 Sat Dec 6 23:02:41 CET 2014
Running kld_90 for twitter method x1 Sat Dec 6 23:02:53 CET 2014
Running kld_90 for twitter method x2 Sat Dec 6 23:03:02 CET 2014
Running kld_95 for twitter method x0 Sat Dec 6 23:03:11 CET 2014
Running kld_95 for twitter method x1 Sat Dec 6 23:03:23 CET 2014
Running kld_95 for twitter method x2 Sat Dec 6 23:03:32 CET 2014
Running kld_100 for twitter method x0 Sat Dec 6 23:03:41 CET 2014
Running kld_100 for twitter method x1 Sat Dec 6 23:03:53 CET 2014
Running kld_100 for twitter method x2 Sat Dec 6 23:04:02 CET 2014
```

Sample of running times for hundreds users, vocabulary selection strategy.

```
Running kld_5  for ap method x0 Sun Aug 10 03:22:19 CEST 2014
Running kld_10 for ap method x0 Sun Aug 10 10:44:14 CEST 2014
Running kld_15 for ap method x0 Sun Aug 10 21:44:37 CEST 2014
Running kld_20 for ap method x0 Mon Aug 11 11:40:53 CEST 2014
Running kld_25 for ap method x0 Tue Aug 12 04:41:51 CEST 2014
Running kld_30 for ap method x0 Tue Aug 12 21:49:44 CEST 2014
Running kld_35 for ap method x0 Wed Aug 13 21:16:07 CEST 2014
Running kld_40 for ap method x0 Thu Aug 14 17:35:22 CEST 2014
Running kld_45 for ap method x0 Fri Aug 15 12:53:03 CEST 2014
Running kld_50 for ap method x0 Sat Aug 16 11:23:28 CEST 2014
Running kld_55 for ap method x0 Sun Aug 17 14:48:45 CEST 2014
Running kld_60 for ap method x0 Mon Aug 18 13:34:05 CEST 2014
Running kld_65 for ap method x0 Tue Aug 19 16:47:14 CEST 2014
Running kld_70 for ap method x0 Thu Aug 21 00:56:57 CEST 2014
Running kld_75 for ap method x0 Fri Aug 22 03:29:35 CEST 2014
Running kld_80 for ap method x0 Sat Aug 23 06:53:09 CEST 2014
Running kld_85 for ap method x0 Sun Aug 24 05:28:13 CEST 2014
Running kld_90 for ap method x0 Mon Aug 25 09:07:58 CEST 2014
Running kld_95 for ap method x0 Tue Aug 26 12:11:06 CEST 2014
Running kld_100 for ap method x0 Wed Aug 27 14:12:07 CEST 2014


Running kld_5  for krijn method x0 Wed Apr 9 21:46:26 CEST 2014
Running kld_5  for krijn method x1 Wed Apr 9 21:51:15 CEST 2014
Running kld_5  for krijn method x2 Wed Apr 9 21:56:45 CEST 2014
Running kld_10 for krijn method x0 Wed Apr 9 22:02:31 CEST 2014
Running kld_10 for krijn method x1 Wed Apr 9 22:04:55 CEST 2014
Running kld_10 for krijn method x2 Wed Apr 9 22:08:59 CEST 2014
Running kld_15 for krijn method x0 Wed Apr 9 22:14:05 CEST 2014
Running kld_15 for krijn method x1 Wed Apr 9 22:16:55 CEST 2014
Running kld_15 for krijn method x2 Wed Apr 9 22:18:41 CEST 2014
Running kld_20 for krijn method x0 Wed Apr 9 22:35:21 CEST 2014
Running kld_20 for krijn method x1 Wed Apr 9 22:38:40 CEST 2014
Running kld_20 for krijn method x2 Wed Apr 9 22:46:02 CEST 2014
Running kld_25 for krijn method x0 Wed Apr 9 22:56:52 CEST 2014
Running kld_25 for krijn method x1 Wed Apr 9 23:01:24 CEST 2014
Running kld_25 for krijn method x2 Wed Apr 9 23:12:48 CEST 2014
Running kld_30 for krijn method x0 Wed Apr 9 23:27:13 CEST 2014
Running kld_30 for krijn method x1 Wed Apr 9 23:33:47 CEST 2014
Running kld_30 for krijn method x2 Wed Apr 9 23:49:39 CEST 2014
Running kld_35 for krijn method x0 Thu Apr 10 00:12:20 CEST 2014
Running kld_35 for krijn method x1 Thu Apr 10 00:21:26 CEST 2014
Running kld_35 for krijn method x2 Thu Apr 10 00:45:05 CEST 2014
Running kld_40 for krijn method x0 Thu Apr 10 01:16:55 CEST 2014
Running kld_40 for krijn method x1 Thu Apr 10 01:30:58 CEST 2014
Running kld_40 for krijn method x2 Thu Apr 10 02:06:53 CEST 2014
Running kld_45 for krijn method x0 Thu Apr 10 02:54:07 CEST 2014
Running kld_45 for krijn method x1 Thu Apr 10 03:13:35 CEST 2014
Running kld_45 for krijn method x2 Thu Apr 10 04:00:51 CEST 2014
```

```
Running kld_50 for krijn method x0 Thu Apr 10 05:08:26 CEST 2014
Running kld_50 for krijn method x1 Thu Apr 10 05:36:27 CEST 2014
Running kld_50 for krijn method x2 Thu Apr 10 06:46:37 CEST 2014
Running kld_55 for krijn method x0 Thu Apr 10 08:20:03 CEST 2014
Running kld_55 for krijn method x1 Thu Apr 10 08:56:52 CEST 2014
Running kld_55 for krijn method x2 Thu Apr 10 10:24:26 CEST 2014
Running kld_60 for krijn method x0 Thu Apr 10 12:15:27 CEST 2014
Running kld_60 for krijn method x1 Thu Apr 10 13:05:20 CEST 2014
Running kld_60 for krijn method x2 Thu Apr 10 14:57:17 CEST 2014
Running kld_65 for krijn method x0 Thu Apr 10 17:17:35 CEST 2014
Running kld_65 for krijn method x1 Thu Apr 10 18:08:34 CEST 2014
Running kld_65 for krijn method x2 Thu Apr 10 19:57:37 CEST 2014
Running kld_70 for krijn method x0 Thu Apr 10 22:39:05 CEST 2014
Running kld_70 for krijn method x1 Thu Apr 10 23:57:40 CEST 2014
Running kld_70 for krijn method x2 Fri Apr 11 02:29:47 CEST 2014
Running kld_75 for krijn method x0 Fri Apr 11 05:42:46 CEST 2014
Running kld_75 for krijn method x1 Fri Apr 11 07:11:25 CEST 2014
Running kld_75 for krijn method x2 Fri Apr 11 10:01:15 CEST 2014
Running kld_80 for krijn method x0 Fri Apr 11 13:35:57 CEST 2014
Running kld_80 for krijn method x1 Fri Apr 11 15:18:19 CEST 2014
Running kld_80 for krijn method x2 Fri Apr 11 18:20:17 CEST 2014
Running kld_85 for krijn method x0 Fri Apr 11 22:08:20 CEST 2014
Running kld_85 for krijn method x1 Sat Apr 12 00:01:36 CEST 2014
Running kld_85 for krijn method x2 Sat Apr 12 03:20:58 CEST 2014
Running kld_90 for krijn method x0 Sat Apr 12 07:32:43 CEST 2014
Running kld_90 for krijn method x1 Sat Apr 12 09:45:24 CEST 2014
Running kld_90 for krijn method x2 Sat Apr 12 12:59:40 CEST 2014
Running kld_95 for krijn method x0 Sat Apr 12 16:44:22 CEST 2014
Running kld_95 for krijn method x1 Sat Apr 12 19:09:21 CEST 2014
Running kld_95 for krijn method x2 Sat Apr 12 23:03:16 CEST 2014
Running kld_100 for krijn method x0 Sun Apr 13 03:20:09 CEST 2014
Running kld_100 for krijn method x1 Sun Apr 13 05:47:12 CEST 2014
Running kld_100 for krijn method x2 Sun Apr 13 09:34:30 CEST 2014


Running kld_5  for twitter method x0 Mon Jun 23 00:54:46 CEST 2014
Running kld_5  for twitter method x1 Mon Jun 23 00:57:59 CEST 2014
Running kld_5  for twitter method x2 Mon Jun 23 01:02:48 CEST 2014
Running kld_10 for twitter method x0 Mon Jun 23 01:07:56 CEST 2014
Running kld_10 for twitter method x1 Mon Jun 23 01:11:29 CEST 2014
Running kld_10 for twitter method x2 Mon Jun 23 01:17:10 CEST 2014
Running kld_15 for twitter method x0 Mon Jun 23 01:24:15 CEST 2014
Running kld_15 for twitter method x1 Mon Jun 23 01:27:52 CEST 2014
Running kld_15 for twitter method x2 Mon Jun 23 01:34:58 CEST 2014
Running kld_20 for twitter method x0 Mon Jun 23 01:44:19 CEST 2014
Running kld_20 for twitter method x1 Mon Jun 23 01:48:32 CEST 2014
Running kld_20 for twitter method x2 Mon Jun 23 01:57:34 CEST 2014
Running kld_25 for twitter method x0 Mon Jun 23 02:13:24 CEST 2014
Running kld_25 for twitter method x1 Mon Jun 23 02:19:09 CEST 2014
Running kld_25 for twitter method x2 Mon Jun 23 02:35:03 CEST 2014
Running kld_30 for twitter method x0 Mon Jun 23 02:56:32 CEST 2014
```

```
Running kld_30 for twitter method x1 Mon Jun 23 03:04:35 CEST 2014
Running kld_30 for twitter method x2 Mon Jun 23 03:29:02 CEST 2014
Running kld_35 for twitter method x0 Mon Jun 23 04:06:01 CEST 2014
Running kld_35 for twitter method x1 Mon Jun 23 04:18:36 CEST 2014
Running kld_35 for twitter method x2 Mon Jun 23 04:53:40 CEST 2014
Running kld_40 for twitter method x0 Mon Jun 23 06:05:41 CEST 2014
Running kld_40 for twitter method x1 Mon Jun 23 06:27:28 CEST 2014
Running kld_40 for twitter method x2 Mon Jun 23 07:42:53 CEST 2014
Running kld_45 for twitter method x0 Mon Jun 23 09:35:18 CEST 2014
Running kld_45 for twitter method x1 Mon Jun 23 10:14:48 CEST 2014
Running kld_45 for twitter method x2 Mon Jun 23 12:29:11 CEST 2014
Running kld_50 for twitter method x0 Mon Jun 23 15:36:32 CEST 2014
Running kld_50 for twitter method x1 Mon Jun 23 16:39:52 CEST 2014
Running kld_50 for twitter method x2 Mon Jun 23 20:10:44 CEST 2014
Running kld_55 for twitter method x0 Tue Jun 24 01:15:07 CEST 2014
Running kld_55 for twitter method x1 Tue Jun 24 02:57:20 CEST 2014
Running kld_55 for twitter method x2 Tue Jun 24 08:30:14 CEST 2014
Running kld_60 for twitter method x0 Tue Jun 24 15:41:04 CEST 2014
Running kld_60 for twitter method x1 Tue Jun 24 18:10:20 CEST 2014
Running kld_60 for twitter method x2 Wed Jun 25 01:36:15 CEST 2014
Running kld_65 for twitter method x0 Wed Jun 25 12:12:29 CEST 2014
Running kld_65 for twitter method x1 Wed Jun 25 15:52:26 CEST 2014
Running kld_65 for twitter method x2 Thu Jun 26 02:19:31 CEST 2014
Running kld_70 for twitter method x0 Thu Jun 26 16:14:38 CEST 2014
Running kld_70 for twitter method x1 Thu Jun 26 21:10:08 CEST 2014
Running kld_70 for twitter method x2 Fri Jun 27 09:38:06 CEST 2014
Running kld_75 for twitter method x0 Sat Jun 28 01:32:25 CEST 2014
Running kld_75 for twitter method x1 Sat Jun 28 07:49:01 CEST 2014
Running kld_75 for twitter method x2 Sat Jun 28 21:31:43 CEST 2014
Running kld_80 for twitter method x0 Sun Jun 29 18:25:36 CEST 2014
Running kld_80 for twitter method x1 Mon Jun 30 02:36:25 CEST 2014
Running kld_80 for twitter method x2 Mon Jun 30 19:16:29 CEST 2014
Running kld_85 for twitter method x0 Tue Jul 1 17:04:51 CEST 2014
Running kld_85 for twitter method x1 Wed Jul 2 02:46:25 CEST 2014
Running kld_85 for twitter method x2 Wed Jul 2 23:33:47 CEST 2014
Running kld_90 for twitter method x0 Thu Jul 3 22:37:55 CEST 2014
Running kld_90 for twitter method x1 Fri Jul 4 10:27:32 CEST 2014
Running kld_90 for twitter method x2 Sat Jul 5 03:43:29 CEST 2014
Running kld_95 for twitter method x0 Sun Jul 6 04:11:50 CEST 2014
Running kld_95 for twitter method x1 Sun Jul 6 15:47:39 CEST 2014
Running kld_95 for twitter method x2 Mon Jul 7 12:44:43 CEST 2014
Running kld_100 for twitter method x0 Tue Jul 8 16:16:29 CEST 2014
Running kld_100 for twitter method x1 Wed Jul 9 07:37:16 CEST 2014
Running kld_100 for twitter method x2 Thu Jul 10 04:00:31 CEST 2014
```

# Bibliography


[1] S. Argamon. Interpreting burrows's delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147, 2008.

[2] S. Argamon and P. Juola. Overview of the international authorship identification competition at PAN-2011. In V. Petras, P. Forner, and P. D. Clough, editors, *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[3] S. Argamon and S. Levitan. Measuring the usefulness of function words for authorship attribution. In *In Proceedings of the 2005 ACH/ALLC Conference*, 2005.

[4] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 27–34. AUAI Press, 2009.

[5] D. V. Ayala, E. Castillo, D. Pinto, I. Olmos, and S. León. Information Retrieval and Classification based Approaches for the Sexual Predator Identification - Notebook for PAN at CLEF 2012. In Forner et al. [37].

[6] A. Bacchelli, T. Dal Sasso, M. D'Ambros, and M. Lanza. Content classification of development emails. In *Software Engineering (ICSE), 2012 34th International Conference on*, pages 375–385, 2012.

[7] R. Bache, F. Crestani, D. Canter, and D. Youngs. Mining police digital archives to link criminal styles with offender characteristics. In *Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers*, ICADL'07, pages 493–494. Springer-Verlag, 2007.

[8] R. Bache, F. Crestani, D. Canter, and D. Youngs. A language modelling approach to linking criminal styles with offender characteristics. *Data & Knowledge Engineering*, 69(3):303–315, 2010.

[9] B. W. Bader and P. A. Chew. *Algebraic Techniques for Multilingual Document Clustering*, pages 21–36. John Wiley & Sons, Ltd, 2010.

[10] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology behind Search*. Addison-Wesley Professional, 2 edition, 2011.

[11] K. Balog, M. Bron, J. He, K. Hofmann, E. J. Meij, M. de Rijke, E. Tsagkias, and W. Weerkamp. The University of Amsterdam at TREC 2009: Blog, Web, Entity, and Relevance Feedback. In *TREC 2009 Working Notes*. NIST, 2009.

[12] L. Barbosa and J. Feng. Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44. Association for Computational Linguistics, 2010.

[13] M. Berry. *Survey of Text Mining I: Clustering, Classification, and Retrieval*. Springer, 2004.

[14] M. Berry. *Survey of Text Mining II: Clustering, Classification, and Retrieval*. Springer, 2007.

[15] M. W. Berry and J. Kogan, editors. *Text Mining*. John Wiley & Sons, Ltd, 2010.

[16] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[17] D. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[19] S. R. Boutwell. Authorship attribution of short messages using multimodal features. Master's thesis, Naval Postgraduate School, 2009.

[20] A. Cano, M. Fernandez, and H. Alani. Detecting Child Grooming Behaviour Patterns on Social Media. *Social Informatics*, 2014.

[21] G. Carenini, G. Murray, and R. Ng. Methods for Mining and Summarizing Text Conversations. *Synthesis Lectures on Data Management*, 3(3):1–130, 2011.

[22] C.D.Manning and H.Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[23] Y. Cha and J. Cho. Social-network analysis using topic models. In *SIGIR '12: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 565–574, 2012.

[24] T. Chen and M.-Y. Kan. Creating a live, public short message service corpus: the NUS SMS corpus. *Language Resources and Evaluation*, 47(2):299–335, 2013.

[25] P. Clough, N. Ferro, P. Forner, J. Gonzalo, B. Huurnink, J. Kekäälåinen, M. Lalmas, V. Petras, and M. de Rijke. CLEF 2011: Conference on Multilingual and Multimodal Information Access Evaluation. *SIGIR Forum*, 45(2):32–37, 2012.

[26] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. John Wiley & Sons, 2006.

[27] B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison Wesley, 2009.

[28] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[29] N. A. Diakopoulos and D. A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1195–1198. ACM, 2010.

[30] J. S. Durham. Topic detection in online chat. Master's thesis, Naval Postgraduate School, 2009.

[31] M. Elsner. *Generalizing Local Coherence Modeling*. PhD thesis, Brown University, 2011.

[32] M. Elsner and E. Charniak. You Talking to Me? A Corpus and Algorithm for Conversation Disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842. Association for Computational Linguistics, 2008.

[33] M. Elsner and E. Charniak. Disentangling Chat. *Computational Linguistics*, 36 (3):389–409, 2010.

[34] G. Eriksson and J. Karlgren. Features for Modelling Characteristics of Conversations- Notebook for PAN at CLEF 2012. In Forner et al. [37].

[35] H. Escalante, I. LabTL, and L. No. Sexual predator detection in chats with chained classifiers. In *WASSA 2013*, pages 46–54, June 2013.

[36] A. Field and G. Hole. *How to Design and Report Experiments*. SAGE Publications Ltd, 2003.

[37] P. Forner, J. Karlgren, and C. Womser-Hacker, editors. *CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers, 17-20 September 2012, Rome, Italy*, 2012.

[38] E. N. Forsythand and C. H. Martell. Lexical and Discourse Analysis of Online Chat Dialog. In *International Conference on Semantic Computing (ICSC 2007)*, pages 19–26. IEEE, 2007.

[39] D. Gayo-Avello, D. J. Brenes, D. Fernández-Fernández, M. E. Fernández-Menéndez, and R. García-Suárez. De retibus socialibus et legibus momenti. *EPL (Europhysics Letters)*, 94(3):38001, 2011.

[40] S. Gerani. *Proximity-based approaches to blog opinion retrieval*. PhD thesis, Università della Svizzera italiana, 2012.

[41] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *ACL (Short Papers)*, pages 42–47. The Association for Computer Linguistics, 2011.

[42] T. Gollub, M. Potthast, A. Beyer, M. Busse, F. Rangel, P. Rosso, E. Stamatatos, and B. Stein. Recent trends in digital text forensics and its evaluation. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 282–302. Springer Berlin Heidelberg, 2013.

[43] A. Grewal, T. Allison, S. Dimitrov, and D. Radev. Multi-document summarization using off the shelf compression software. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop - Volume 5*, HLT-NAACL-DUC '03, pages 17–24. Association for Computational Linguistics, 2003.

[44] J. Grieve. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270, 2007.

[45] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, 2004.

[46] S. C. H. Haichao Dong and Y. He. Structural analysis of chat messages for topic detection. *Online Information Review*, 30(5):496–516, 2006.

[47] E. Hatcher and O. Gospodnetic. *Lucene in Action (In Action Series)*. Manning Publications Co., 2004.

[48] H. S. Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., Orlando, FL, USA, 1978.

[49] M. A. Hearst. Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 3–10. Association for Computational Linguistics, 1999.

[50] J. M. G. Hidalgo and A. A. C. Díaz. Combining Predation Heuristics and Chat-Like Features in Sexual Predator Identification - Notebook for PAN at CLEF 2012. In Forner et al. [37].

[51] D. I. Holmes. Authorship attribution. *Computers and the Humanities*, 28(2): 87–106, 1994.

[52] L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 832–840. ACM, 2011.

[53] D. L. Hoover. Delta prime? *Literary and Linguistic Computing*, 19(4):477–495, 2004.

[54] A. Hotho, A. Nürnberger, and G. Paaß. A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 2005.

[55] Y. How and M. yen Kan. Optimizing predictive text entry for short message service on mobile phones. In *in Human Computer Interfaces International (HCII 05). 2005: Las Vegas*, 2005.

[56] G. Inches and F. Crestani. Overview of the International Sexual Predator Identification Competition at PAN-2012. In Forner et al. [37].

[57] G. Inches, M. J. Carman, and F. Crestani. Statistics of online User-generated short Documents. In *ECIR '10: Proceedings of the 32nd European Conference on IR Research on Advances in Information Retrieval*, pages 649–652, 2010.

[58] G. Inches, A. Basso, and F. Crestani. On the generation of rich content metadata from social media. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents - SMUC '11*, pages 85–92, 2011.

[59] G. Inches, M. J. Carman, and F. Crestani. Investigating the Statistical Properties of User-Generated Documents. In *FAQS 2011: Proceedings of the 9th International Conference on Flexible Query Answering Systems*, pages 198–209, 2011.

[60] J.Codina, A.Kaltenbrunner, J.Grivolla, R. E.Banchs, and R.Baeza-Yates. Content analysis in web 2.0. In *18th International World Wide Web Conference*, 2009.

[61] E. P. Jiang. *Content-Based Spam Email Classification using Machine-Learning Algorithms*, pages 37–56. John Wiley & Sons, Ltd, 2010.

[62] P. Juola. What can we do with small corpora? Document categorization via cross-entropy. In *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization*. Department of Artificial Intelligence, University of Edinburgh, Edimburg, UK, 1997.

[63] P. Juola. Cross-entropy and linguistic typology. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, NeMLaP3/CoNLL '98, pages 141–149. Association for Computational Linguistics, 1998.

[64] P. Juola. Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334, 2006.

[65] P. Juola. An overview of the traditional authorship attribution subtask. In Forner et al. [37].

[66] P. Juola and H. Baayen. A controlled-corpus experiment in authorship identification by cross-entropy. In *Literary and Linguistic Computing*, volume 20(Suppl 1), pages 59–67. Kluwer Academic Publishers, 2003.

[67] D. Jurafsky and J. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson international edition. Pearson Prentice Hall/Pearson education international, 2008.

[68] I.-S. Kang, C.-K. Kim, S.-J. Kang, and S.-H. Na. IR-based k-Nearest Neighbor Approach for Identifying Abnormal Chat Users - Notebook for PAN at CLEF 2012. In Forner et al. [37].

[69] R. Kern, S. Klampfl, and M. Zechner. Vote/Veto Classification, Ensemble Clustering and Sequence Classification for Author Identification - Notebook for PAN at CLEF 2012. In Forner et al. [37].

[70] A. Kontostathis, L. Edwards, and A. Leatherman. Text mining and cybercrime. In *Text Mining*, pages 149–164. Wiley Online Library, 2010.

[71] A. Kontostathis, W. West, A. Garron, K. Reynolds, and L. Edwards. Identify Predators Using ChatCoder 2.0 - Notebook for PAN at CLEF 2012. In Forner et al. [37].

[72] M. Koppel, J. Schler, S. Argamon, and E. Messeri. Authorship attribution with thousands of candidate authors. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–660, 2006.

[73] M. Koppel, J. Schler, and S. Argamon. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26, 2009.

[74] T. Kucukyilmaz, B. Cambazoglu, C. Aykanat, and F. Can. Chat mining for gender prediction. *Advances in Information Systems*, pages 274–283, 2006.

[75] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, 44(4):1448–1466, 2008.

[76] M. Latapy. Quantifying Paedophile Queries in a Large P2P System. *System*, pages 401–405, 2011.

[77] R. Layton, P. Watters, and R. Dazeley. Authorship attribution for twitter in 140 characters or less. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second*, pages 1–8, 2010.

[78] R. Layton, S. McCombie, and P. Watters. Authorship attribution of irc messages using inverse author frequency. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2012 Third*, pages 7–13, 2012.

[79] J. Lin. *Automatic Author Profiling of Online Chat Logs*. PhD thesis, Naval Postgraduate School, 2007.

[80] R. T.-W. Lo, B. He, and I. Ounis. Automatically building a stopword list for an information retrieval system. *JDIM*, 3(1):3–8, 2005.

[81] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[82] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 227–236. ACM, 2011.

[83] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski. Learning to Identify Internet Sexual Predation. *International Journal of Electronic Commerce*, 15(3):103–122, 2011.

[84] T. C. Mendenhall. The characteristic curves of composition. *Science*, ns-9(214S): 237–246, 1887.

[85] G. Mikros and K. Perifanos. Authorship attribution in greek tweets using author's multilevel n-gram profiles. In *AAAI Spring Symposium Series*, 2013.

[86] C. Morris. Identifying online sexual predators by svm classification with lexical and behavioral features. Master's thesis, Department of Computer Science, University of Toronto, 2013.

[87] C. Morris and G. Hirst. Identifying Sexual Predators by SVM Classification with Lexical and Behavioral Features - Notebook for PAN at CLEF 2012. In Forner et al. [37].

[88] P. Mutton. *IRC Hacks*. O'Reilly Media, 2004.

[89] S. Nadali, M. A. A. Murad, N. M. Sharef, A. Mustapha, and S. Shojaee. A review of cyberbullying detection: An overview. *2013 13th International Conference on Intellient Systems Design and Applications*, pages 325–330, Dec. 2013.

[90] J. O'Neill and D. Martin. Text chat in action. In *GROUP '03: Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work*, pages 40–49. ACM, 2003.

[91] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proc. of NAACL*, 2013.

[92] J. Parapar, D. E. Losada, and A. Barreiro. A learning-based approach for the identification of sexual predators in chat logs - Notebook for PAN at CLEF 2012. In Forner et al. [37].

[93] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

[94] C. Peersman, W. Daelemans, and L. Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents - SMUC '11*, page 37. ACM Press, 2011.

[95] C. Peersman, F. Vaassen, V. V. Asch, and W. Daelemans. Conversation Level Constraints on Pedophile Detection in Chat Rooms - Notebook for PAN at CLEF 2012. In Forner et al. [37].

[96] N. Pendar. Toward Spotting the Pedophile Telling victim from predator in text chats. In *International Conference on Semantic Computing (ICSC 2007)*, pages 235–241. IEEE, 2007.

[97] M. Popescu and C. Grozea. Kernel Methods and String Kernels for Authorship Analysis - Notebook for PAN at CLEF 2012. In Forner et al. [37].

[98] M. Potthast, B. Stein, A. Eiselt, A. Barrón-Cedeño, and P. Rosso. Overview of the 1st International Competition on Plagiarism Detection. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel, and E. Agirre, editors, *SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, pages 1–9. CEUR-WS.org, 2009.

[99] M. Potthast, T. Gollub, M. Hagen, M. Tippmann, J. Kiesel, P. Rosso, E. Stamatatos, and B. Stein. Overview of the 5th International Competition on Plagiarism Detection. In *Working Notes Papers of the CLEF 2013 Evaluation Labs*, 2013.

[100] A. A. Puretskiy, G. L. Shutt, and M. W. Berry. *Survey of Text Visualization Techniques*, pages 105–127. John Wiley & Sons, Ltd, 2010.

[101] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

[102] D. Ramage, S. Dumais, and D. Liebling. Characterizing Microblogs with Topic Models. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 130–137. AAAI Press, 2010.

[103] B. Ramnath and r. K. Aleksande. "w00t! feeling great today!" chatter in twitter: Identification and prevalence. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, 2013.

[104] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. Overview of the Author Profiling Task at PAN 2013. In *CLEF (Online Working Notes/Labs/Workshop)*, 2013.

[105] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, pages 37–44. ACM, 2010.

[106] A. Reyes and P. Rosso. On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*, pages 1–20, 2013.

[107] A. Ritter, C. Cherry, and B. Dolan. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180. Association for Computational Linguistics, 2010.

[108] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[109] S. Rose, D. Engel, N. Cramer, and W. Cowley. *Automatic Keyword Extraction from Individual Documents*, pages 1–20. John Wiley & Sons, Ltd, 2010.

[110] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, UAI '04, pages 487–494. AUAI Press, 2004.

[111] G. Salton. *Automatic Information Organization and Retrieval*. McGraw Hill Text, 1968.

[112] C. Sanderson and S. Guenter. Short text authorship attribution via sequence kernels, markov chains and author unmasking: an investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 482–491. Association for Computational Linguistics, 2006.

[113] R. L. T. Santos, C. Macdonald, R. McCreadie, I. Ounis, and I. Soboroff. Information retrieval on the blogosphere. *Foundations and Trends in Information Retrieval*, 6(1):1–125, 2012.

[114] J. Savoy. Authorship Attribution Based on Specific Vocabulary. *ACM Transactions on Information Systems*, 30(2):1–30, 2012.

[115] R. Schwartz, O. Tsur, A. Rappoport, and M. Koppel. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1880–1891, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[116] M. Serrano, A. Flammini, and F. Menczer. Modeling statistical properties of written text. *PLoS ONE*, 4(4):e5372–, 2009.

[117] D. A. Shamma, L. Kennedy, and E. F. Churchill. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*, WSM '09, pages 3–10. ACM, 2009.

[118] C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, 2001.

[119] P. Shrestha, C. Jacquin, and B. Daille. Clustering short text and its evaluation. In *Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'12, pages 169–180. Springer-Verlag, 2012.

[120] R. S. Silva, G. Laboreiro, L. Sarmento, T. Grant, E. Oliveira, and B. Maia. ‘twazn me!!! ;(’ automatic authorship analysis of micro-blogging messages. In *Proceedings of the 16th International Conference on Natural Language Processing and Information Systems*, NLDB'11, pages 161–168, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-22326-6.

[121] E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[122] B. Stein, M. Koppel, and E. Stamatatos. Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07). *SIGIR Forum*, 41(2):68–71, 2007.

[123] V. H. Thuc and P. Srinivasan. Topic models and a revisit of text-related applications. In *PIKM '08: Proceeding of the 2nd PhD workshop on Information and knowledge management*, pages 25–32. ACM, 2008.

[124] A. P. S. Trevor K. M. Stone. Detection of topic change in irc chat logs. Website, 1993. http://www.trevorstone.org/school/ircsegmentation.pdf.

[125] V. H. Tuulos and H. Tirri. Combining topic models and social networks for chat data mining. In *WI '04: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 206–213. IEEE Computer Society, 2004.

[126] S. Ueberwasser. Non-standard data in swiss text messages with a special focus on dialectal forms. In M. Zampieri and S. Diwersy, editors, *Non-standard Data Sources in Corpus-based Research*, ZSM-Studien, Schriften des Zentrums Sprachenvielfalt und Mehrsprachigkeit der Universität zu Köln, pages 7–24. Shaker Verlag, Aachen, 2013.

[127] A. Vartapetiance and L. Gillam. Quite Simple Approaches for Authorship Attribution, Intrinsic Plagiarism Detection and Sexual Predator Identification - Notebook for PAN at CLEF 2012. In Forner et al. [37].

[128] A. Vartapetiance and L. Gillam. "Our Little Secret": pinpointing potential predators. *Security Informatics*, 3(1):3, 2014. ISSN 2190-8532.

[129] E. Villatoro-Tello, A. Juárez-González, H. J. Escalante, M. Montes-Y-Gómez, and L. Villaseñor-Pineda. A Two-step Approach for Effective Detection of Misbehaving Users in Chats - Notebook for PAN at CLEF 2012. In Forner et al. [37].

[130] E. Voorhees. The TREC-8 Question Answering Track Report. *TREC*, pages 77–82, 1999. URL http://trec.nist.gov/pubs/trec8/papers/qa_report.pdf.

[131] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT Press, 2005.

[132] H. M. Wallach. *Structured Topic Models for Language*. PhD thesis, University of Cambridge, 2008.

[133] L. Wang and D. W. Oard. Context-based message expansion for disentanglement of interleaved text conversations. In *NAACL '09*, pages 200–208. Association for Computational Linguistics, 2009.

[134] Y. Wang, H. Bai, M. Stanton, W.-Y. Chen, and E. Y. Chang. Plda: Parallel latent dirichlet allocation for large-scale applications. In *Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management*, AAIM '09, pages 301–314. Springer-Verlag, 2009.

[135] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, 2006.

[136] G. Wilcock. Introduction to linguistic annotation and text analytics. *Synthesis Lectures on Human Language Technologies*, 2(1):1–159, 2009.

[137] X. Yi and J. Allan. Evaluating topic models for information retrieval. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1431–1432, 2008.

[138] X. Yi and J. Allan. A comparative study of utilizing topic models for information retrieval. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 29–41, 2009.

[139] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on web 2.0. In *CAW 2.0 '09: Proceedings of the 1st Content Analysis in Web 2.0 Workshop*, 2009.

[140] W. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *ECIR'11: Proceedings of the 33rd European conference on Advances in information retrieval*, pages 338–349, 2011.

[141] Y. Zhao, J. Zobel, and P. Vines. Using relative entropy for authorship attribution. In *Proceedings of the Third Asia conference on Information Retrieval Technology*, AIRS'06, pages 92–105. Springer-Verlag, 2006.

[142] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3): 378–393, 2006.

[143] X. Zhu, Z.-Y. Ming, X. Zhu, and T.-S. Chua. Topic hierarchy construction for the organization of multi-source user generated contents. In *SIGIR '13: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 233–242, 2013.

*Morning Zombies, 2014*

*So should it be with you. When you have done all you have been commanded, say, 'We are unprofitable servants; we have done what we were obliged to do.'*

Lc 17,10