Tracking Public Opinion on Social Media

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
Anastasia Giachanou

under the supervision of
Prof. Fabio Crestani

October 2018

Dissertation Committee

Prof. Luca Maria Gambardella	Università della Svizzera Italiana, Switzerland
Prof. Evanthia Papadopoulou	Università della Svizzera Italiana, Switzerland
Prof. Gianni Amati	Fondazione Ugo Bordoni, Rome, Italy
Prof. Paolo Rosso	Universitat Politècnica de València, Valencia, Spain

Dissertation accepted on 16 October 2018

Research Advisor	PhD Program Director
Prof. Fabio Crestani	**Prof. Walter Binder and Prof. Olaf Schenk**

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.

Anastasia Giachanou
Lugano, 16 October 2018

I am among those who think that science has great beauty. A scientist in his laboratory is not only a technician, he is also a child place before natural phenomenon, which impress him like a fairy tale.

Marie Skłodowska-Curie,
1867-1934

Abstract

The increasing popularity of social media has changed the web from a static repository of information into a dynamic forum with continuously changing information. Social media platforms has given the capability to people expressing and sharing their thoughts and opinions on the web in a very simple way. The so-called User Generated Content is a good source of users opinion and mining it can be very useful for a wide variety of applications that require understanding the public opinion about a concept. For example, enterprises can capture the negative or positive opinions of customers about their services or products and improve their quality accordingly.

The dynamic nature of social media with the constantly changing vocabulary, makes developing tools that can automatically track public opinion a challenge. To help users better understand public opinion towards an entity or a topic, it is important to: a) find the related documents and the sentiment polarity expressed in them; b) identify the important time intervals where there is a change in the opinion; c) identify the causes of the opinion change; d) estimate the number of people that have a certain opinion about the entity; and e) measure the impact of public opinion towards the entity.

In this thesis we focus on the problem of tracking public opinion on social media and we propose and develop methods to address the different subproblems. First, we analyse the topical distribution of tweets to determine the number of topics that are discussed in a single tweet. Next, we propose a topic specific stylistic method to retrieve tweets that are relevant to a topic and also express opinion about it. Then, we explore the effectiveness of time series methodologies to track and forecast the evolution of sentiment towards a specific topic over time. In addition, we propose the LDA & KL-divergence approach to extract and rank the likely causes of sentiment spikes. We create a test collection that can be used to evaluate methodologies in ranking the likely reasons of sentiment spikes. To estimate the number of people that have a certain opinion about an entity, we propose an approach that uses pre-publication and post-publication features extracted from news posts and users' comments respectively. Finally, we

propose an approach that propagates sentiment signals to measure the impact of public opinion towards the entity's reputation. We evaluate our proposed methods on standard evaluation collections and provide evidence that the proposed methods improve the performance of the state-of-the-art approaches on tracking public opinion on social media.

Acknowledgements

I am grateful to my advisor Prof. Fabio Crestani who gave me the opportunity to join his research group and explore interesting research directions with freedom. I am very thankful to him for his constant trust and encouragement during my Ph.D. studies. I am also grateful to him for all the opportunities he provided to me to attend conferences and to meet well known researchers in the field of Information Retrieval.

I am grateful to my doctoral thesis committee, Gianni Amati, Paolo Rosso, Luca Maria Gambardella, and Evanthia Papadopoulou for reading my thesis and providing me with valuable feedback.

I gratefully acknowledge the financial support of the SNSF OpiTrack Project funded by Swiss National Science Foundation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Problem

Recent years have witnessed the rapid growth of social media platforms that have changed the way that people communicate and exchange information. Social media such as social network sites (e.g., Facebook, LinkedIn) and microblogs (e.g., Twitter) gave people the capability to express and share their thoughts and opinions on the web in a very simple way. The increasing popularity of social media has changed the web from a static repository of information into a dynamic forum with continuously changing information. The so-called *User Generated Content* varies a lot, from simple comments in Facebook posts to long publications in blogs.

Social media contain a tremendous amount of user generated content such as text, images, audio or video that is a good source of opinion and can be valuable for a variety of applications which require understanding the public standpoint about a concept [5]. One typical example that illustrates the importance of public opinion refers to enterprises that can capture the views of customers about their products or their competitors. This information can be used to improve the quality of their services or products accordingly. In addition, potential customers of a product can use the opinionated information to decide whether to buy the product or not. Public opinion is also useful for a government to understand the public view regarding different social issues and act promptly.

Until recently, the main sources of opinionated information were friends and specialized websites. Now, consumers can consult opinions published by others before buying a specific product. However, mining opinion and sentiment from social media is very challenging due to the vast amount of data generated by the different sources. Much opinionated information about a topic is hidden within

the data and therefore it is nearly impossible for a person to look through the different sources and extract useful information. In this thesis we develop tools that can help users understand and track the public opinion expressed towards a topic by other users, through the use of the dynamic nature of social media.

The huge amount of data posted in the different social media platforms makes tracking public opinion a challenging problem. Although opinion mining models are useful in finding documents that are relevant and opinionated about a certain topic, the retrieved set can still be very large and reading and analysing all documents can be difficult for users. This makes it almost impossible for users to understand the strength of the opinion expressed, how many people support that opinion, the impact of that opinion and when and how caused such opinion to be expressed about the topic.

Another challenge in developing tools that can track opinion is the text that is used in social media. The social media documents are usually short (e.g., tweets, Facebook posts). The short length and the informal type of the platforms have caused the emergence of textual informalities (e.g., emoticons, emphatic lengthening) that are extensively encountered in Twitter but also in other media. Thus, the methods proposed for tracking opinion in social media should take into account these unique characteristics.

One important aspect of tracking public opinion in social media is to estimate the number of people that support a specific opinion or that will react to a specific post. In general, some posts trigger massive reactions whereas others do not. To estimate the number of reactions triggered by a post is not trivial since there are a lot of factors involved such as the structure of the network. Different information signals should be considered that have to do not only with the content of the post but also with the early reactions of the users regarding the post (e.g., number of comments regarding the post). An approach that combines the different information signals is a promising direction that needs to be explored to effectively estimate the number of people that will react to a specific post.

The impact of the posts that are published online is another important aspect of tracking public opinion in social media. Reputation analysts are highly interested in understanding if the posts that are published in social media can have a positive or negative impact on the entity and more specifically on its reputation. This task that is known as *reputation polarity analysis* is challenging since there are posts that describe facts (i.e., do not express sentiment) and which have an impact on the entity. Propagating sentiment signals from tweets that express sentiment to those that do not is a direction that needs to be investigated for the reputation polarity analysis problem.

In summary, the dynamic nature of social media with the constantly changing

vocabulary, makes developing tools that can automatically track public opinion a challenge. To help users better understand public opinion, it is important to analyse the set of documents to: a) find the relevant documents and the sentiment polarity expressed in them; b) identify the important time intervals where there is a change in the opinion; c) identify the causes for the opinion change; d) estimate the number of people that have a certain opinion about the entity; and e) measure the impact of public opinion towards the entity.

In the rest of this chapter we first present the research questions guiding this thesis. Next, we present the main contributions of this thesis. Finally we present an overview of the thesis and the publications resulted from the research.

1.2 Research Questions

This thesis addresses the following question: *How can we track the public opinion expressed on social media towards a topic?* Before we can proceed with tracking public opinion we need to understand how can opinion be effectively extracted from social media and microblogs (**RQ1**). We then develop a methodology to effectively track public opinion over time, forecast opinion in the future, extract sentiment spikes and understand why these spikes occurred (**RQ2**). Understanding public opinion also requires predicting how many people support a certain opinion. Therefore, we propose an approach to predict the number of triggered emotional reactions (**RQ3**). Finally, it is important to estimate the impact of the public opinion and therefore we propose sentiment signals to explore the impact of public opinion on an entity's reputation (**RQ4**).

In more detail, as a prerequisite to tracking opinion, we need to develop a methodology that can retrieve relevant and opinionated documents from a popular social media platform. In Chapter 3 we address the following question:

RQ1 How can we find documents that are opinionated and express opinion about a topic in a microblogging collection? Can we make use of the textual peculiarities that are present in posts such as tweets to improve Twitter opinion retrieval?

This general research question leads to the following detailed questions:

RQ1.1 How many topics are discussed in a single tweet?

RQ1.2 What is the most effective combination of stylistic variations regarding topic-specific Twitter opinion retrieval?

RQ1.3 Is the importance of stylistic variations in indicating opinion topic dependent?

In Chapter 4 we develop a tool that can track the opinion over time, forecast opinion, extract the opinion spikes and explain the reasons that likely caused the spikes. We answer the following research question:

RQ2 How can we model opinion evolution and identify the important causes of opinion change?

This general research question leads to the following detailed questions:

RQ2.1 Can conventional time series methods be applied to track sentiment evolution over time and forecast sentiment in the future?

RQ2.2 Can outlier detection be applied to identify sentiment spikes?

RQ2.3 How does an approach based on a combination of topic model with KL-divergence perform in extracting the likely reasons that caused a sentiment spike?

In Chapter 5 we develop a tool that can estimate how many people support a specific opinion. We answer the following research question:

RQ3 How can we predict how many people will react with a specific emotion when a news post is published?

This general research question leads to the following detailed questions:

RQ3.1 Can we improve the effectiveness of baseline classifiers by adding additional pre-publication information based on news post content?

RQ3.2 Can we improve the effectiveness of baseline classifiers by adding additional post-publication information extracted from users' comments?

RQ3.3 How does a model that combines textual and early commenting features perform?

RQ3.4 What is the added value of the commenting features in terms of effectiveness in the task of emotional reactions prediction?

In Chapter 6 we focus on estimating the impact of the public opinion on the reputation of an entity. We address the following research question:

RQ4 How can we estimate the impact of posts on an entity? Can we use sentiment signals propagation to estimate the impact of posts on an entity's reputation?

The last general research question leads to the following detailed questions:

RQ4.1 Can we use training material to detect terms with reputation polarity and use them to augment a general sentiment lexicon?

RQ4.2 What is the right level of generalization for a reputation lexicon?

RQ4.3 Can we propagate sentiment to text that is similar in terms of content to improve reputation polarity?

RQ4.4 What is the best way to select the set of pairwise similar tweets that can be used to learn the sentiment that will be propagated?

1.3 Main Contributions

The main contributions of this thesis are the following:

- Some novel algorithms to retrieve tweets that are relevant and opinionated with respect to a topic.
- A test collection of labelled tweets that can be used for evaluating methods for ranking the likely causes of sentiment spikes.
- A new method that can be used for extracting and ranking the causes of a sentiment spike.
- Some novel algorithms to measure the emotional reactions of users by news posts.
- Some novel algorithms to measure the impact of tweets on the reputation polarity of an entity.

1.4 Thesis Overview

This thesis is organised in 7 chapters.

Chapter 2 introduces related work for social media and information retrieval. Also, the chapter presents related work on both opinion mining in social media

and current approaches to estimating sentiment dynamics and online reputation analysis.

Chapter 3 focuses on Twitter opinion retrieval. First, the chapter explores the topical distribution of tweets to determine the number of topics discussed in a single tweet. Next, it proposes a Twitter opinion retrieval model which uses information about the topics of tweets to retrieve those that are relevant and contain opinion about a user's query. The proposed model calculates opinionatedness by combining information from the tweet's terms and the topic-specific stylistic variations that are extensively used in Twitter. We compare several combinations of stylistic variations, including emoticons, emphatic lengthening, exclamation marks and opinionated hashtags.

Chapter 4 focuses on tracking opinion over time. First, we explore time series approaches to investigate if they can be used for opinion tracking. Next, we plot signals that show a topic's popularity and sentiment evolution towards the topic under examination. We explore the effectiveness of state-of-the-art time series tools in predicting sentiment in future. In addition, the chapter proposes a new method that can be used for extracting and ranking the causes behind a sentiment spike. The approach combines LDA topic model with Relative Entropy. The former allows to extract the topics discussed in the time window before the sentiment spike and the latter to detect the topics which probably caused the sudden change. We finally rank these topics according to their contribution to the sentiment spike. In addition, the chapter presents a labelled collection of tweets that can be used for extracting and ranking the likely causes of a sentiment spike.

Chapter 5 presents a methodology for predicting the emotional reactions that are triggered on users by news posts. We propose features that are extracted from news posts' content and users' comments about the post to predict the number of triggered emotional reactions. In addition, we combine the features extracted from comments published shortly after the post (within the first 10, 20 or 30 minutes after the publication of the news post) with the terms of the news post to explore if this combination can effectively address the problem of emotional reactions prediction.

Chapter 6 focuses on estimating the impact of online posts on the reputation of an entity. More specifically, we propose sentiment signals propagation to estimate reputation polarity of tweets. We consider two ways of propagating sentiment signals: (i) augmenting the sentiment lexicons with terms that indicate reputation polarity even if they do not convey sentiment polarity; and (ii) direct propagation to texts with similar content. We hypothesize that tweets that are about a specific topic tend to have the same reputation polarity. In this way, if there are many tweets about a specific topic, then some of those tweets will

explicitly express some sentiment towards the topic and can be used to annotate the polar facts. We explore different approaches for estimating the similarity and propagating sentiment. Finally, we propose a polar fact filter that can differentiate between polar facts and reputation-neutral tweets.

Chapter 7 concludes the thesis. We revisit the research questions introduced earlier and answer them. We look forward and formulate open questions in automatic opinion mining and emotional reactions.

1.5 Publication Overview

The material of this thesis was published in conferences and journals listed below:

- Chapter 2 is partially based on:
  - A. Giachanou and F. Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):28, 2016

- Chapter 3 is based on:
  - A. Giachanou and F. Crestani. Opinion retrieval in twitter: Is proximity effective? In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, SAC '16, pages 1146–1151, 2016
  - A. Giachanou, M. Harvey, and F. Crestani. Topic-specific stylistic variations for opinion retrieval on twitter. In *Proceedings of the 38th European Conference on Information Retrieval Research*, ECIR '16, pages 466–478, 2016

- Chapter 4 is based on:
  - A. Giachanou and F. Crestani. Tracking sentiment by time series analysis. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1037–1040, 2016
  - A. Giachanou, I. Mele, and F. Crestani. Explaining sentiment spikes in twitter. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2263–2268, 2016
  - A. Giachanou, I. Mele, and F. Crestani. A collection for detecting triggers of sentiment spikes. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1249–1252, 2017

- Chapter 5 is based on:
  - A. Giachanou, P. Rosso, I. Mele, and F. Crestani. Emotional influence pre-

diction of news posts. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, ICWSM '18, pages 592–596, 2018
- A. Giachanou, P. Rosso, I. Mele, and F. Crestani. Emotional reactions prediction of news posts. In *Proceedings of the 9th Italian Information Retrieval Workshop*, IIR '18, 2018
- A. Giachanou, P. Rosso, I. Mele, and F. Crestani. Early commenting features for emotional reactions prediction. In *Proceedings of the 25th International Symposium on String Processing and Information Retrieval*, SPIRE '18, 2018

- Chapter 6 is based on:
- A. Giachanou, J. Gonzalo, I. Mele, and F. Crestani. Sentiment propagation for predicting reputation polarity. In *Proceedings of the 39th European Conference on Information Retrieval Research*, ECIR '17, pages 226–238, 2017

1.6 Additional Publications

Additional papers were published during this thesis. These publications originated either from other collaborations and projects or from participation in evaluation campaigns such as Text REtrieval Conference[1] (TREC).

- A. Giachanou, I. Markov, and F. Crestani. Opinions in federated search: University of Lugano at TREC 2014 federated web search track. In *Proceedings of the 23rd Text REtrieval Conference*, TREC 2014, 2014

- M. Aliannejadi, S. A. Bahrainian, A. Giachanou, and F. Crestani. University of Lugano at TREC 2015: Contextual suggestion and temporal summarization tracks. In *Proceedings of the 24th Text REtrieval Conference*, TREC 2015, 2015

- K. D. Varathan, A. Giachanou, and F. Crestani. Temporal analysis of comparative opinion mining. In *Proceedings of the 18th International Conference on Asian Digital Libraries*, ICADL '16, pages 311–322, 2016

- A. Giachanou and F. Crestani. Opinion retrieval in twitter using stylistic variations. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, SAC '16, pages 1077–1079, 2016

- A. Giachanou, F. Rangel, F. Crestani, and P. Rosso. Emerging sentiment language model for emotion detection. In *Proceedings of the 4th Italian Conference on Computational Linguistics*, CLiC-it '17, 2017

[1] https://trec.nist.gov/

- K. D. Varathan, A. Giachanou, and F. Crestani. Comparative opinion mining: A review. *Journal of the Association for Information Science and Technology*, 68(4):811–829, 2017

- A. Giachanou, I. Mele, and F. Crestani. USI participation at SMERP 2017 text retrieval task. In *Proceedings of the 1st Exploitation of Social Media for Emergency Relief and Preparedness Workshop (Data Challenge Track)*, SMERP@ECIR '17, 2017

- A. Giachanou, I. Mele, and F. Crestani. USI participation at SMERP 2017 text summarization task. In *Proceedings of the 1st Exploitation of Social Media for Emergency Relief and Preparedness Workshop (Data Challenge Track)*, SMERP@ECIR '17, 2017

- A. Lomi, A. Giachanou, F. Crestani, and S. Angelopoulos. Table for two: Explaining variations in the evaluation of authenticity by restaurant critics. In *Proceedings of the 12th Organization Studies Workshop*, 2018

- N. Fuhr, A. Giachanou, G. Grefenstette, I. Gurevych, A. Hanselowski, K. Jarvelin, R. Jones, Y. Liu, J. Mothe, W. Nejdl, et al. An information nutritional label for online documents. *ACM SIGIR Forum*, 51(3):46–66, 2018

Chapter 2

Related Work

In this chapter we introduce the background and the underlying concepts related to the thesis. First, we introduce social media and their applications in Section 2.1. Then, we present an overview of Information Retrieval and the recent developments in this field in Section 2.2. Next we move to related work on sentiment and opinion analysis and present the background regarding this field in Section 2.3. Prior work on sentiment dynamics is presented in Section 2.4. Finally, Section 2.5 presents the related work on reputation analysis.

2.1 Social Media

Recent years have witnessed the rapid growth of social media platforms that has transformed the interaction and communication of individuals throughout the world. Social media platforms gave the capability to users to publish content and interact with each other in a very easy way. According to the Oxford dictionary[1], *social media* is defined as:

> Websites and applications that enable users to create and share content or to participate in social networking

Social media is different to social networking although a lot of people use the two terms interchangeably. The key difference is that social media is a media that is primarily used to transmit or share information to a broad audience, whereas social networking is an act of engagement among people with common interests who build relationships through community [1].

[1] https://en.oxforddictionaries.com/definition/social_media

The first social media were created in the 1990s. Six Degrees is considered the first social media website since it combined popular features such as profiles and friends lists [42]. In the following years, blogging services such as Blogger and Epinions gained a lot of popularity. Blogging sites are mainly used from people to post their experiences and opinions on any topic (e.g., travelling experiences, sports, reviews on products). Since 2000s many social media platforms have emerged allowing different types of interactions among users. For example, some sites are completely based on sharing multimedia content such as photos and videos (e.g., Instagram, Flickr), whereas other focus on connections among professionals (e.g., LinkedIn). Another very popular type of social media is the online games and virtual worlds such as SecondLife on which users create virtual communities and interact with each other.

Two of the most popular social media platforms are Twitter[2] and Facebook[3]. Twitter is a microblogging service that allows users to exchange short messages, images, or video links. It has about 320 million active users who can post short messages (i.e., tweets) that are limited to 140 character[4]. Twitter allows unidirectional connections where users can follow other users without being befriended. Facebook is also very popular and has about 1.65 billion monthly active users. The users can create a personal profile, add other users as friends, exchange messages, and share photos and comments. In addition, users can join or follow several pages of their interest (e.g., news pages).

Social media platforms are extensively used in many countries around the world. According to Statista [2], approximately 2 billion people used social media platforms and apps in 2015. A number of different research papers emerged with the aim to answer questions about the use, influence and impact of social media on users. For example, there have been several studies that tried to understand the motivation of social media users and the context they are used [161]. Hughes et al. [77] examined how the personality traits (neuroticism, extraversion, openness-to-experience, agreeableness and conscientiousness) correlate to the social and informational use of Twitter and Facebook and found a correlation between sociability and the use of Twitter and Facebook. Fu et al. [45] focused on incentives of content-sharing and found that self-interest incentives (i.e., achievement, self-expression, and loneliness) could lead users to share commercial messages and opinions, whereas communal incentives (i.e., connection, altruism, and group joy) could lead users to share lifestyles affairs and opinions.

In health domain, Signorini et al. [140] showed that tweets can be used to

[2] `https://twitter.com/`

[3] `https://www.facebook.com/`

[4] this limitation has been recently removed.

track users' interest on H1N1 influenza and estimated disease activity in real time, whereas Kapp et al. [83] reported on Facebook advertising as a tool of recruiting patients for online or clinical studies. More recently, some studies started focusing on vaccination in an attempt to understand the current attitudes and beliefs towards vaccines [76, 81].

Information extracted from social media is also very important for marketing and business. Jansen et al. [78] found that Twitter is an online tool for customer word of mouth communications (WOM) and that 19% of tweets mention one brand. De Vries et al. [39] investigated the factors that lead brand post popularity and showed that it is positively correlated to the number of positive and negative comments. Many academics over the last years tried to predict trends in financial markets using data form social media. Bollen et al. [25] performed a sentiment analysis of all public tweets posted from the 1st of August to the 20th of December, 2008 and found a correlation between the mood level (happy, calm, anxiety) in posts and the value of the Dow Jones Industrial Index. More recently, Pagolu et al. [116] showed that there is a strong correlation between rise/fall in stock prices of a company to the public opinions about that company.

Apart from their numerous benefits and applications, social media have been criticised a lot about their negative effects. Health professionals and researchers reported that the excessive use of digital technology, like social media, by adolescents can cause disruptions in their physical and mental health, sleeping behaviour and academic performance [128]. Zagorski [169] showed that the use of multiple social media platforms is more associated with depression and anxiety among young adults than time spent online. In addition, social media have been criticised for spreading fake news, a problem that affects different domains such as politics, finance, and health. For example, anti-vaccine campaigns, mainly propagated via social media, led to a decrease of *Measles, Mumps, & Rubella* (MMR) vaccination rates causing in 2017 one of the worst measles outbreak in decades, for a disease that was almost eradicated [137].

2.2 Information Retrieval

The field of Information Retrieval (IR) is *concerned with the structure, analysis, organization, storage, searching, and retrieval of information* [135]. The most common application of IR is web search where someone types a query to a search engine and receives in response a ranked list of documents relevant to the query. Thus, one of the most important concepts in IR is relevance. A *relevant document* is a document that contains information that a person is looking for, when

submitting the query to the search engine [37].

A retrieval model can be defined as a formal representation of the process of matching a query and a document. Retrieval models can be classified into a number of classes, such as for example Boolean [146], Vector Space [135], Probabilistic [155] and Language [125] models. In the boolean retrieval systems, documents that contain the exact query terms are retrieved. Vector space models introduced by Salton [135] represent a document d and a query q as a vector where each term represents a dimension. Documents are then ranked according to their distance to the query. Term frequency (*tf*) of a term can be used as a value of each dimension. The inverse document frequency (*idf*) was introduced to score the terms according to their importance regarding the collection. The combination $tf \cdot idf$ is the most commonly used term statistic for vector space models. Then the similarity between the document and the query can be measured by distance metrics such as the cosine similarity.

According to language models, documents are ranked according to their likelihood given a query. The likelihood of a document d given a query q can be calculated as follows:

$$P(d|q) \propto P(q|d)P(d) = P(d)\prod_{t \in q} P(t|d)$$

where $P(q|d)$ is the likelihood that the document generated the query and $P(d)$ is the (query independent) prior probability of retrieving that document. A common assumption is that the document is generated according to a multinomial distribution, and that each term in the query is chosen independently of each other, leading to the product of term probabilities $P(t|d)$ which can be calculated using a variety of estimation techniques. Maximum Likelihood estimates (i.e., basic relative frequency) do not work well because of the finite length of most documents. This problem is alleviated by the application of various forms of smoothing (such as Jelinek-Mercer or Dirichlet smoothing) where some probability mass is given to unseen words for each document, according to the frequency of the words across all documents in the corpus. The term distribution for a single document d can be smoothed using the general frequency for terms in all documents $P(t|\bigcup_i d_i)$ or the frequency of the term in documents that are similar in one aspect (e.g. documents with the same author or the same time stamp). For example,

$$P(t|d) \approx \lambda_1 P_{ML}(t|d) + \lambda_2 P_{ML}(t|\bigcup\nolimits_i d_i)$$

is a Jelinek-Mercer smoothed estimate calculated by weighted summation of Maximum Likelihood estimates.

Topic Modeling is an evolution of Language Models that assumes a more complicated statistical process for generating documents. In topic modeling approaches the documents are represented as distributions over a latent topic space $P(z|d)$ which has much lower dimensionality than the original term space. In other words, the documents are represented on a low dimension topic frequency vector (usually in the order of hundreds of topics) rather than a very high dimensional term frequency vector (usually with tens of thousands of words).

Latent Dirichlet Allocation (LDA), one of the most well known topic models, is a generative document model which uses a "bag of words" approach and treats each document as a vector of word counts [23]. Each document is a mixture of topics and is represented by a multinomial distribution over those topics. Each document d is associated with a multinomial distribution over K topics, denoted θ. Each topic z is associated with a multinomial distribution over words, denoted ϕ. Both θ and ϕ have Dirichlet prior with hyperparameters α and β respectively. For each word in a document d, a topic z is sampled from the multinomial distribution θ associated with the document and a word w from the multinomial distribution ϕ associated with topic z. This generative process is repeated N_d times, where N_d is the total number of words in the document d. LDA defines the following process for each document in the collection:

1. Choose $\theta_d \sim \text{Dir}(\alpha)$,
2. Choose $\phi_z \sim \text{Dir}(\beta)$,
3. For each of the N words w_n:
   (a) Pick a topic z_n from the multinomial distribution θ_d
   (b) Pick a word w_n from the multinomial distribution ϕ_z

In recent years, there has been an increase of interest in developing topic models for topic evolution over time. The Dynamic Topic Model (DTM) uses state space models on the natural parameters of the multinomial distributions that represent the topics [24]. The continuous time dynamic topic model (cDTM) replaces the discrete state space model of the DTM with its continuous generalisation, Brownian motion [162]. The DTM and cDTM employ a Markov assumption over time that the distribution at current epoch only depend on the previous epoch distribution. Topic Over Time model (TOT) explicitly models time jointly with word co-occurrence patterns [163]. The TOT model does not discretise time and does not make Markov assumptions over state transitions in time. Rather, TOT parameterises a continuous distribution over time associated with each topic, and topics are responsible for generating both observed timestamps as well as words. Parameter estimation is thus driven to discover topics

that simultaneously capture word co-occurrences and locality of those patterns in time. Bahrainian et al. [17] presented the discrete Dynamic Topic Modeling (dDTM) approach that, similarly to DTM, is based on LDA. Different to DTM, dDTM relaxes the assumption that each topic should appear on all the time slices. A Hidden Markov Model is used to infer topic chains over different time slices. This approach has been effectively applied on a range of collections such as news and tweets [17, 102].

2.3 Opinion in Social Media

This section presents prior work on opinion mining in Section 2.3.1 and briefly describes approaches for opinion retrieval in Section 2.3.2.

2.3.1 Opinion Mining

Opinion Mining (OM) and *Sentiment Analysis* (SA) are two of the emerging fields that aim to help users find opinionated information and detect the sentiment polarity. OM and SA are commonly used interchangeably with roughly the same meaning. However, some researchers state that they tackle two slightly different problems. According to Tsytsarau and Palpanas [153], OM is about determining whether a piece of text contains opinion, a problem that is also known as *subjectivity analysis*, whereas the focus of SA is the detection of the sentiment polarity by which the opinion of the examined text is assigned a positive or negative sentiment. More formally, OM or SA is the *computational study of opinions, feelings and subjectivity in text* [120].

According to a more complete definition given by Liu [93] "an opinion is a quintuple $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k". To illustrate the different parts of the definition we use an example. Consider the following review posted on *10.06.2015* by the user *Helen*:

The picture quality of my new Nikon V3 camera is great.

In this example, *Nikon V3* is the entity for which the opinion is expressed, *picture quality* is an aspect of the entity, the sentiment of the opinion about this particular aspect is *positive*, the opinion holder is the user *Helen* and the time that the opinion is expressed is *10.06.2015*. The opinion quintuple *(Nikon_V3, picture_quality, positive, Helen, 10.06.2015)* can be generated after analysing this example.

OM and SA have been studied on many media including reviews, forum discussions, and blogs. The sentiment analysis methods can be roughly divided into lexicon and classification based approaches. Lexicon based approaches [40, 147, 154] rely on opinion and sentiment lexicons and do not require any training. One of the most well-known lexicon based algorithms developed for social media is SentiStrength [150]. SentiStrength can effectively identify the sentiment strength of informal text including tweets using a human-coded lexicon that contains words and phrases frequently found in social media. Besides the sentiment lexicon, SentiStrength uses a list of emoticons, negations and boosting words to assign the sentiment to a text. The algorithm was extended in [151] by introducing idiom lists, new sentiment words in the lexicon and by strength boosting using emphatic lengthening. SentiStrength was compared with many machine-learning approaches and tested on six different datasets including a dataset with tweets.

The classification based approaches employ classifiers that are trained on several features to do the prediction [132, 87]. One of the most well known studies is by Go et al. [64] who treated the problem as a binary classification problem, classifying the tweets as either positive or negative. They compared Naïve Bayes, Maximum Entropy and Support Vector Machine (SVM), among which SVM with unigrams achieved the best result. Later Pak and Paroubek [117] used emoticons to label the training data from which they built a multinomial Naïve Bayes classifier using N-gram and Part-Of-Speech tags as features. Other studies apart from emoticons also considered hashtags as class labels [38, 88] and showed that combining them improves the performance of the classifier. Also, a number of researchers employed features such as emoticons, abbreviations and emphatic lengthening to study their impact on sentiment analysis. Emphatic lengthening (e.g., cooooool) was studied by Brody and Diakopoulos [28] who showed that it is strongly associated with sentiment.

Feature selection that is the process of selecting a subset of the most useful features is not a trivial task. To this end, Agarwal et al. [3] and Kouloumpis et al. [88] analysed the usefulness of different features for Twitter sentiment analysis. Agarwal et al. [3] proposed a feature-based model and performed a comprehensive set of experiments to examine the usefulness of various features including POS and lexicon features. The analysis showed that the most useful combination is the one of POS with the polarity of words. Kouloumpis et al. [88] also analysed the impact of different features on sentiment analysis. This study was mostly focused on semantic and stylistic features including emoticons, abbreviations and the presence of intensifiers. Combination of features that reveal the polarity of the terms with n-grams managed to achieve the best performance.

However, this study showed that POS had a negative impact on sentiment analysis, in contrast to the conclusions of Agarwal et al. [3].

Deep learning is a new field of machine learning concerned with algorithms that are based on learning data representations [66]. Recently, researchers started exploring deep learning approaches for sentiment analysis. Deep learning can be used to learn word embeddings from large amounts of text data [96]. Tang et al. [149] proposed to learn sentiment specific word embeddings (SSWE) from tweets that were collected using distant supervision. They developed three neural networks to learn SSWE that were then used as features. The best result was obtained by combining SSWE with sentiment lexicons and the same features used by Mohammad et al. [105]. Dong et al. [41] proposed an Adaptive Recursive Neural Network (AdaRNN) for entity-level sentiment analysis. This method used a dependency tree in order to find the words syntactically related with the target and to propagate the sentiment from sentiment words to the targets. AdaRNN was evaluated on a manually annotated dataset consisting of 6,248 training and 692 testing tweets. More recently, Goel et al. [65] managed to obtain the best performance in the emotion intensity task [107], that focused on predicting the intensity of emotions in tweets (a score that ranges from 0 to 1), using an ensemble of three neural-network approaches.

A related field to sentiment analysis is *Tweet sentiment quantification* that estimates the distribution of tweets across different classes. However, the two tasks are not the same. All the differences between classification and quantification are discussed by Gao and Sebastiani [47]. Gao and Sebastiani aimed to differentiate between sentiment classification and sentiment quantification and they argued that the latter is more appropriate when the goal is to estimate the class prevalence. In their study, they performed a series of experiments on various tweet collections and showed that quantification-specific algorithms outperform, at prevalence estimation, state-of-the-art classification algorithms. Amati et al. [11] modified Hopkins and King's approach to estimate the sentiment distribution towards different topics. They proposed to use features that are learned during the training phase. These features composed the sentiment dictionary. Their experiments showed that their proposed approach can be effectively applied for real time sentiment estimation.

A large amount of work has also been done on emotion analysis. The difference between sentiment and emotion is that sentiment reflects a feeling whereas emotion reflects an attitude [153]. According to Plutchik [124] there are eight basic emotions: anger, joy, sadness, fear, trust, surprise, disgust and anticipation. Emotion detection aims at identifying various emotions from text. Mohammad [106] proposed to consider hashtags that show an emotion (e.g., #anger, #sur-

prise) for emotion detection. After creating a corpus that could be used for emotion detection, Mohammad [106] conducted experiments which showed that the self-labeled hashtag annotations were consistent and matched with the annotations of the trained judges. Also, he created an emotion lexicon that could be used as available source of information when detecting emotions in text. Roberts et al. [132] used the six Ekman's basic emotions (joy, anger, fear, sadness, surprise, disgust) proposed in [43]. Roberts et al. [132] extended the original list with an additional emotion: love. In their study, they created a series of binary SVM emotion classifiers to address the problem of emotion analysis.

2.3.2 Opinion Retrieval

Opinion Retrieval deals with the retrieval of documents that are relevant to a query but also contain opinions about it [119, 97]. A typical approach to opinion retrieval is based on a three-step process. In the first step, a standard IR model is applied to rank the documents in relation to their relevance to the query. The retrieval models can be any of the models discussed in Section 2.2. Next, the top retrieved documents are analysed based on their opinionatedness on the topic of the query. The opinionatedness score of the documents can be calculated with any of the techniques discussed in Section 2.3.1. In the last step, the relevance and opinionatedness scores are combined to generate the final ranking of the documents. A common approach is the linear combination of the scores on which a parameter determines the weight of each score [75].

A typical approach is described by Yang et al. [167] who presented a method based on multiple modules. Yang et al. used a vector space model to rank documents based on their relevance to the query which are then analysed in reference to their affective content using different sources of evidence. A different approach was followed by Amati et al. [10] who proposed an information-theoretic approach to automatically select the most effective terms from an opinionated lexicon. Candidate opinion terms are selected according to a method that is similar to the Weighted Log-Likelihood Ratio. Terms more uniformly distributed among opinionated documents are preferred, as they are more likely to convey an opinion regardless of a particular topic. Finally, opinion terms are submitted to a retrieval system as a query to get scores for the documents.

Other researchers considered proximity in their methodologies. Zhang and Ye [170] calculated the proximity of opinion terms to query terms by computing the probability that a query term co-occurs with an opinion term within a short window of text. Proximity is also used by Gerani et al. [48] who considered proximity-based opinion density functions to capture the proximity information

between the opinionated terms and the query terms. They used different kernel functions to measure the opinionatedness at the position of a query term in the document.

The majority of prior work on opinion retrieval was focused on blogs. Research on blogs opinion retrieval was facilitated by TREC Blog Track that provided an evaluation framework on opinion finding for the researchers [113, 97, 114]. Ranking opinionated tweets based on their relevance and opinionatedness towards a topic was first considered by Luo et al. [95]. In their work, they proposed a learning to rank algorithm to rank tweets based on their relevance and opinionatedness to a user's query. To address the problem, they explore the effectiveness of social aspects of the author (e.g., number of friends), information derived from the body of tweets and opinionatedness. Experimental results showed that social features can improve retrieval performance.

2.4 Sentiment Dynamics

Public opinion changes over time, therefore tracking opinion and sentiment evolution is very critical for the interested parties. One interesting study was presented by Bollen et al. [25] who performed sentiment analysis on all public tweets posted from the 1st of August, 2008 to the 20th of December, 2008. Bollen et al. used a psychometric instrument called Profile of Mood States (POMS) to extract and analyse different moods (tension, depression, anger, vigor, fatigue, confusion) detected in tweets. They found that the mood level in Twitter posts was correlated with cultural, political and other world global events. An et al. [14] combined classical sentiment analysis algorithms, data mining techniques and time series methods with the aim to detect and track sentiment regarding climate change from Twitter feeds.

Another line of research proposes to use models that combine topic detection and sentiment analysis. One example is the work proposed by Mei et al. [101] who modeled how positive/negative opinions about a given subtopic changes over time. They first used a topic-sentiment mixture model to estimate background, content, positive and negative topics based on the occurrence of words in the collection's documents. Then, they used Hidden Markov Model [22] to assign topic and sentiment polarity to each word. Finally, the topic life cycles and sentiment dynamics were extracted by counting the words with corresponding labels over time. He et al. [70] introduced the dynamic Joint Sentiment-Topic model (dJST) to capture and track shifts in topics and sentiments. The dJST model assumed that the generation of documents at each timestamp was influ-

enced by documents at previous timestamps.

Jiang et al. [80] proposed the Topic Sentiment Change Analysis (TSCA) model in which they considered as the start of a time interval the point when the number of documents started increasing. The intuition behind this model is that the change of a topic's sentiment is usually accompanied by hot discussions related to the topic. In TSCA, a sudden change in the number of documents was considered as an indicator of a possible sentiment change and therefore, the associated time stamp was regarded as a good candidate for being the start of a time interval.

The analysis of sentiment evolution gives the opportunity to identify sudden changes of sentiment and, more importantly, to get insights on what has caused sentiment spikes. *Sentiment spikes* occur when a large amount of documents of a specific sentiment is posted. Detecting sentiment spikes allows to take quick reactions in response to user sentiments, whereas understanding the reasons that likely caused a sentiment spike provides valuable information for governments and companies to be proactive and improve their tactics. For example, suppose that the negative sentiment towards a product increases, then the respective marketing department can extract the causes for such increase in negative opinion and act promptly to avoid a further increase in the negative sentiment.

Understanding the reasons that likely caused a sentiment spike is still underexplored. One work in this direction is the one by Balog et al. [19] who tried to identify causes of spikes in users' mood. Balog et al. used LiveJournal posts which mention the user mood and they extracted unusually common words in order to find the causes of the identified mood changes. These words were then used for searching the related events in a news dataset. Their method managed to identify major news as the causes of mood changes. However, they did not conduct a full evaluation to measure the performance of their approach.

Montero et al. [110] focused on identifying the likely causes of emotion spikes of influential users. They used empirical heuristics to identify the emotional spikes and keyphrases to extract the causes of the identified spikes. Their evaluation on emotion flow visualization, emotion spikes identification and likely cause extraction showed that their methodology is effective. Tan et al. [148] tried to analyse sentiment variations and to extract possible causes of such variations. They proposed two models: (1) the Foreground and Background LDA (FB-LDA) model to extract the foreground topics and (2) the Reason Candidate and Background LDA (RCB-LDA) model to rank the extracted foreground topics according to their popularity within the sentiment variation period. However, they did not perform any evaluation regarding the reasons that caused the spikes.

2.5 Reputation Polarity Analysis

Recently, with the rise of user-generated content online, social media analysis for reputation management analysis of companies is gaining importance [91]. Online Reputation Analysis aims at the development of computational tools that allow filtering of relevant documents mined from UGC and estimation of the reputation impact towards an entity [98].

One of the core tasks of Online Reputation Management is *reputation polarity* that refers to analysing the impact of the posts published on the entity of interest. Determining the impact of a post on the reputation of an entity is a challenging task. The task of reputation polarity analysis has several similarities with the sentiment analysis task. Therefore, prior work on reputation polarity analysis has evolved from sentiment analysis. However, the two tasks are not the same. A key difference refers to the posts that do not explicitly express a sentiment but have an impact on the entity's reputation. For example, the tweet *BS becomes first UK bank to back Visa V.me wallet* which has a positive impact on *RBS* does not express any sentiment.

Using pure sentiment analysis as a subsitute of reputation polarity has a lot of limitations. Therefore, there were efforts towards evaluation campaigns that can facilitate research on this field [12, 13]. The first well-known evaluation campaign for Online Reputation Management Systems is the Replab 2012 [12]. One of the tasks was polarity classification with regards to the reputation of the company. Many of the participants tried to detect reputation polarity using sentiment as their starting point. Kaptein used SentiStrength [150] to extract sentiment features [84], whereas Yang et al. [166] added a happiness feature to detect the reputation polarity. Albornoz et al. [7] extracted the WordNet concepts from the tweets and then an emotion from an affective lexicon was assigned to them. In addition to the lexicon-based features, Chenlo et al. [32] bootstrap more data from the background to learn hashtags for positive and negative polarity, whereas Peetz et al. [121] focused on how tweets were perceived by analysing the sentiment in retweets and replies.

Greenwood et al. [67] detected reputation polarity using a text classification approach with tokens as one of the features. Balahur and Tanev [18] used lexicon resources in addition to other features including emoticons, negation and intensifiers. Jeong and Lee [79] calculated correlation coefficients for each term and polarity. Karlgren et al. [86] considered a *customer satisfaction* semantic pole consisting of manually selected terms in addition to a semi-automatic enlargement via a semantic model. Villena-Román et al. [160] used a sentiment analysis tool in addition to linguistic features.

Several teams participated in the following year campaign, the 2013 RepLab evaluation [13] campaign. Several participants used bag-of-words approaches and combined them with information from lexicon resources [133, 111, 44] or clustering to detect the reputation polarity [69]. Castellanos et al. [29] used KL-divergence to extract the most discriminating terms for the polarity classification whereas, Cossu et al. [36] based their approach on TF-IDF measure. Distributional term representations were used by Villatoro-Tello et al. [159] who used term co-occurrence statistics to represent also the contextual information. Others have applied commercial systems based on distributional semantics represented in a semantic space [85]. Finally, Spina et al. [142] based their approach on domain specific semantic graphs and automatically expanded a general purpose lexicon.

2.6 Summary

In this chapter we presented the main research areas that are related to this thesis. We started with an introduction to social media and their different applications. Then we briefly explained retrieval models and introduced the opinion retrieval task. Next, we presented previous work on sentiment and opinion analysis that was followed by a review of previous studies on sentiment dynamics. Finally, we gave an overview of the previous work in the area of reputation polarity detection.

Chapter 3

Topic Specific Opinion Retrieval in Twitter

Tracking public opinion requires retrieving the documents that are relevant and opinionated towards the topic of interest. Therefore, this chapter focuses on Twitter opinion retrieval and aims to identify tweets that are both relevant and express opinion about a query. First, we use the topic model LDA to explore the topical distribution of tweets and we estimate the number of topics discussed in a single tweet. Next, we propose a model which uses information about the topics of tweets to retrieve those that are relevant and opinionated about a query. The proposed model calculates opinionatedness by combining information from the terms and the topic-specific stylistic variations. The stylistic variations include emoticons, emphatic lengthening, exclamation marks and opinionated hashtags. Experimental results show that stylistic variations are topic-specific and that incorporating them in the ranking function significantly improves the performance of Twitter opinion retrieval.

3.1 Introduction

The first step for tracking public opinion towards an entity or a topic is to find the documents that are relevant to that entity and also express an opinion about it. Hence, in this chapter, we focus on *opinion retrieval* and more specifically on *Twitter opinion retrieval* that aims to identify tweets that are both relevant to a topic and express opinion about it. Twitter opinion retrieval is different to standard opinion retrieval that focuses on blogs or review sites since it deals with a different type of text (i.e., tweets) with a lot of peculiarities. Addressing Twitter opinion retrieval is important because it can be used as a tool to understand

public opinion about a specific topic, which is helpful for a variety of applications. One typical example refers to enterprises that can capture the views of customers about their product or their competitors. This information can be then used to improve the quality of their services or products accordingly.

Retrieving tweets that are opinionated about a specific topic is a non-trivial task. One of the many reasons is the informal nature of the medium, which has effected the emergence of new stylistic conventions such as emoticons, emphatic lengthening and slang terms widely used in Twitter. These informal stylistic conventions can, however, be a valuable source of information when retrieving tweets that express opinion towards the topic of interest. The use of emoticons usually implies an opinion [88] and emphatic lengthening has been shown to be strongly associated with opinionatedness [28]. For the rest of the chapter, we use the phrases *stylistic conventions* and *stylistic variations* interchangeably to denote the emerged textual conventions in Twitter such as the emoticons and the emphatic lengthening. Table 3.1 shows examples of some tweets with the stylistic variations that they contain. The stylistic conventions are only a subset of the writing style of users in Twitter. Writing style refers to a much wider manner that is used in writing [144].

Table 3.1. Examples of tweets with stylistic variations.

Example tweet	Stylistic variation
Hope you all have a fun day in the sun ☺	emoticon
Surprise party is getting ready to start**!!!!!!!**	exclamation marks
My new iPhone is so **cooooooool**	emphatic lengthening
One week till holidays! **#excited**	opinionated hashtag

The extent to which stylistic variations are used varies considerably among the different topics discussed in Twitter. That is, the number of the stylistic variations present in each tweet is dependent on its topic. For example, tweets about entertainment (i.e., movies, TV series) tend to use more stylistic variations than those that express opinion about social issues (i.e., immigration) or products (i.e., Google glass). This implies that stylistic variations do not have the same importance in revealing opinion across different topics.

In this chapter, we focus on the problem of Twitter opinion retrieval. The chapter starts with a topical analyis of Twitter data to try to understand how many topics are discussed in a single tweet. This step is important before proposing a new Twitter opinion retrieval model. Next, we propose a model to address

Twitter opinion retrieval which uses information about topics to retrieve those that are relevant and contain opinion about a user's query. The proposed model calculates opinionatedness by combining information from the tweet's terms and the topic-specific stylistic variations that are extensively used in Twitter. We compare several combinations of stylistic variations, including emoticons, emphatic lengthening, exclamation marks and opinionated hashtags. We evaluate the proposed model on the opinion retrieval dataset proposed by Luo et al. [95]. Experimental results show that stylistic variations are topic-specific and that incorporating them in the ranking function significantly improves the performance of opinion retrieval on Twitter.

In this chapter we address the first research question presented in Chapter 1:

RQ1 How can we find documents that are opinionated and express opinion about a topic in a microblogging collection? Can we make use of the textual peculiarities that are present in posts such as tweets to improve Twitter opinion retrieval?

This research question leads to the following more specific research questions:

RQ1.1 How many topics are discussed in a single tweet?

RQ1.2 What is the most effective combination of stylistic variations regarding topic-specific Twitter opinion retrieval?

RQ1.3 Is the importance of stylistic variations in indicating opinion topic dependent?

We proceed with a description of a topic classification task in Twitter in Section 3.2. Section 3.3 introduces the proposed topic-specific Twitter opinion retrieval approach. We present our experimental setup in Section 3.4 and our results and analysis in Section 3.5. Finally, the conclusions are presented in Section 3.6.

3.2 Topic Classification in Twitter

In this section we analyse the topical distribution of tweets and we investigate the question whether an individual tweet deals with a single or with multiple topics. Our main motivation is to gain a better understanding of the text of

tweets and get valuable information that can help us addressing the problem of opinion retrieval in Twitter. Although tweets are short because of the length limitation that was in force until recently, it is still not clear if an individual tweet deals with one or more topics. To this end, we apply Latent Dirichlet Allocation (LDA) proposed by Blei et al. [23] from the domain of topic modelling to explore the question of how many topics appear in a single tweet. The answer to this question is very important so we can understand how Twitter opinion retrieval should be addressed.

Topic models aim to identify text patterns in document content. Standard topic models include Latent Dirichlet Allocation (LDA) [23] and Probabilistic Latent Semantic Indexing (pLSI) [72]. LDA, one of the most well known topic models, is a generative document model which uses a “bag of words" approach and treats each document as a vector of word counts. Each document is a mixture of topics and is represented by a multinomial distribution over those topics. More formally, each document d in the collection is associated with a multinomial distribution over K topics, denoted θ. Each topic z is associated with a multinomial distribution over words, denoted ϕ. Both θ and ϕ have Dirichlet prior with hyperparameters α and β respectively. For each word in a document d, a topic z is sampled from the multinomial distribution θ associated with the document and a word w from the multinomial distribution ϕ associated with topic z. This generative process is repeated N_d times, where N_d is the total number of words in the document d. LDA defines the following process for each document in the collection:

1. Choose $\theta_d \sim \text{Dir}(\alpha)$,
2. Choose $\phi_z \sim \text{Dir}(\beta)$,
3. For each of the N words w_n:
   (a) Pick a topic z_n from the multinomial distribution θ_d
   (b) Pick a word w_n from the multinomial distribution ϕ_z

Topic models have been applied in a wide range of areas including Twitter. Hong and Davison [73] conducted an empirical study to investigate the best way to train models for topic modeling on Twitter. They showed that topic models learned from aggregated messages of the same user may lead to superior performance in classification problems. Zhao et al. [171] proposed a Twitter-LDA model that considered the shortness of tweets to compare topics discussed in Twitter with those in traditional media. Their results showed that Twitter-LDA works better than LDA in terms of semantic coherence. Inspired by the popularity of LDA, Krestel et al. [89] proposed using LDA for tag recommendation.

Based on the intuition that tags and words are generated from the same set of latent topics, they used the distributions of latent topics to represent tags and descriptions and to recommend tags.

The studies that deal with topic modeling in Twitter assume that an individual tweet deals with a single topic. However, to the best of our knowledge, there is no empirical study that validates this assumption. Thus we decided to use topic models as a way to gain a better understanding of the tweets with the aim to address Twitter opinion retrieval and not for identifying and analysing the topics that are discussed in Twitter. We also use LDA [23] to determine the topics of tweets, which are then used to learn the importance of the stylistic variations for each topic.

3.3 Topic-Specific Twitter Opinion Retrieval

Twitter opinion retrieval aims to develop an effective retrieval function which retrieves and ranks tweets accordingly to the likelihood that they express an opinion about a particular query. The proposed approaches for opinion retrieval usually follow a three step framework. In the first step, traditional IR methods are applied to rank documents by their relevance to the query. In the second, opinion scores are generated for the documents that were retrieved during the first step and, in the last step, a final ranking of the documents is produced based both on their relevance and opinionatedness towards the query.

In this section, we propose a new opinion retrieval model which leverages topic-specific stylistic variations of short informal texts such as tweets to calculate their opinionatedness. The proposed model calculates the opinionatedness of a document by combining two different opinion scores. The *term-based* component is based on the opinionatedness of the document's terms, whereas the *stylistic-based* component instead considers the stylistic variations present in the document.

Let $S_d(o)$ be the opinion score of a document (tweet) d based on its terms and $S_{ls,d}(o)$ be the opinion score of a document d based on the stylistic variations that d contains. Then the opinionatedness of the document d is the weighted sum of the two opinion score components and is calculated as follows:

$$S_{q,d}(o) = \lambda * S_d(o) + (1-\lambda) * S_{ls,d}(o)$$

where $S_{q,d}(o)$ denotes the opinion score o of a document d towards the query q, $S_d(o)$ is the opinion score based on the terms of the document, $S_{ls,d}(o)$ is the

opinionatedness of the document when it contains a subset of stylistic variations and $\lambda \in [0, 1]$.

3.3.1 Term-Based Opinion Score

The presence of opinionated terms in a document, and their probability of expressing opinion, is a popular approach to calculate the document's opinionatedness. A simple method is to calculate this score as the average opinion score over all terms in the document, thus:

$$S_d(o) = \sum_{t \in d} opinion(t) p(t|d) \tag{3.1}$$

where $p(t|d) = c(t, d)/|d|$ is the relative frequency of term t in document d and $opinion(t)$ shows the opinionatedness of the term. To calculate $opinion(t)$ we use the opinion scores that are given in the AFINN lexicon [112].

Since this is one of the most widely used methods to calculate the opinionatedness of a document, we also use this method as one of our baselines.

3.3.2 Stylistic-Based Opinion Score

Our method incorporates several stylistic variations of tweets into a ranking function to rank tweets according to their opinionatedness. The stylistic-based component of our model calculates an opinion score using the stylistic variations that a document contains. Let l be a stylistic variation taken from the list $L = (l_1, ..., l_i, ..., l_{|L|})$ which includes all the possible stylistic variations that reveal opinions. We then calculate the stylistic-based component as follows:

$$S_{ls,d}(o) = \sum_{l \in LS} LF(l, d) * ILF(l, d)$$

where LS is a subset of stylistic variations ($LS \subset L$), $LF(l, d)$ represents the frequency of the stylistic variation l in the document d and $ILF(l, d)$ represents the importance of the variation l, that is whether the stylistic variation is common across the documents or not. The inverse frequency ILF of the stylistic variation l controls the amount of opinion information that the specific variation holds.

We explore various ways of calculating the frequency LF of the stylistic variations. These are the following:

$$LF_{Bool}(l, d) = \begin{cases} 0, & \text{if } f(l, d) = 0 \\ 1, & \text{if } f(l, d) > 0 \end{cases}$$

$$LF_{Freq}(l,d) = f(l,d)$$

$$LF_{Log}(l,d) = 1 + \log f(l,d)$$

where $f(l,d)$ is the number of occurrences of variation l in document d.

To model the relative importance of each stylistic variation l across the documents we consider the following methods:

$$ILF_{Inv}(l,d) = \log \frac{N}{1+n_l} \tag{3.2}$$

$$ILF_{Prob}(l,d) = \log \frac{N-n_l}{n_l} \tag{3.3}$$

where n_l can also be written as $|d \in D : l \in d|$ and denotes the number of documents that belong to the collection D and contain the stylistic variation l.

Thus, the importance of a given stylistic variation l depends on how frequently it is used in the collection D.

3.3.3 Topic Specific Stylistic-Based Opinion Score

The assumption made in the existing literature, that the stylistic variations are used with the same frequency across documents of different topics, is not accurate. Informal stylistic variations are used with differing frequencies depending on the topic discussed. For example, tweets that are relevant to a TV series probably contain more stylistic variations than those that are relevant to a social issue, such as immigration. That means that the probability that stylistic variations imply opinion depends on the topic of the tweet. In other words, if emoticons are extensively used in tweets about a specific topic, then their ability to imply opinion decreases.

Based on this assumption, we propose using *topic-specific stylistic variations*. To this end, we first apply topic modeling to determine the topic of a tweet and then we use this information to calculate the stylistic-based component of our approach, that is the opinionatedness of a tweet when it contains a specific stylistic variation. More formally, let $T = (T_1, ..., T_i, ..., T_{|T|})$ be the topics extracted after applying a topic model on the collection D, and $D_T = (d_1, ..., d_t)$ the documents that are assigned to the topic T_i. Then, the relative importance ILF of each stylistic variation l is calculated using equations 3.2 and 3.3 with the difference that n_l denotes the number of documents that belong to collection D_T and contain the stylistic variation l. In other words, n_l is calculated as $|d \in D_T : l \in d|$, where D_T is a collection of documents that were assigned the same topic T_i.

3.3.4 Combining Relevance and Opinion Scores

To generate the final ranking of documents according to their relevance and opinionatedness, we combine the relevance score with the opinionatedness of the tweet:

$$S_{o,q}(d) = S_d(q) * S_{q,d}(o)$$

where $S_d(q)$ is the relevance score of d given topic t and $S_{q,d}(o)$ is the opinionatedness of d. The relevance score $S_d(q)$ can be estimated using any existing IR model.

3.4 Experimental Design

In this section, we describe the experimental design of our study. We start with a description of the dataset followed by the experimental settings. Then we present the opinion lexicon and stylistic variation we used and finally the evaluation process we followed to measure the effectiveness of our model on the Twitter opinion retrieval task.

3.4.1 Dataset

To evaluate our methods we used the dataset created by Luo et al. [95], which is, so far and to the best of our knowledge, the only dataset that has been used for Twitter opinion retrieval. Initially, Luo et al. crawled around 30 million tweets using Twitter API. Then they implemented a search engine that was based on Lucene-BM25 and they asked seven people (i.e., annotators) to use it and submit queries. The annotators labelled the retrieved tweets regarding relevance and opinionatedness to the query. Finally, they managed to collect 50 topics and 5000 judged tweets.

We should note that there is another dataset which could be used for opinion retrieval in Twitter. This dataset was created by Paltoglou and Buckley [118] who annotated part of the Microblog dataset [115] provided by TREC with subjectivity annotations. However, as this dataset has not yet been used in any study, we would not be able to make direct comparisons of our methods. Therefore, we decided to use the first.

3.4.2 Experimental Settings

To create the index, we removed URLs, hashtag symbols (#) placed in front of some terms and character repetitions that appear consecutively more than twice in a term. We indexed the collection with the Terrier IR system[1]. Our preprocessing also involved stop-word removal using the snowball stop word list[2] and stemming using the Porter stemmer [127].

To avoid overfitting the data we performed 5 fold cross-validation on the 50 queries. For each fold we used 40 queries for the training phase and 10 for testing. The training and test data were kept separate in all phases of our experiments. We performed our experiments under two different settings: *non topic-based* and *topic-based*. For the non topic-based settings, we applied the proposed method on the whole collection without considering the tweet's topic. For the topic-based settings we first applied LDA to detect the topics and then we applied the proposed method on tweets of the same topic. To estimate the LDA parameters we used a Gibbs sampler. Since the Gibbs sampler is a stochastic method, and therefore produced different outputs per run, we report the mean performance of the methods based on ten runs.

3.4.3 Opinion Lexicon and Stylistic Variations

To identify the opinionated terms we used the AFINN Lexicon, as proposed by Nielsen [112]. AFINN contains more than 2000 words, each of which is assigned a valence from -5 to -1 for terms with a negative sentiment or from 1 to 5 for terms with a positive sentiment. We chose this lexicon as it contains affective words that are used in Twitter. We took the absolute values of the scores since we do not consider sentiment polarity in our study. We used $MinMax$ normalisation [68] to convert the valence score of a term to opinion score.

To calculate the stylistic-based component of our model, we identified, for each tweet, the number of emoticons, exclamation marks, terms under emphatic lengthening and opinionated hashtags as follows:

- *Emoticons*: Number of emoticons in a tweet. We used the list of emoticons provided in Wikipedia[3]. We considered all emoticons to be opinion-bearing. Therefore, we did not distinguish them by their subjectivity, sentiment or emotion they express.

[1]Available at: `http://terrier.org/`
[2]Available at: `http://snowball.tartarus.org/`
[3]See `http://en.wikipedia.org/wiki/List_of_emoticons`

- *Exclamation marks*: Number of exclamation marks in a tweet.

- *Emphatic lengthening*: Number of terms under emphatic lengthening in a tweet. Emphatic lengthening refers to terms that contain more than two repeated letters (e.g., cooooooool).

- *Opinionated hashtags*: Number of opinionated hashtags. As opinionated hashtags we considered any hashtag that is contained in the AFINN opinion lexicon. For example, the hashtag *#love* is considered an opinionated hashtag because the term *love* appears in the AFINN opinion lexicon.

3.4.4 Evaluation

We compare the proposed opinion retrieval method with two baselines. The first, *BM25*, is the method with the best performance in Twitter opinion retrieval according to the results presented in [95]. The *Relevance-Baseline* is based purely on topical relevance and does not consider opinion. As a second baseline, we use the term-based opinion score (equation 3.1). The *Opinion-Baseline* considers opinion and therefore it is a more appropriate baseline to compare our results with. To evaluate the methods, we report *Mean Average Precision* (MAP), which is the only metric reported in previous work [95] on Twitter opinion retrieval. MAP is a standard evaluation metric for information retrieval tasks [99]. For a single information need, Average Precision is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved. More formally, if $q_j \in Q$ $\{d_1, ..., d_{mj}\}$ is the set of relevant documents for a query q_j and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k then:

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk})$$

Finally, to compare the different methods we used the Wilcoxon signed ranked matched pairs test [164] with a confidence level of 0.05.

3.5 Results and Analysis

In this section, first we describe the results on the topic classification task and next, we report the results using the topic specific opinion retrieval approach.

Table 3.2. Topic descriptions.

Sample topics from Twitter	Sample topics from blogs
obama barack barrack news	back room house car
music awards carpet red	face head eyes hand
men half man lol	baby gift young feel
disney world walt top	photo online family world
steve jobs apple biography	dance join date dressed

3.5.1 Number of Topics in a Tweet

In order to identify the number of topics that are discussed in a single tweet, we applied the LDA [23] topic model on the dataset proposed by Luo et al. [95]. In addition, we applied LDA on a blogs collection with the aim to compare the number of topics that are discussed in a single blog to those of a single tweet. To create the blogs collection, we randomly chose blogs from the TREC 2008 blog track [114]. The tweets and blogs collections used in this study contain the same number of documents. For the analysis, we applied Gibbs sampling for the LDA model parameter estimation and inference as proposed by Yao et al. [168].

We considered each tweet and each blog as a document. We tried a number of different values for the K parameter that represents the number of topics, ranging from 1 to 500. We should note that a significant number of studies have reported that their best results are achieved when the number of topics is set to 100 [73]. We set the number of iterations to 1000. Table 4.2 shows a list of five topics that were discovered in the collection of tweets and the collection of blogs when the number of topics was set to 100. From the topics descriptions, we observe that LDA manages to cluster together terms that refer to the same topic.

Figure 3.1 shows the rate on which the percentage of tweets and blogs with a single topic changes as the K parameter increases. We observe that the percentage of tweets that are about a single topic remains high even when the parameter is set to 500 topics. In contrast, this percentage drops really quickly in blogs even when the parameter is lower than 100. This fast drop implies that a single blog is more likely to contain information about more than one topics compared to a single tweet.

The fact that the majority of tweets is about a single topic implies that if a tweet is opinionated then it is likely that it will be opinionated for this topic.

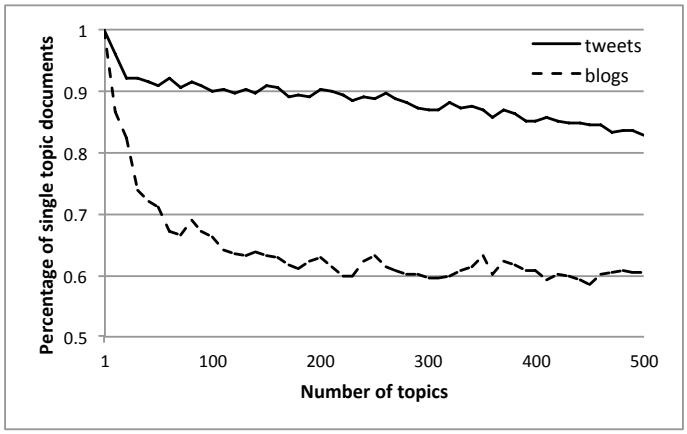

Figure 3.1. Rate on which percentage of tweets/blogs with a single topic changes as the number of topics is increased.

This means that proximity-based opinion retrieval in Twitter may not be as effective as it is for blogs. In fact, experiments with proximity-based method were performed [50] and showed that proximity is not useful for Twitter opinion retrieval.

The topic analysis is also used for the topic specific Twitter opinion retrieval approach. To this end, we tried a number of different values for the K parameter, which represents the number of topics, ranging from 1 to 200 with a step of 5. We set the number of iterations to 2000. The minimum log likelihood is obtained for 65 topics.

3.5.2 Topic Specific Twitter Opinion Retrieval

In this section, we present the results of our approach. Table 3.3 presents the results of Twitter opinion retrieval when different stylistic variations are combined. Any of the approaches of calculating LF and ILF presented in Section 3.3.2 can be used to evaluate the effectiveness of the different combinations. For the results displayed in Table 3.3 we applied LF_{Log} and ILF_{Inv} under topic-based settings. We observe that all the three examined combinations ($LF_{Log}ILF_{Inv}$-Emot-Excl, $LF_{Log}ILF_{Inv}$-Emot-Excl-Emph, $LF_{Log}ILF_{Inv}$-Emot-Excl-Emph-OpHash) perform significantly better than both the relevance and opinion baselines. The best performance is achieved when we combined *emoticons*, *exclamation marks* and *emphatic lengthening*. This is a very interesting result that shows that integrating the most useful stylistic variations with the opinionatedness of the terms

Table 3.3. Performance results of the $LF_{Log}ILF_{Inv}$ method under topic-based settings using different combinations of stylistic variations over the baselines. A star($*$) and dagger($\dagger$) indicate statistically significant improvement over the relevance and opinion baselines respectively.

	MAP
Relevance-Baseline	0.2835
Opinion-Baseline	0.3807$*$
$LF_{Log}ILF_{Inv}$-Emot-Excl	0.4314$*$ $\dagger$
$LF_{Log}ILF_{Inv}$-Emot-Excl-Emph	0.4413$*$ $\dagger$
$LF_{Log}ILF_{Inv}$-Emot-Excl-Emph-OpHash	0.4344$*$ $\dagger$

Table 3.4. Performance results of different LF and ILF combinations, based on emoticons, exclamation marks and emphatic lengthening. A star($*$) indicates statistically significant improvement over the non topic-based settings for the same approach.

LF - ILF	Non Topic-Based	Topic-Based
$LF_{Bool}ILF_{Inv}$	0.4279	0.4419$*$
$LF_{Freq}ILF_{Inv}$	0.4279	0.4398
$LF_{Log}ILF_{Inv}$	0.4275	0.4413$*$
$LF_{Bool}ILF_{Prob}$	0.4279	0.4427$*$
$LF_{Freq}ILF_{Prob}$	0.4279	0.4421$*$
$LF_{Log}ILF_{Prob}$	0.4275	0.4429$*$

into a ranking function can be very effective for Twitter opinion retrieval.

Table 3.4 shows the performance of the proposed model on non topic-based and topic-based settings for Twitter opinion retrieval. We evaluate the effectiveness of different combinations of approaches in calculation of LF and ILF. We observe that most of the approaches perform statistically better under the topic-based settings compared to the non topic-based settings. This is a very interesting result which shows that stylistic variations are indeed topic-specific and the amount of the opinion information they hold depends on the topic of the tweet. We also observe that there is no statistical difference between the different LF and ILF approaches when they are compared under the same settings.

In addition, we performed a per topic analysis to compare the model under topic-based versus non topic-based settings. Figure 3.2 shows the increase and

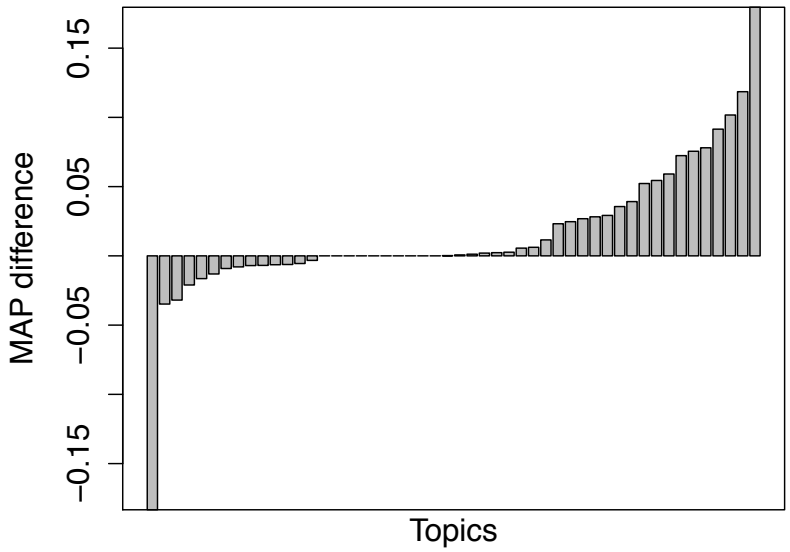

Figure 3.2. Difference in performance between the topic-based $\text{LF}_{Log}\text{ILF}_{Prob}$ and the non topic-based $\text{LF}_{Log}\text{ILF}_{Prob}$ model. Positive/negative bars indicate improvement/decline over the non topic-based $\text{LF}_{Log}\text{ILF}_{Prob}$ model in terms of MAP.

decrease in Average Precision (AP) when comparing the best run ($\text{LF}_{Log}\text{ILF}_{Prob}$) of the proposed model under topic-based against non topic-based settings. The plot shows that there are topics for which there is an improvement over the non topic-based settings as well as topics for which the topic-based setting is not helping. However, the topic-based $\text{LF}_{Log}\text{ILF}_{Prob}$ model has more topics for which it improves performance compared to the number of topics that it hurts. This shows that in general considering topic-specific stylistic variations in ranking opinionated tweets is helpful.

In addition, Table 3.5 shows the three topics that were helped or hurt the most when the $\text{LF}_{Log}\text{ILF}_{Prob}$ model was used under the topic-based compared to the non topic-based settings. We observe that the topics that were helped are those related to topics about products (e.g., Lenovo, galaxy note), whereas the topics that were hurt the most have to do with science (e.g., big bang).

Finally, we compare the performance of our proposed approach with the performance of the best run presented by Luo et al. [95] and report the comparison result in Table 3.6. We observe that our best runs largely improve their best

Table 3.5. Topics that are helped or hurt the most in the $\text{LF}_{Log}\text{ILF}_{Prob}$ model under topic-based compared to non topic-based settings.

Helped		Hurt	
Title	ΔMAP	Title	ΔMAP
iran	0.1795	new start-ups	-0.1833
Lenovo	0.1185	iran nuclear	-0.0480
galaxy note	0.1017	big bang	-0.0319

Table 3.6. Results on Δ MAP for best runs over Opinion-Baseline.

Run	Map	Δ MAP
Opinion-Baseline	0.3807	-
BM25_Best	0.4181	9.82%
$\text{LF}_{Log}\text{ILF}_{Prob}$-Emot-Excl-Emph	0.4429	16.33%
$\text{LF}_{Bool}\text{ILF}_{Prob}$-Emot-Excl-Emph	0.4427	16.28%

reported result (denoted as BM25_Best). Finally, we should mention that their method uses SVM^{Rank} and their best run (BM25_Best) is trained using a number of social features (URL, Mention, Statuses, Followers) together with BM25 score, and Query-Depedent opinionatedness (Q_D) features.

3.6 Conclusions

In this chapter, we considered the problem of Twitter opinion retrieval that can be used to find tweets that express an opinion about a specific topic. First, we analysed the topical distribution of tweets and investigated the question whether a single tweet deals with a single or with multiple topics with the aim to get valuable information for addressing the problem of opinion retrieval in Twitter. The results showed that the vast majority of tweets deal with a single topic.

More importantly, we proposed a topic-based method that uses topic-specific stylistic variations to address the problem of Twitter opinion retrieval. We studied the effect of different approaches and stylistic variations in the performance of Twitter opinion retrieval. The results showed that stylistic variations are good indicators for identifying opinionated tweets and that opinion retrieval performance is improved when emoticons, exclamation marks and emphatic lengthening are taken into account. Additionally, we demonstrated that the importance of stylistic variations in indicating opinionatedness is indeed topic dependent as

our topic model-based approaches significantly outperformed those that assumed importance to be uniform over topics.

Chapter 4

Tracking Sentiment Evolution

This chapter focuses on tracking sentiment about a specific topic over time. First, we plot signals that show the topic's popularity and sentiment evolution. We explore the effectiveness of state-of-the-art time series tools (mean, naïve, ARIMA) in predicting positive and negative sentiment in future. In addition, we use outlier detection to identify spikes that are likely related to events that cause a sentiment change. Next, we propose a method that combines LDA topic model with KL-divergence to extract and rank the causes behind a sentiment spike. In addition, the chapter presents a labelled collection of tweets that can be used for extracting and ranking the likely causes of a sentiment spike. Our results show that the state-of-the-art time series approaches are very useful in tracking sentiment over time. Frequency analysis is very useful to observe positive and negative sentiment evolution over time. Also, we show that in some cases the naïve that is a simple forecasting approach can outperform ARIMA in predicting sentiment in the future. Finally, we show that LDA & KL-divergence approach can effectively be used to extract and rank the topics identified on sentiment spikes.

4.1 Introduction

Public opinion changes over time, therefore tracking opinion and sentiment evolution is very critical for the interested parties. The analysis of sentiment evolution not only gives the opportunity to find patterns and seasonality in sentiment but also to forecast sentiment in the future that is a very important tool for the interested parties. Forecasting sentiment towards a specific entity is very challenging since it depends on many external factors that in some cases are very difficult to be predicted.

Tracking sentiment evolution gives the opportunity to identify sudden spikes

of sentiment and, more importantly, to get insights on what has caused these sentiment spikes. Detecting sentiment spikes allows to take quick reactions in response to user sentiment (especially a negative one). Understanding the reasons that likely caused a sentiment spike provides valuable information for governments and companies to be proactive and improve their tactics. For example, suppose that the negative sentiment towards a political person increases during an electoral campaign, then the respective Press Office can extract the causes for such increase and avoid, in future, those situations that may have caused the negative sentiment. Another example is related to movie production companies that may want to better understand what people think of the actors that participated in their movies. Sentiment spikes towards an actor may impact the success of a movie and analysing these spikes is crucial for movie production companies.

Extracting causes of sentiment spikes is very challenging since sentiment spikes can be also caused by external factors that are very difficult to capture. Previous work by Montero et al. [110] focused on identifying the likely causes of emotion spikes using the most popular keyphrases. Tan et al. [148] proposed an extension of LDA to rank the extracted foreground topics according to their popularity within the sentiment variation period. Although using the popularity of the topics seems sensible, its effectiveness requires deeper analysis.

The problem of tracking sentiment evolution with respect to an entity and extracting and ranking the causes of a sentiment spike can be viewed as a five-step approach as follows: (i) first, we need to observe the sentiment evolution towards the entity of interest; (ii) given this sentiment evolution, we forecast sentiment in the future; (iii) we identify the most important sentiment spikes; (iv) next we extract the topics that were discussed when the sentiment spike occurred; and (v) finally, we rank the extracted topics based on their contribution to the sentiment spike.

In this chapter, we focus on the problem of tracking sentiment evolution towards the following entities: *android lollipop*, *pretty little liars*, *Michelle Obama*, *Angela Merkel* and *Angelina Jolie*. In particular, first we propose conventional time-series approaches to track the evolution of the sentiment over time, forecast sentiment and extract sentiment spikes. Time series models seem to be an appropriate tool for sentiment tracking and can be used to understand data and identify trends and seasonality. First, we apply state-of-the-art forecast approaches to predict positive and negative sentiment in the future. In addition, we use outlier detection to identify outliers which are likely to be related to events that caused a sentiment change. Next, we propose a new method that combines LDA topic model with KL-divergence to extract and rank the causes behind a sentiment spike. LDA allows to extract the topics discussed in the time window before

the sentiment spike and the KL-divergence to detect the topics which probably caused the sudden change. We finally rank these topics according to their contribution to the sentiment spike. We create a collection of tweets to assess the effectiveness of our methodology. The ground truth was collected using a popular crowdsourcing platform.

In this chapter, we address the following research question:

RQ2 How can we model opinion evolution and identify the important causes of opinion change?

This research question leads to the following more specific research questions:

RQ2.1 Can conventional time series methods be applied to track sentiment evolution over time and forecast sentiment in the future?

RQ2.2 Can outlier detection be applied to identify sentiment spikes?

RQ2.3 How does an approach based on a combination of topic model with KL-divergence perform in extracting the likely reasons that caused a sentiment spike?

In the rest of the chapter we proceed with applying time series to track sentiment evolution over time and forecast sentiment in Section 4.2. Section 4.3 introduces the proposed LDA & KL-divergence approach that extracts and ranks likely reasons of sentiment spikes. We present our collection in Section 4.4 and our results and analysis in Section 4.5. Finally, the conclusions are presented in Section 4.6.

4.2 Modelling Sentiment Evolution

In this section, we explain how we applied time series approaches to track sentiment evolution, to forecast sentiment and to identify sentiment spikes given an entity.

4.2.1 Tracking Sentiment Evolution

Here, we describe the approach we use to track the sentiment evolution towards a specific entity. To address this problem, we have to estimate the overall sentiment

towards the entity on a given point in time. Therefore, we use frequency analysis approach.

Frequency Analysis. The initial step in time series analysis is to explore the data and observe how the frequencies change. Let $X = (x_1, x_2, .., x_n)$ be a set of data observed at consecutive and equal time intervals denoted as $\{t_1, t_2, ..., t_n\}$. Based on this, we can observe how the number of tweets about an entity z changes per day, denoted as $N_t(z)$. This time series is an indicator of the popularity of the entity independent from the sentiment that is expressed.

To explore sentiment trends we need to measure the frequencies of tweets that express a specific sentiment. We only consider positive and negative sentiment. However, the approach can be also applied on sentiment that represents emotions such as love, anger, sadness etc. Let $N_t(z,s)$ be the number of tweets that express a sentiment s towards a specific entity z posted during a particular time period t and $N_t(z)$ the number of total tweets posted towards z at t. Then, we can define the ratio of tweets that share a common sentiment s as:

$$r_t(z,s) = \frac{N_t(z,s)}{N_t(z)}$$

Based on this, we can measure the *sentiment velocity* that represents the rate of sentiment change and is defined as: $Vel_t(z,s) = N_{t+1}(z,s) - N_t(z,s)$. We can also measure *sentiment acceleration* that represents the rate of change of sentiment velocity at a particular time t. The sentiment acceleration of entity z and sentiment s is defined as: $Acc_t(z,s) = Vel_{t+1}(z,s) - Vel_t(z,s)$. Plotting sentiment velocity and acceleration is useful not only to observe how a specific sentiment changes but also to detect if there is any emerging sentiment. For example, a negative emerging sentiment about a specific entity means that the company should be alerted and act promptly.

Time Series Decomposition. To get a better understanding of the data we further apply time series decomposition. Decomposition is a statistical tool that deconstructs a time series into several components and is of crucial importance for the subsequent analysis and modeling. Time series data can be decomposed into three components: the trend (T_t), the seasonal (S_t) and the random (R_t) components. Based on this, we can define the time series X_t as a function of these components: $X_t = f(T_t, S_t, R_t)$. The decomposition is such that the three components add up to the original time series.

One well known decomposition method is the additive decomposition defined as: $y_t = T + S + R$. The trend component reflects the long-term increase or decrease in the data. Another way to find the trend is to smooth the data and remove any wide variation that can be considered related to seasonality. *Moving*

average is one of the most well known smoothing techniques and can be very helpful to identify patterns and trends in time series because it evens out short term fluctuations and makes the trend more apparent. According to this approach, the value of data at time t is the unweighted mean of the data observed at the k previous time periods. This is defined as:

$$MA_t = \frac{x_{t-(k-1)} + ... + x_{t-1} + x_t}{k}$$

The seasonal component represents patterns that are repeated at fixed periods like days, weeks, months etc. *Seasonal adjustment* is a method for removing the seasonal component of a time series. This is useful to observe the data without the seasonal effects that may have an influence on them. One typical example is that users may tweet more at specific days or specific time. Seasonally adjusted data can be constructed as: $SeasAdj_t = X_t - S_t$. Finally, the random component represents noise in data and can be constructed by removing the trend and seasonal components as: $R_t = X_t - T_t - S_t$.

4.2.2 Sentiment Forecast

We apply the following state-of-the-art forecasting approaches to predict the negative and positive sentiment: *mean*, *naïve* and *Auto-Regressive Integrated Moving Averages (ARIMA)*. We now provide some details about how these approaches estimate the predictions.

Mean. According to this approach the forecast of all the future values are equal to the mean of the historical data. Let $X = (x_1, x_2, .., x_n)$ be a set of data observed at consecutive and equal time intervals denoted as $\{t_1, t_2, ..., t_n\}$. Then the forecasts can be estimated as:

$$y_{n+1} = (x_1 + ... + x_n)/n$$

Naïve. This method applies the value of the last observation to all the forecasts. In particular, all the future values are set to x_n, where x_n is the last observed value. More formally, we have:

$$y_{n+1} = x_n$$

Auto-Regressive Integrated Moving Averages (ARIMA). ARIMA is a forecasting technique that predicts the future values of a series based on its own past values. A stationary time series is required to apply ARIMA. A time series is stationary when its statistical properties such as mean, variance, autocorrelation

are all constant over time. If the data are not stationary, they can be differenced that is a way of transforming a nonstationary series to a stationary one.

Another important component of ARIMA are the autocorrelations which are numerical values that indicate how a data series is related to itself over time. More precisely, autocorrelations measure how strongly data values at a specified number of periods apart are correlated to each other over time.

In general, an ARIMA model is defined as an *ARIMA(p,d,q)* model, where:

- p is the number of autoregressive components,
- d is the number of differencing operators, and
- q is the number of lagged forecast errors in the prediction equation.

More formally, the general model for ARIMA is as follows:

$$y_t = c + \phi_1 Y_{t-1} + ... + +\phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + +\phi_q y_{t-q}$$

where Y_t is the differenced time series value.

4.2.3 Identifying Sentiment Spikes

To identify sentiment spikes, we use an outlier detection approach adopted from the field of time series. Outliers are visually depicted as sudden peaks and in most of the cases are caused by some important events. These important events not only influence the popularity of an entity but also the public sentiment towards the entity. In order to identify outliers we use the following equation:

$$e_i = x_i - x_i'$$

where x_i is the observation i and x_i' is the prediction of the observation i. In other words, this equation calculates the ordinary residuals for each observation. We use *LOESS* [34] that is also known as the locally weighted polynomial regression model and interquartile range to detect the residuals. Let Q_1 and Q_3 be the lower and upper quartiles, respectively, then the outliers are the observations that are outside the following range:

$$[Q_1 - k * (Q_3 - Q_1), Q_3 + k * (Q_3 - Q_1)]$$

where k represents the span of the range, and it is usually set between 1.5 and 3.0.

4.3 LDA & KL-divergence Approach

As explained before, our ultimate aim is, given a sentiment spike, to get a ranked list of topics that can reflect their contribution in having caused the specific spike.

More formally, let $P_{e,s}$ be a sentiment spike that is related to an entity e and a sentiment s. The first step is to decide the time window that we need to focus. The starting point of the time window can be the point when the sentiment s had started increasing. This time point is denoted as t_{start}. As an ending point of the time window, we consider the time point that the spike occurred and this is denoted as t_{prev}. Figure 4.1 shows the phases of the sentiment spike from the point when the sentiment started increasing to the time that has become the prevalent sentiment.

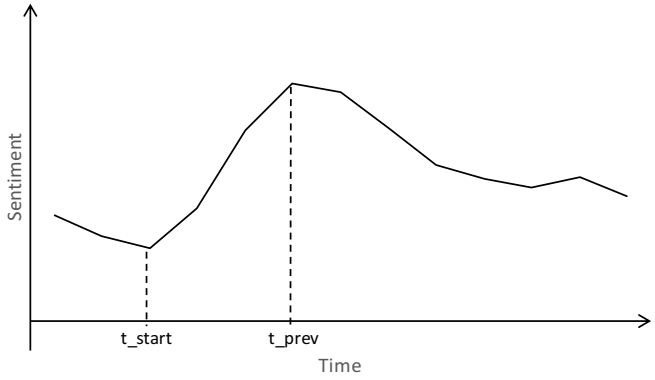

Figure 4.1. Starting and prevalent time point of a sentiment spike.

The topics that were discussed within the time window $[t_{start}, t_{prev}]$ represent the topics that contributed to the sentiment spike. We apply LDA to all the tweets within the time window $[t_{start}, t_{prev}]$ to extract the topics discussed within this window. The next step is to rank the extracted topics according to their contribution to the sentiment spike. For this we use Kullback-Leibler divergence [90]. Kullback-Leibler divergence (KL-divergence) is a measure of the difference between two probability distributions, where one typically represents the actual distribution of observations and the other one is an approximation of it. For this reason, KL-divergence is usually applied to measure the information lost between one distribution and its approximation. Although the KL-divergence is not a true metric, it is often interpreted as a way of measuring the distance between two probability distributions. Inspired by this, we show how it is possible to use the

KL-divergence to determine the causes of a sentiment spike in Twitter.

In particular, we present an approach based on computing the total distribution of the sentiment towards an entity and compare it against another distribution of the sentiment towards the same entity, but which is obtained by removing tweets belonging to a specific topic. The intuition is that if a sentiment has changed for a specific reason, such reason can be captured by the tweets that are about a related topic. We can quantify this by using the difference between the distribution of the sentiment using all tweets and the distribution of sentiment computed using all the tweets except for those belonging to a specific topic. This can be done for all topics, namely, considering one topic at a time we remove the tweets about it and compare the obtained distribution against the original one. As a result, we can determine for which topic the two distributions are more different and rank the topics accordingly. The final ranking should reflect the contribution of the topics to the sentiment spike.

To present the approach more formally, we introduce some notation. Let $S_e = (s_1, ..., s_j, ..., s_{|N|})$ be the distribution of the sentiment s over time towards an entity e where s_1 is calculated based on the number of tweets that express sentiment s at the first time point. In addition, let $T = (T_1, ..., T_z, ..., T_{|T|})$ be the topics extracted after applying LDA on the tweets' collection D. Also, the collection D is made of the tweets that are related to the sentiment spike $P_{e,s}$ and which were actually posted within the time window $[t_{start}, t_{prev}]$.

In addition, let $D_{T_z} = (d_1, ..., d_t)$ be the tweets that were assigned to the topic T_z. Then, for each topic T_z we define a new distribution that we denote as $S^*_{e,T_z} = S_e - S_{e,T_z}$ and which is calculated based on the sentiment and after removing the tweets that belong to D_{T_z}. Based on this, we calculate the KL-divergence between the distribution S_e and S^*_{e,T_z} as:

$$KL(S_e, S^*_{e,T_z}) = \sum_i S_e(i) \log \frac{S_e(i)}{S^*_{e,T_z}(i)}$$

Finally, the topics are ranked based on their KL-divergence value. The highest the value, the highest the contribution of the specific topic to the sentiment spike. For example, if we track the negative sentiment towards an entity, then the topic with the highest KL-divergence value is considered to be the most responsible to the sentiment spike. Figure 4.2 shows an example of one initial and of three different distributions related to three topics. In this example, *Topic3* has the largest contribution to the sentiment spike, because removing it makes the difference between the two distributions bigger compared to the other two topics.

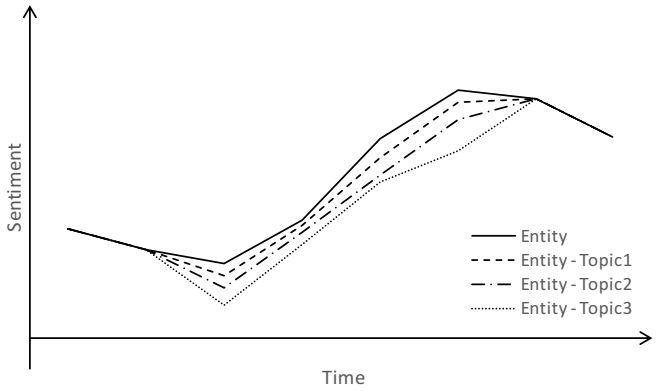

Figure 4.2. Example of the distributions of three topics compared to the entity's distribution.

4.4 Sentiment Spikes Collection

In order to create a collection that can be used for detecting and ranking the triggers of sentiment spikes, we need to analyse the sentiment evolution towards the entity of interest. We build the collection based on the following pipeline: (i) collect the tweets towards three different entities; (ii) annotate tweets by sentiment polarity (iii) identify sentiment spikes; (iv) identify candidate topics that triggered each sentiment spike; and (v) produce a ranking of the candidate topics based on their contribution to the sentiment change. In the following sections, we present details of some of these steps.

4.4.1 Data Collection

The task of extracting and ranking sentiment spikes' triggers requires a dataset that spans over several months. There are other available collections of tweets but most of them are over a short period of time. Due to Twitter's restrictions, it was not possible to extend the available datasets by additional months, since given an entity you can collect the tweets that are published no more than two weeks earlier. Hence, we used the Twitter API to collect our data. Our collection spans from the 10th of April to the 31st of December 2015. To create the collection, we focused on three entities that are well known personalities: *Michelle Obama* (1,076,690 tweets), *Angela Merkel* (1,369,306 tweets), and *Angelina Jolie* (1,264,828 tweets). We have decided to focus only on three entities because annotating the collection requires a lot of resources.

To measure the sentiment of a tweet we used SentiStrength [150] that has been shown to be effective in different social media platforms, and it does not need any training. The main reason to use SentiStrength is that we did not have any training data and it would a very costly process to obtain training data for our collection. Given that our collection contains millions of tweets, we would manage to annotate only a very small part of the collections that it would not be representative. Therefore, we chose to use the empirical scores of SentiStrength that do not depend on any context. Although SentiStrength can also assign a sentiment score to each tweet, we only considered the three following classes: *positive, neutral,* and *negative*. For the entities *Michelle Obama* and *Angela Merkel* we focused on negative sentiment polarity, whereas for *Angelina Jolie* we focused on positive sentiment polarity. The reason for this decision is that users tend to be critical with people or topics related to politics whereas they tend to post tweets that mostly express positive opinions about celebrities.

To identify the topics discussed around the date that the sentiment spike occurred, we applied the Latent Dirichlet Allocation (LDA) topic model on the tweets posted between the date that the number of tweets started increasing and the actual date of the sentiment spike. We treated each tweet as a document and we extracted 10 topics for each spike. Before applying LDA, we removed all the occurrences of terms that were referring to the entity (e.g., for the entity *Michelle Obama* we removed all the occurrences of the terms *Michelle* and *Obama* as well as their variations). For the analysis, we applied Gibbs sampling for the LDA model parameter estimation and inference. We set the number of iterations to 2000.

The next step was to identify the sentiment spikes. For this problem, we used the approach presented in Section 4.2.3. Table 4.1 summarises some of the most important statistics about the sentiment spikes of our collection.

4.4.2 Design of the CrowdFlower Experiment

As already mentioned, our final aim was to get a ranking of the extracted topics based on the sentiment polarity and strength of the tweets that belong to each topic. First, we extracted the topics discussed before and on the specific date of the sentiment spike using LDA. Each topic was related to a list of keywords and a set of tweets. One problem that we encountered was that it was not possible to show to the annotators the whole set of tweets since most of the sets contained thousands of tweets. Therefore, we decided to get a sample of the tweets that belong to each set. This number ranged from 3 - 6 % depending on the popularity of the topics. The same percentage was used for all the topics that belong to the

Table 4.1. Statistics of the different sentiment spikes.

	Start date	Date of spike	Number of tweets
Michelle Obama	03 May 2015	07 May 2015	20,839
	30 May 2015	31 May 2015	7,496
	13 July 2015	16 July 2015	12,521
	10 Aug. 2015	12 Aug. 2015	4,890
	06 Nov. 2015	08 Nov. 2015	29,214
Angela Merkel	21 May 2015	23 May 2015	10,264
	01 July 2015	03 July 2015	31,454
	26 July 2015	29 July 2015	8,627
	15 Aug. 2015	16 Aug. 2015	2,507
	25 Nov. 2015	29 Nov. 2015	28,578
Angelina Jolie	17 Apr. 2015	19 Apr. 2015	9,357
	02 June 2015	04 June 2015	18,360
	20 July 2015	21 July 2015	7,099
	24 Sep. 2015	26 Sep. 2015	12,220
	29 Sep. 2015	30 Sep. 2015	9,654

same spike. This is important since we wanted annotators to have an estimate of the popularity of each topic.

Another challenge was how to select each sample of tweets. One possible solution would be to rank them chronologically and then use systematic sampling, that would be adding a tweet into the sample using a constant step (e.g., select one tweet every 100 tweets). However, there would still be a risk of creating a sample that was not representative. Therefore, we ranked the tweets based on their similarity with the topic. To do this, we used the representative keywords of the topic generated by LDA and we ranked its tweets based on the number of their common words. Also, we tried to exclude as many retweets as possible since repeated content could be annoying for the annotators. However, in some cases showing retweets was inevitable. In particular, for those topics in which we had only a small number of distinct tweets and the majority of the tweets were simple retweets.

For our evaluation we needed a ranking of the extracted topics, however, it would be very tricky and difficult for the annotators to go through the lists of tweets, each one representing a topic, and give back a ranking of these topics. Therefore, we asked the annotators to rate only one topic at a time. Given the list of keywords of the topic and the sample of the tweets, we asked the following

question: *What is the polarity and the strength of the sentiment/emotions expressed in this set of tweets?* and one annotator had to give a rating for all the topics that we extracted from the period before and related to a sentiment spike. Note that the order we showed the topics was completely random and different for each annotator.

4.4.3 Annotators

Collecting human judgements with crowdsourcing has the risk of low-quality submissions. The most popular technique for checking the quality of submissions is having test questions which are used in the quiz page (e.g., the first page displayed to the annotator) to train the annotators, and which are also randomly displayed in each page for checking the performance of each annotator. In our case, each annotator had to evaluate all the topics that were extracted from a specific spike (i.e., all the ten topics extracted from a specific date must be displayed on the same page) and therefore we had to annotate at least 8 test questions (out of 10) which could be used in the quiz page. Due to this design restrictions of CrowdFlower and the design of our experiment, it was not possible to have enough test questions during the training and the execution of the task.

Instead of having test questions, we measured the accuracy of the submissions afterwards and removed annotators with low-quality or biased submissions. To do so, we followed a specific process. In particular, for each topic, we first tried to understand the trend in the rating, that is the majority class, and then to identify any ratings that deviated from the trend. For example, if one annotator rated a topic as expressing a positive sentiment (ranking *+1, +2*, or *+3*) whereas the large majority of the annotators rated this topic as expressing a negative sentiment (ranking *-1, -2*, or *-3*) then this rating was labeled as *deviated*. If the annotator had at least two deviated ratings on a specific date, then his/her contributions on the specific date were removed. Here, we want to notice that there were cases where one annotator submitted ratings for more than one spikes. If the annotator had two or more deviated ratings for only one spike, we removed only the specific contributions and not all his/her contributions.

One interesting case with many deviated ratings was one annotator of the entity *Angela Merkel*. The specific annotator used a high percentage of the score *-3* whereas the rest of the annotators were more conservative for most of those topics. For example, for the spike on the 16th of August, 70% percent of his ratings were equal to *-3*, 20% percent of his ratings were equal to *-2*, and 10% to *-1*. Trying to explain this weird behavior, we looked at the demographic data and realised that the annotator was Greek and that most of the topics that he/she

rated with *-3* were about Greece (e.g., topics about Greek crisis, greek referendum, greek debt). Even if this annotator probably was not a spammer, we considered that his/her replies as biased for the specific topics and we removed his/her contributions.

After we removed biased and low-quality contributions we ended up with the following annotations per entity; for *Michelle Obama* we had 30 different annotators that submitted 470 evaluations for the 50 sets of tweets with an average of 15.66 sets per annotator; for *Angela Merkel* we had 22 different annotators and 420 total annotations with an average of 19.1 sets per annotator; for *Angelina Jolie* we had 16 different annotators and 500 total annotations with an average of 31.25 sets per annotator.

4.4.4 Analysis of the Collection

As previously mentioned, for each of the entities we focused on five different sentiment spikes. To understand the causes of a sentiment spike, we looked through the tweets that were published not only on the specific day but also a bit before. Since we considered 5 different spikes we had 5 different time windows per entity, and we extracted the topics discussed in all of them using LDA.

Table 4.2. Sample of extracted topics from different sentiment spikes.

Michelle Obama	
07 May 2015	ruined lunch cookies college signing
31 May 2015	grieving tonight bidens beau death
16 July 2015	mayor gorilla resign face racist
12 Aug. 2015	stand years feminist attacks miss
08 Nov. 2015	kids drag fam roast mom
Merkel	
23 May 2015	scandal political crisis spy germany
03 July 2015	debt unsustainable greek wikileaks phone
29 July 2015	girl palestinian cry caused abolish
16 Aug. 2015	bigger challenge migrants crisis european
29 Nov. 2015	syria military french downing aircraft
Angelina Jolie	
19 Apr. 2015	life structure cheekbones appreciation amazing
04 June 2015	birthday happy beautiful women inspirational
21 July 2015	blood imagine veins donate celebrities
26 Sep. 2015	awards academy flaws bones kardashians
30 Sep. 2015	pitt brad smith movie amazing

Table 4.2 shows one of the topics that was extracted from the five different sentiment spikes of each entity. We could observe that LDA managed to group terms that were about the same topic together. Some of those topics are related to important news (e.g., the topic detected for *Angela Merkel* on the 3rd of July that is about *German chancellor admitting in a 2011 phone call that Greek debt is unsustainable*) whereas other are less important events (e.g., the topic on the 4th of June that is about *wishing happy birthday* to *Angelina Jolie*). This is due to the informal style of Twitter in which users frequently retweet a message that does not refer to an important event but the users may find it interesting or funny.

In total we collected 1,390 relevance assessments. Figure 4.3 shows the average inter-annotator agreement for each sentiment spike and each entity. We considered two different settings to calculate the inter-annotator agreement. In the first setting (*Setting_1*) we considered all the possible ratings (*-3, -2, -1, 0, 1, 2, 3*) such that two annotators agree if and only if they gave the exact same rating. In the second setting (*Setting_2*) we considered three different classes (*positive, neutral, negative*). In this case we considered that two annotators agree if both of them have given a positive (*1, 2, 3*), a neutral (*0*) or a negative (*-1, -2, -3*) rating. As can be observed, the percentage of agreement increased when we considered only three classes. In addition, we observe that there is higher agreement for the entity *Angelina Jolie* in most of the spikes compared to *Michelle Obama* and *Angela Merkel*. We believe that one reason is that positive sentiment is easier to understand compared to negative and therefore is more likely the annotators to give similar ratings.

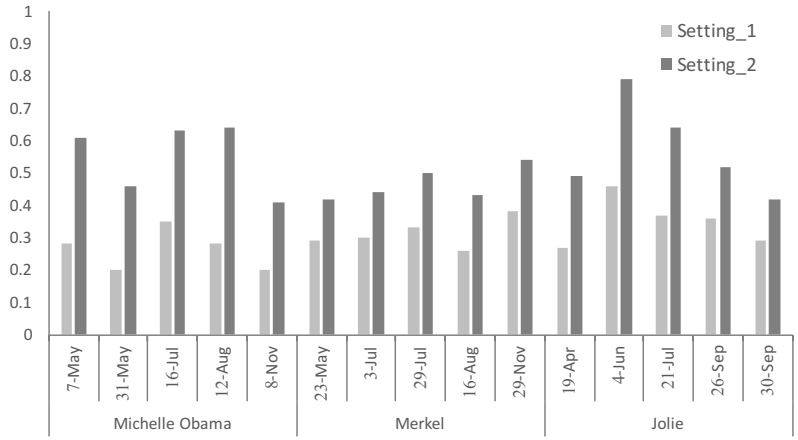

Figure 4.3. Average inter-annotator agreement per sentiment spike and entity.

Figure 4.4 shows the overall distribution of the relevance assessments for all

the extracted topics per sentiment spike and entity. We observe that the majority of the assessments given for the topics about *Michelle Obama* and *Angela Merkel* were rated as negative whereas the majority of the assessments about *Angelina Jolie* were positive. One likely reason for this is that users tend to express positive opinion when they post about celebrities usually showing their admiration whereas they tend to be more critical on persons or topics related to politics.

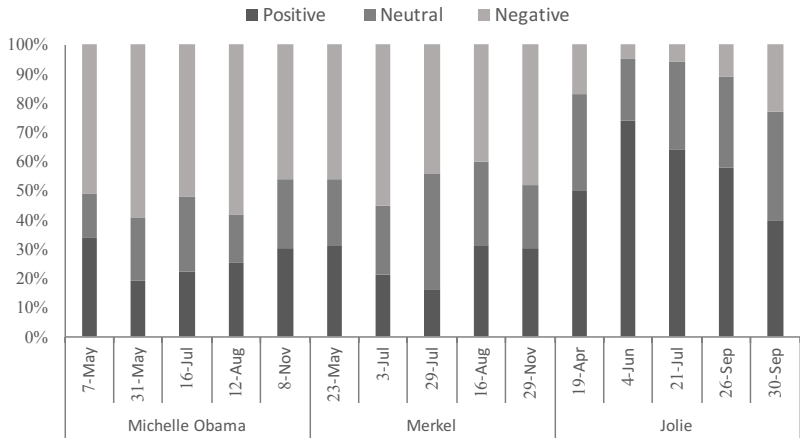

Figure 4.4. Distribution of relevance assessments in reference to the sentiment polarity class per spike and entity.

Finally, after additional analysis, we found that some topics have a high standard deviation. For example, the topic with the highest standard deviation is *Topic 1* of *Michelle Obama* on *May 31, 2015* for which the collected ratings have a standard deviation of 1.856, whereas the average standard deviation on the specific spike was 1.347. This topic contains tweets about the death of Beau Biden (i.e., *Michelle and I are grieving tonight. Beau Biden was a friend of ours.*). This topic is very subjective and some annotators considered that if one is grieving for having lost a person, this can be seen as a positive sentiment towards the person who passed away, whereas others considered it as expressing negative sentiment.

4.5 Results and Discussion

In this section, we describe the results of our study. We begin by presenting our results on sentiment tracking and sentiment spike detection. Next, we present

our results on extracting topics from the sentiment spikes. Finally, we present and discuss our results on ranking the likely causes of sentiment spikes.

4.5.1 Sentiment Tracking

We examine different cases and try to understand if the time series tools are useful for sentiment tracking. For the sentiment tracking we focus on five different entities: *android lollipop, pretty little liars, michelle obama, angela merkel* amd *angelina jolie*. The first step is to plot the data and examine if there are any patterns. Figure 4.5 shows the number of total, positive and negative tweets published every day for the entity under examination. From this figures, we can observe if there are any specific patterns on the data. For example, from Figure 4.5a that shows the frequencies about *android lollipop* we observe that the popularity of the entity is decreasing. However, the positive and negative tweets do not follow the same trend. In other cases, the popularity of the entities is consistent to one of the sentiment. For example, in Figure 4.5e we observe that there are several peaks in the number of positive tweets that are about *angelina jolie* that occur when there is an increase in the number of tweets that are about *angelina jolie*. These peaks may be related to some important events or to seasonal effects.

To have better understanding of the data we isolate the different components. Figure 4.6 shows the decomposition of the various entities into trend, seasonal and noise. First, we observe that all the five entities have seasonality. Also, we observe from Figure 4.6a that the trend of the entity *android lollipop* is decreasing. Figure 4.6b shows that the entity *pretty little liars* has seasonality and this is probably related to the weekly episodes. Patterns regarding *michelle obama*, *merkel* and *angelina jolie* are more difficult to be identified since they are influenced mainly from external events or news.

Some important information is also reflected with sentiment velocity and acceleration that help us to understand how quickly an entity is gaining or losing preference. Figure 4.7 shows the positive and negative velocity and acceleration of the entities that we examine. For example, Figure 4.7d shows that the negative sentiment towards *angela merkel* grows faster in the middle of July whereas in August the positive sentiment has greater acceleration that lasts only few days. In addition, we observe from Figure 4.7e that the positive sentiment towards *angelina jolie* grows in general faster compared to the negative sentiment.

Figure 4.8 shows the number of total, positive and negative tweets published every day for the entities together with the identified peaks. From this figure we can observe the time when there was a sudden change in entity's popularity or

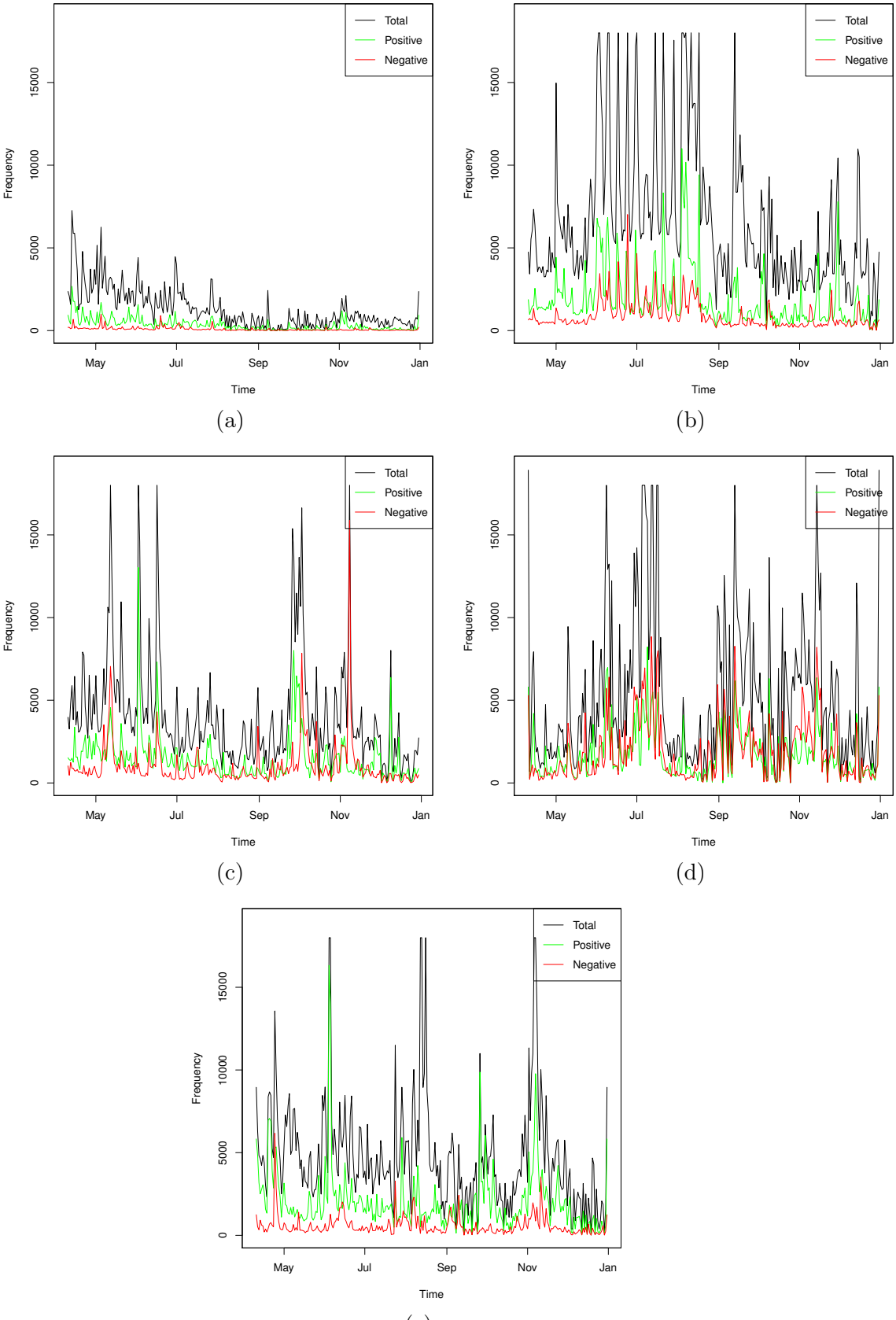

Figure 4.5. Number of total, positive and negative tweets of (a) android lollipop, (b) pretty little liars, (c) michelle obama, (d) angela merkel and (e) angelina jolie entity per day.

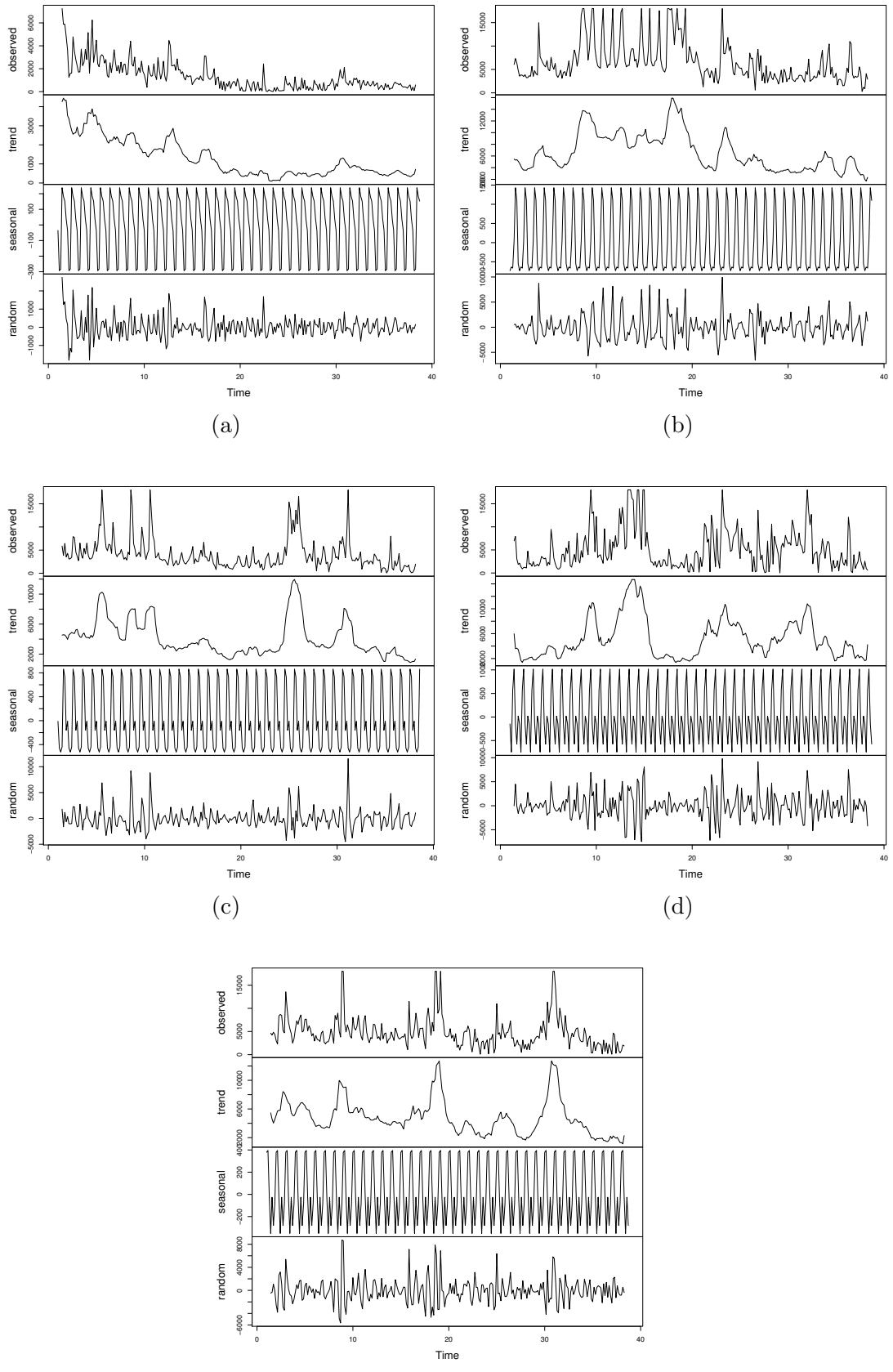

Figure 4.6. Decomposition plots of (a) android lollipop, (b) pretty little liars, (c) michelle obama, (d) angela merkel and (e) angelina jolie entity per day.

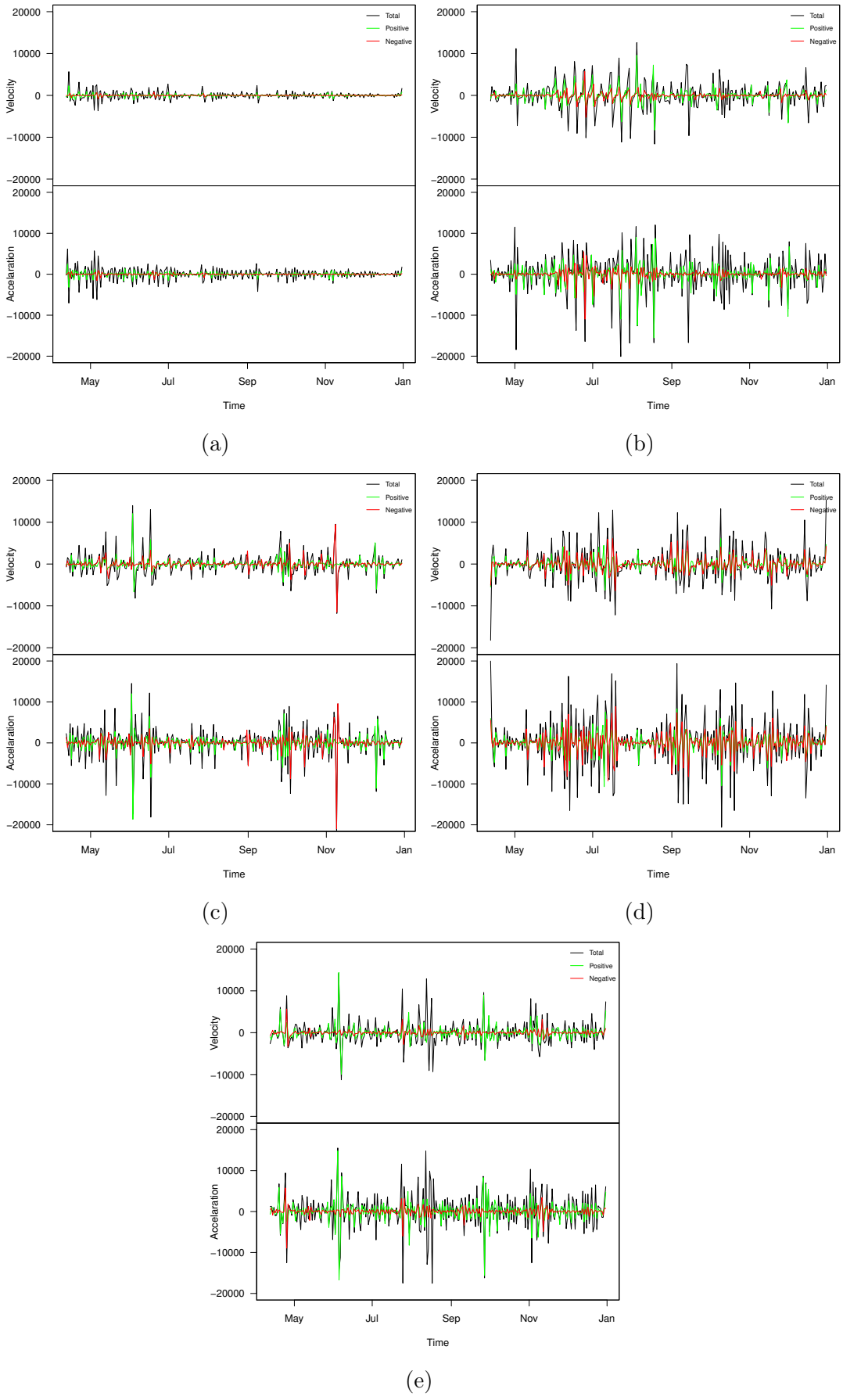

Figure 4.7. Velocity and acceleration plots of (a) android lollipop, (b) pretty little liars, (c) michelle obama, (d) angela merkel and (e) angelina jolie entity per day.

in sentiment towards the specific entity. Knowing the time of a potential peak is useful in identifying the events that caused those peaks. Apart from that, we observe that the peaks in entity's popularity, positive and negative peaks occur at different time periods. This implies that a sudden peak in an entity's popularity does not mean that there will be emerging sentiment. However, there are many cases that a peak in entity's popularity is followed by emerging sentiment. This occurs a lot in the entity *angelina jolie* shown in Figure 4.8e.

Next, we focus on predicting the sentiment using state-of-the-art time series tools. We compare naïve, mean and ARIMA models. Figures 4.9 and 4.10 show the predictions regarding the five examined entities of positive and negative sentiment respectively. The predictions are performed for 15 days, from the 16th to 31st of December 2015. From the figures, we observe that the different approaches perform in a different way across the entities.

In order to understand how the different forecast approaches perform, we show the results in terms of Mean Absolute Error (MAE). Calculation of MAE is relatively simple and is based on summing the magnitudes of errors and then dividing the total error with the number of instances. The MAE is measured as:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

where n is the number of classified instances, x_i is the actual label of instance i and y_i is the predicted label for instance i.

Table 4.3 shows the predictions regarding the positive sentiment in terms of Mean Absolute Error (MAE). From the results we observe that the least errors are obtained for the *android lollipop* entity. Also, we observe that ARIMA outperforms the mean and naïve approaches for the *pretty little liars* entity. However, for the rest of the entities, the naïve approach manages to outperform both mean and ARIMA. This is an interesting result that shows that very simple forecast approaches can perform better than other more sophisticated approaches.

Similar observations can be made regarding the negative sentiment prediction. Table 4.4 shows that ARIMA outperforms mean and naïve regarding *android lollipop* and *pretty little liars* entities. However, for the rest of the entities the naïve approach manages to do the predictions with the least errors. One explanation is that regarding the entities *michelle obama*, *angela merkel* and *angelina jolie*, we have data that are more challenging to be predicted using the past values compared to *android lollipop* and *pretty little liars* that have more clear seasonality patterns.

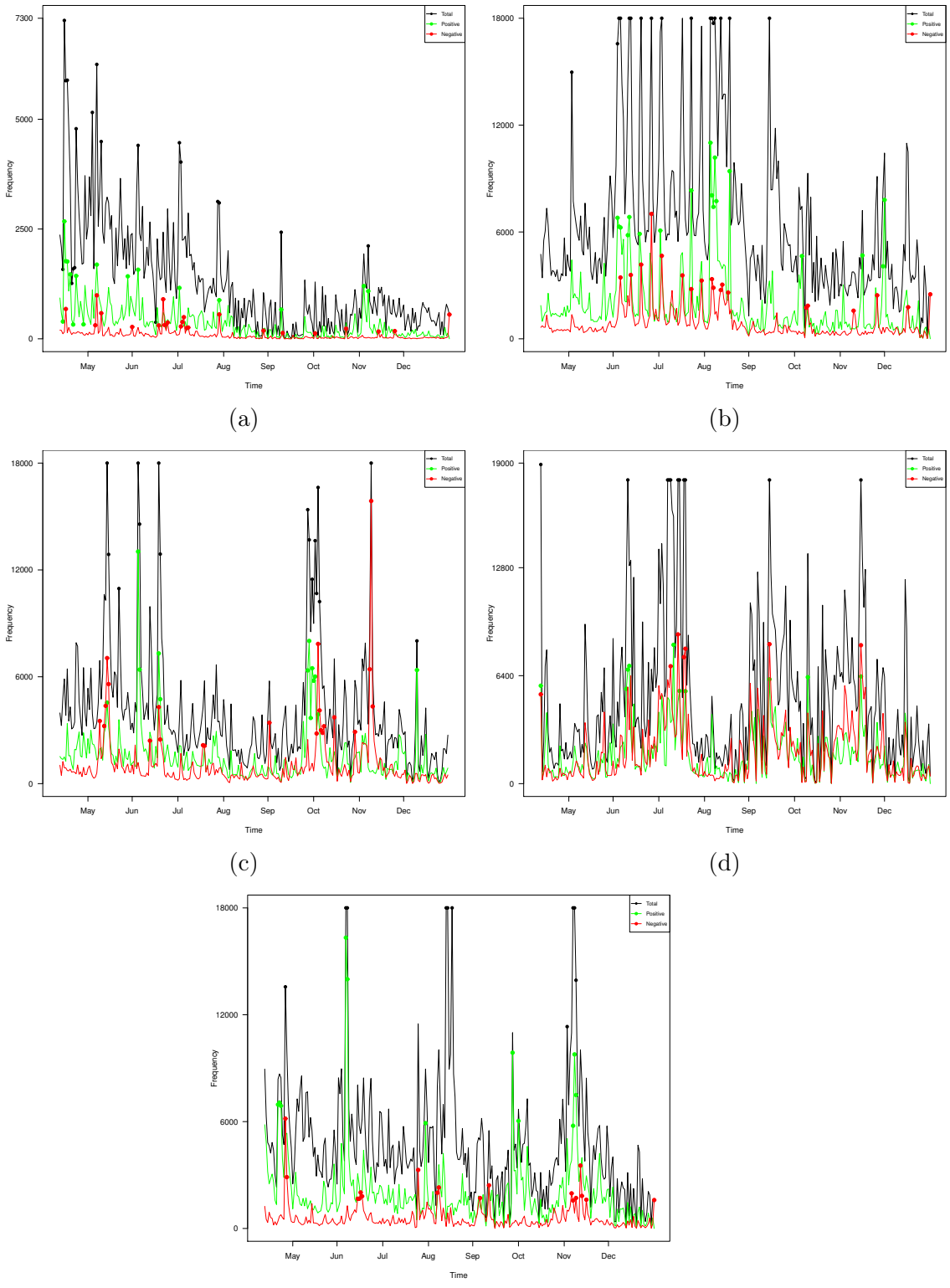

Figure 4.8. Outliers of (a) android lollipop, (b) pretty little liars, (c) michelle obama, (d) angela merkel and (e) angelina jolie entity per day.

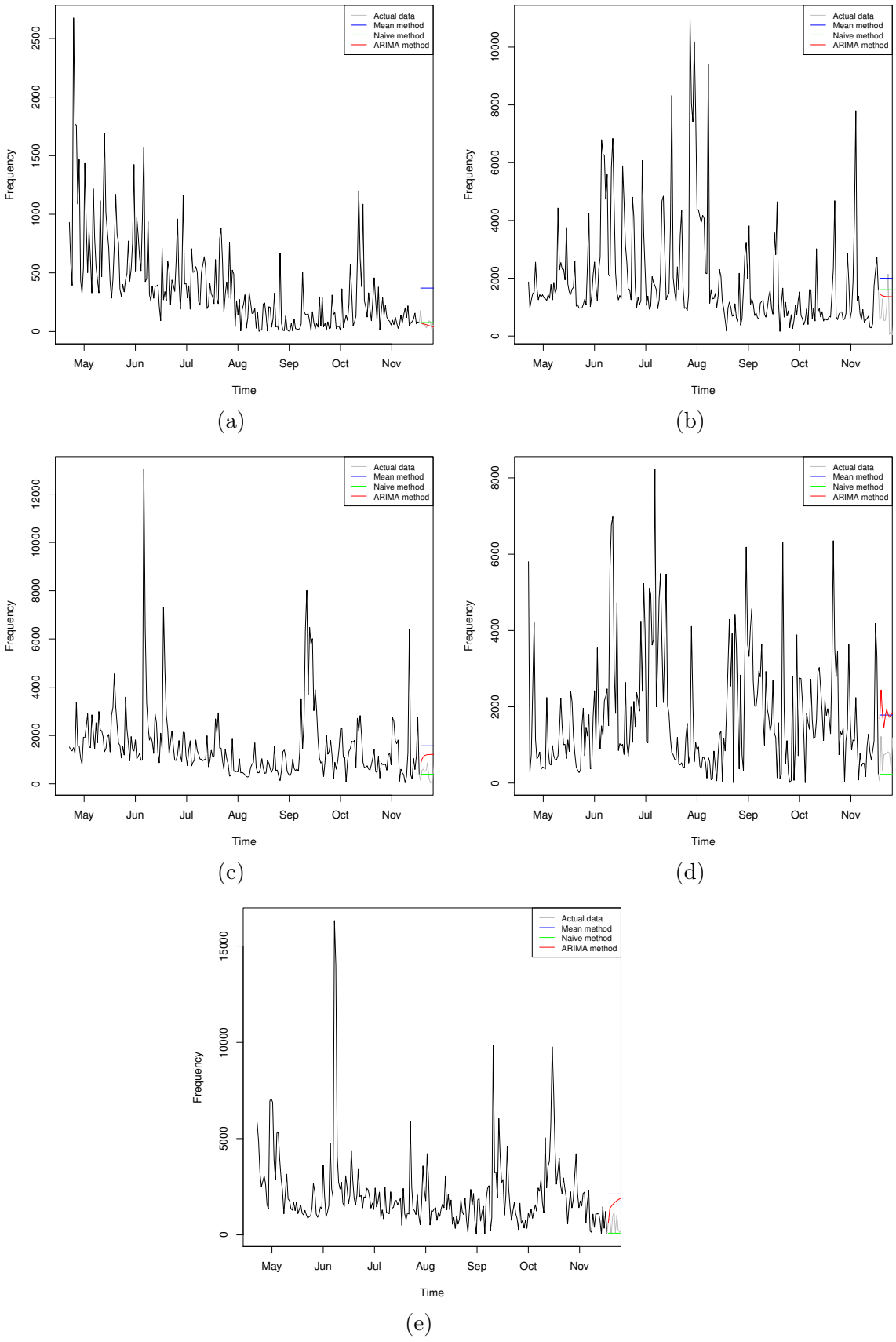

Figure 4.9. Predictions of the positive sentiment of (a) android lollipop, (b) pretty little liars, (c) michelle obama, (d) angela merkel and (e) angelina jolie entity per day.

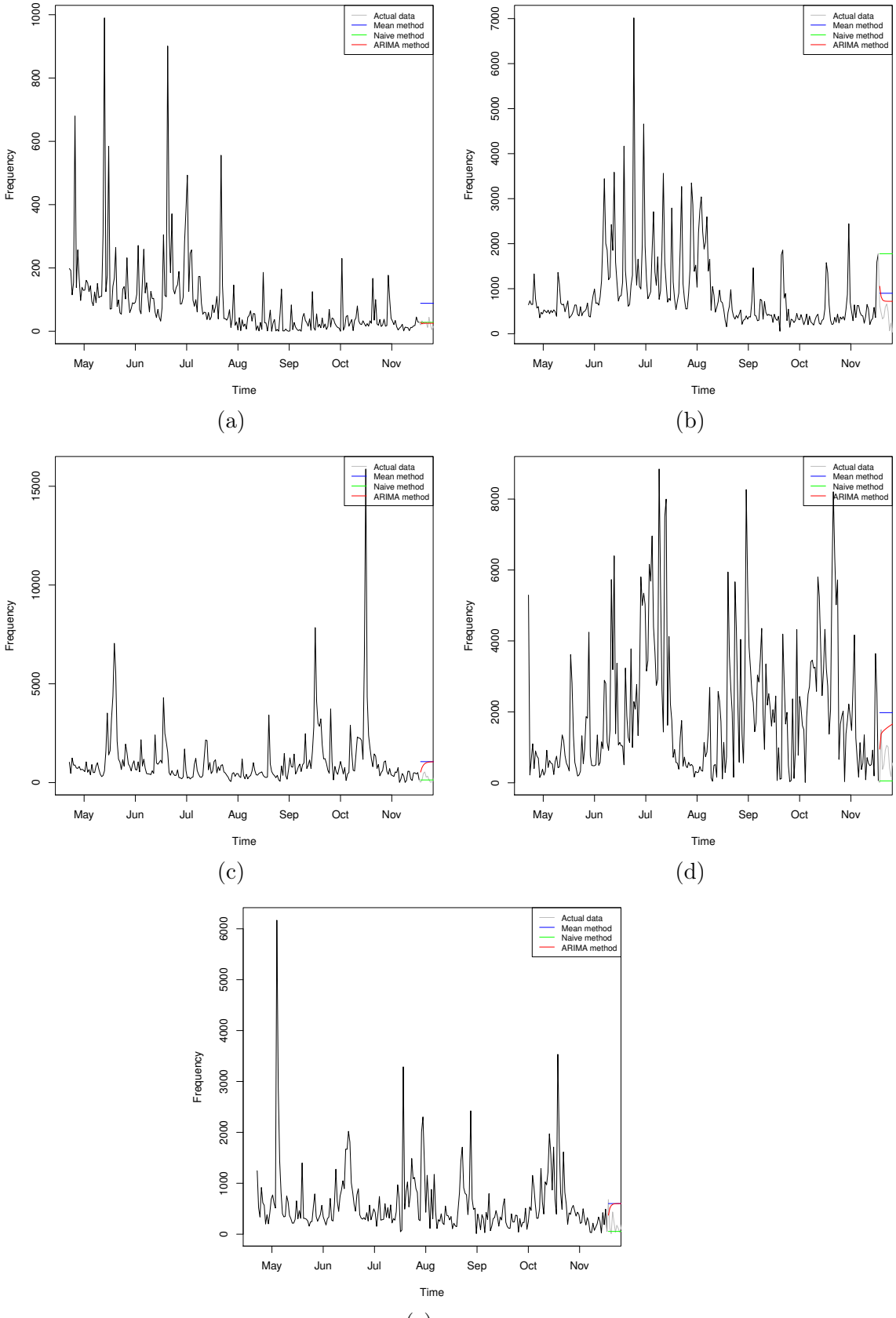

Figure 4.10. Predictions of the negative sentiment of (a) android lollipop, (b) pretty little liars, (c) michelle obama, (d) angela merkel and (e) angelina jolie entity per day.

Table 4.3. Performance results (MAE) for the positive sentiment prediction.

	android lollipop	pretty little liars	michelle obama	angela merkel	angelina jolie
Mean	288.05	1391.09	1073.73	1108.79	1579.30
Naïve	46.64	1051.85	257.42	482.35	465.28
ARIMA	48.78	863.13	662.85	1138.64	1164.98

Table 4.4. Performance results (MAE) for the negative sentiment prediction.

	android lollipop	pretty little liars	michelle obama	angela merkel	angelina jolie
Mean	64.73	529.87	797.48	1323.06	391.52
Naïve	10.21	1408.14	185.21	607.78	184.07
ARIMA	9.99	385.83	705.09	901.97	396.37

4.5.2 Topic Classification on Sentiment Spikes

To create the collection with the likely causes of sentiment spikes, first we identified several sentiment spikes regarding the three following entities: *Michelle Obama*, *Angela Merkel* and *Angelina Jolie*. As already mentioned, some of the most important statistics that are related to the sentiment spikes we analysed in this study can be found at Table 4.1.

To understand the causes of a sentiment spike, we need to look through the tweets that were published not only on the specific day but also a bit before. As a starting point, we considered the day when the number of tweets with the same sentiment had started increasing, that is, the most recent day before the sentiment spike which presented the lowest number of tweets with the same sentiment. For each entity we took into account 5 different spikes. Hence, we have 5 different time windows per entity, and we want to extract the topics discussed in all of them.

In order to extract the topics, we applied LDA on each of the sentiment spike. As mentioned in Section 4.4.4, Table 4.2 shows one of the topics that was extracted on the five different sentiment spikes and for each of the three entities. We observe that LDA managed to group terms that were about the same topic together. Also, we observe that some of those topics are related to important news (i.e., the topic detected for *Angela Merkel* on the 23rd of May that is about a political scandal) whereas other are about less important events (i.e., the topic on the 4th of June that is about wishing happy birthday to *Angelina Jolie*). This is a consequence of the informal style of Twitter in which users frequently retweet

Table 4.5. Performance results for the entity Michelle Obama. A tick mark (✓) means that there is a correlation whereas a x mark (✗) means that there is no correlation compared to the ground truth.

	Michelle Obama					
	Baseline					
	Spearman	Pearson	MAP@5	MAP	P@3	P@5
07 May 2015	✗	✗	0.353	0.574	0.667	0.600
31 May 2015	✗	✗	0.040	0.410	0.000	0.200
16 July 2015	✗	✗	0.040	0.418	0.000	0.200
12 Aug. 2015	✗	✗	0.170	0.487	0.000	0.600
08 Nov. 2015	✗	✗	0.237	0.504	0.333	0.600
	LDA & KL-divergence					
07 May 2015	✓	✓	0.493	0.692	0.667	0.800
31 May 2015	✗	✗	0.247	0.521	0.333	0.400
16 July 2015	✓	✓	0.793	0.815	0.667	0.800
12 Aug. 2015	✗	✗	0.180	0.475	0.000	0.400
08 Nov. 2015	✗	✗	0.337	0.597	0.333	0.600

a message that does not refer to an important event but the users may find the topic interesting or funny.

4.5.3 Extracting Likely Causes of Sentiment Spikes

Next, we present the results on the task of ranking the causes of sentiment spikes. The ground truth in terms of ranking is based on the evaluations of annotators collected via crowdsourcing. Table 4.5 shows the results for the entity *Michelle Obama* on the task of ranking the causes of sentiment spikes using the baseline and the LDA & KL-divergence approach. From this table, we observe that on some days the LDA & KL-divergence approach managed to obtain a similar ranking compared to the one generated by human annotators. Also, we observe that the LDA & KL-divergence approach performs better on all the sentiment spikes compared to baseline.

We analysed the results on the sentiment spikes for which the LDA & KL-divergence method returned a ranking similar to the one given by the ground truth. Table 4.6 shows the three topics that were ranked as the most negative by human annotators for the sentiment spikes observed on *May 7* and *July 16, 2015*. As we can see from the lists of the keywords, the tweets belonging to those topics express negative sentiment. To get a better understanding, we manually

Table 4.6. The three most negative topics based on human judgments for two sentiment spikes (07 May 2015 and 16 July 2015) related to the entity Michelle Obama.

07 May 2015
Topic 4: school museums lunches ruined people
Topic 6 : trip ski cost taxpayers aspen daughters
Topic 3 : don't hey feel care barack wrong anymore
16 July 2015
Topic 7: mayor gorilla resign face racist won't admitting
Topic 9: sailors fried food ban navy blame lady
Topic 8: calls monkey gorilla mayor face barack man

checked the set of tweets that belong to the topics to examine if they truly express a negative sentiment. One example is *Topic 4: school, museums, lunches, ruined, people* that was ranked as the most negative according to human annotators for the spike on *May 7*. This topic contains a great amount of tweets saying: *Michelle Obama ruined our lunch*. Looking in the web, it is easy to find out that this is related to the fact that Michele Obama changed schools' lunch program, but the students did not like the change and so they started complaining on Twitter. Note that this topic was ranked as the second most negative by the LDA & KL-divergence approach.

Another interesting case is *Topic 7: mayor, gorilla, resign, face, racist, wont, admitting* that was ranked as the most negative by human annotators and by the LDA & KL-divergence approach for the spike *July 16, 2015*. This topic contains several tweets saying *I thought racism ended? Mayor won't resign over 'Gorilla face' Michelle Obama rant b/c 'that's admitting I'm racist'* or tweets with a similar content. This topic is related to the fact that the mayor of a city in Washington state posted on Facebook a comment on which he compared President Barack Obama's family to gorillas and monkeys and as a consequence people started commenting on this event and his refusal to resign.

However, we observed that for some other sentiment spikes the ranking generated from the LDA & KL-divergence approach is different from the ranking collected from human annotators. More specifically, on the spikes occurred on *May 31, Aug. 12,* and *Nov. 08, 2015* there is no correlation between the two rankings. In order to get a deeper understanding on the reasons for this disagreement, we manually looked and compared the results on these spikes.

First, we analysed the sentiment spike on *May 31, 2015*. To this end, we measured the absolute difference between the rankings per each topic. From

the results we observed that the biggest difference is on the *Topic 4* whereas the second largest is for the *Topic 1* and *Topic 2*. More specifically, our method ranked *Topic 4* as the second most positive topic, whereas annotators considered that it is the second most negative topic. Looking through the set of tweets that belong to this topic, we could see that the tweets were about a picture that was edited by photoshop and the majority of the tweets were ironic towards *Michelle Obama*. For example, most of them were retweets of the message *So I found this picture of Michelle Obama playing tennis*, and it is obvious that the picture is not real. Irony, that is a way to express the opposite of the literal meaning, is very difficult to be detected automatically [120]. There are a few studies focusing on irony detection on Twitter [130], but we believe that this problem is still under-explored. As we could expect, annotators managed to detect irony whereas our approach could not.

Table 4.7. Performance results for the entity Angela Merkel. A tick mark (✓) means that there is a correlation whereas a x mark (✗) means that there is no correlation compared to the ground truth.

	Angela Merkel					
	Baseline					
	Spearman	Pearson	MAP@5	MAP	P@3	P@5
23 May 2015	✗	✗	0.287	0.529	0.333	0.600
03 July 2015	✗	✗	0.130	0.467	0.000	0.400
29 July 2015	✗	✗	0.337	0.583	0.333	0.600
16 Aug. 2015	✗	✗	0.130	0.463	0.000	0.400
29 Nov. 2015	✗	✗	0.327	0.578	0.333	0.800
	LDA & KL-divergence					
23 May 2015	✗	✗	0.537	0.670	0.333	0.600
03 July 2015	✗	✗	0.387	0.564	0.333	0.600
29 July 2015	✗	✗	0.287	0.583	0.333	0.600
16 Aug. 2015	✗	✓	0.327	0.620	0.333	0.800
29 Nov. 2015	✗	✗	0.220	0.526	0.000	0.600

We have also noticed a difference in *Topic 1* and *Topic 2* that were ranked as negative by the LDA & KL-divergence approach whereas annotators did not consider that they express negative sentiment. The tweets that belong to these topics were about the death of Beau Biden (i.e., *Michelle and I are grieving tonight. Beau Biden was a friend of ours.*). As already mentioned in Section 4.4.2 the annotators were asked to take also into account negative emotions, but they did not consider the content of such tweets as negative, since if one is grieving for

Table 4.8. Performance results for the entity Angela Merkel after considering tweets that refer to news. A tick mark (✓) means that there is a correlation whereas a x mark (✗) means that there is no correlation compared to the ground truth.

Angela Merkel						
LDA & KL-divergence						
	Sp.	Pear.	MAP@5	MAP	P@3	P@5
23 May 2015	✗	✗	0.587	0.695	0.333	0.600
03 July 2015	✗	✗	0.287	0.541	0.333	0.600
29 July 2015	✓	✓	0.743	0.797	0.666	0.800
16 Aug. 2015	✓	✓	0.393	0.671	0.667	0.800
29 Nov. 2015	✓	✗	0.443	0.667	0.667	0.800

Table 4.9. Channels used to detect the tweets that refer to news.

abcnews
ajenglish
ap
bbc
breaking
channel4news
cnn
nbcnews
reuters
rt_com
skynews
telegraph

the loss a person, this can be seen as a positive sentiment towards the person who passed away. In general, the interpretation of emotions is very personal and sometimes it is difficult to obtain a full agreement. This is more intense in topics that are controversial as for example the two topics that are about the death of Beau Biden. The fact that *Topic 1* and *Topic 2* are controversial is also evident from the fact that the collected ratings have a standard deviation of 1.856 and 1.802, respectively, whereas the average standard deviation on the specific spike was 1.347.

Table 4.7 shows the results for the entity *Angela Merkel*. From the results we observe that for most of the days the rankings generated from the LDA & KL-divergence approach are very different compared to the rankings given from

the human annotators. In case of *Angela Merkel* we observed that some of the topics can attract biased contributions. One example is the sentiment spike on *July 3rd, 2015*. In this case, the majority of topics were about Greece: the Greek debt, rumors for grexit and the Greek referendum that took place on the 5th of July 2015. This case is a typical example that highlights the risk of attracting biased annotations due to the subjectivity of the task.

In an effort to understand the disagreement of our approach compared to human annotators we observed that in the entity *Angela Merkel* there are a lot of tweets that are retweets of news and which do not express any sentiment. For example, the tweet *BREAKING: German Chancellor Angela Merkel briefly collapsed while attending an event in Bayreuth - local media* is considered by SentiStrength as negative, probably due to the word *collapsed*. However, the human annotators considered this tweet as neutral, probably due to the fact that it is an event reported in the breaking news.

In order to have a better understanding on the effect of the false polarity detection of the tweets with news, we generated a new ranking of topics. The new ranking considered that tweets containing news are neutral. To do so, we used a list of news channels and keywords that are usually used to post news in Twitter. Table 4.9 shows the list of keywords used to detect the tweets that refer to news. Tweets that were posted from these channels, and their plain retweets were considered as neutral to produce the new ranking. Table 4.8 shows the results for the entity *Angela Merkel* after considering tweets that express news. We observe that in most of the sentiment spikes, there is a big improvement. More importantly, the new ranking seems to be similar compared to the human annotators for three sentiment spikes; *July 29, Aug. 16,* and *Nov. 29, 2015*.

Finally, Table 4.10 shows the results for the entity *Angelina Jolie*. In this case, we observe that the LDA & KL-divergence approach was effective for most of the spikes, and there is an agreement compared to the human judgments. Also, in most of the cases, the LDA & KL-divergence approach managed to perform better compared to the baseline. We believe that one reason for this is that maybe positive sentiment is clearly expressed and therefore it is easier to understand. Also, it is more likely that humans will agree on a positive sentiment compared to a negative. One example is the sentiment spike on the *June 4, 2015* related to *Angelina Jolie*. In this case, there are three topics that all the annotators considered positive and in particular the 80% of the annotators considered them very positive (score of *+3*). We believe that topics with positive sentiment are less controversial and therefore it is easier to be ranked with an automatic approach.

Table 4.10. Performance results for the entity Angelina Jolie. A tick mark (✓) means that there is a correlation whereas a x mark (✗) means that there is no correlation compared to the ground truth.

	Angelina Jolie					
	Baseline					
	Spearman	Pearson	MAP@5	MAP	P@3	P@5
19 Apr. 2015	✗	✗	0.693	0.763	0.667	0.800
04 June 2015	✗	✗	0.413	0.648	0.667	0.400
21 July 2015	✗	✗	0.197	0.507	0.333	0.400
26 Sep. 2015	✗	✗	0.503	0.668	0.667	0.600
30 Sep. 2015	✗	✗	0.703	0.759	0.667	0.600
	LDA & KL-divergence					
19 Apr. 2015	✓	✓	0.743	0.828	0.667	0.800
04 June 2015	✓	✓	0.643	0.805	0.667	0.800
21 July 2015	✗	✗	0.080	0.454	0.667	0.400
26 Sep. 2015	✓	✓	0.593	0.756	0.667	0.800
30 Sep. 2015	✓	✗	0.753	0.817	0.667	0.600

4.6 Conclusions

In this chapter, we analysed a large collection of tweets in order to get a better understanding of sentiment evolution and causes of sentiment spikes. We focused on five different entities: *android lollipop*, *pretty little liars*, *Michelle Obama*, *Angela Merkel* and *Angelina Jolie*. We applied state-of-the-art time series tools to extract patterns and forecast sentiment. In addition, we applied an outlier detection approach to identify the sentiment spikes that emerged within a period of seven months. More importantly, we proposed the LDA & KL-divergence approach that combines LDA and Relative Entropy to extract the topics that are discussed on the sentiment spikes and to rank those topics according to their contribution on the sentiment spike. To evaluate the LDA & KL-divergence approach, we built a collection of tweets which are grouped by topics and labelled based on the polarity and strength of sentiment they express. For the collection, we focused on three entities, *Michelle Obama*, *Angela Merkel* and *Angelina Jolie*. We used crowdsourcing to collect the ground truth.

Our results showed that the state-of-the-art time series approaches are very useful in tracking sentiment over time. Frequency analysis was useful to observe how positive and negative sentiment evolved over time. Also, the decomposition was useful to observe the entities that follow trends and their seasonality. In

addition, we compared the effectiveness of naïve, mean and ARIMA forecasting tools and we showed that in some cases the naïve that is a simple forecasting approach can outperform ARIMA that is a more sophisticated approach. Next, we presented a process to build a collection that can be used to extract the likely causes of sentiment spikes and we discussed several challenges we addressed to build the collection. Finally, we showed that LDA & KL-divergence approach can effectively be used to extract and rank the topics identified on sentiment spikes.

Chapter 5

Emotional Reactions Prediction

This chapter focuses on predicting users' emotional reactions that are triggered by online news articles. More specifically, we propose different pre-publication information such as similarity, reactions entropy and semantics. Next, we propose two different groups of features extracted from users' comments. These features capture the commenting activity (e.g., when the first comment is published) and the content of the comments (e.g., relevance to the post). In addition, we combine the features extracted from comments published shortly after the post with the post's terms to explore the effectiveness of this combination on predicting the ordinal level of five standard emotional reactions (*love, surprise, joy, sadness, anger*). Finally, we analyse and discuss the contribution of terms and of the proposed commenting features in estimating the number of reactions.

Our results show that the terms of the post is the most important feature for the pre-publication prediction. Also, we show that features extracted from users' comments are very important for the emotional prediction task. More importantly, we show that the commenting features contain more predictive power compared to terms for all the emotions except for *sadness*. Finally, our results suggest that the most effective models to predict which posts trigger a high number of emotional reactions are those trained on both posts' terms and users' comments.

5.1 Introduction

In recent years, social networks have become an integral part of the news industry. News agents post news articles on social networks, such as Facebook and Twitter. These news articles are accessible to users who can comment or express their opinion about them. Some of the news articles posted on social networks

trigger a large number of emotional reactions (e.g., sadness, surprise) whereas others do not. Predicting the number of emotional reactions that will be triggered on users is very important for a variety of different tasks, such as information spreading and fake news detection. For example, fake news are written to attract users' attention and trigger emotions to a large number of people [138]. Therefore, the number of emotional reactions can be used as an additional information for fake news or clickbait detection.

Emotional reactions prediction is a challenging problem. The structure of the network or other external factors such as users' location are some of the factors that can affect the number of the triggered emotional reactions. Intuitively, the content of the news post is one of the most important factors that influences the emotional reactions that will be triggered [6]. However, content is not sufficient alone since there are other factors that may influence the number of triggered reactions. Semantic features, such as entities and concepts as well as early commenting features (i.e., features extracted from comments published within the first ten minutes after the publication of the news post) can be very useful for an effective prediction.

The problem of emotional reactions prediction is related to online content popularity prediction. Most prior work on news articles' popularity prediction is based on early-stage measurements [4, 145], whereas little effort has been made on the pre-publication prediction scenario [20, 15]. Although the problem of emotional reactions prediction has apparent similarities with predicting the popularity of news, the two problems are not the same. A piece of news that triggers massive emotional reactions has certainly higher probabilities of receiving attention compared to news articles that do not trigger any. However, predicting the number of the triggered emotional reactions depends on many factors such as, for example, the affective words that the news post contains, the structure of the network and the early commenting activity.

A related work was presented by Clos et al. [35] who proposed a unigram mixture model to create an emotional lexicon that was then used to predict the probabilities of five emotions. In addition, Alam et al. [6] focused on mood level prediction of news articles' readers (ranging from 0 to 1) using features such as character, words and affect scores and showed that n-grams and stylometric features were the most important. More recently, Goel et al. [65] focused on predicting the intensity of emotions in tweets using an ensemble of three neural-network approaches. However, our problem is not the same as that of predicting emotional intensity, since an article may trigger an emotion that is intense to only a small amount of people. For example, consider the case of a strike in the means of transportation in a small city. In such a case, some people may feel

very angry (e.g., *I got stuck in traffic for an hour and a half! #busStrike*) but such intense emotion might be triggered only in a small amount of people.

In this chapter, we focus on the emotional reactions that online news articles trigger on users, and we attempt to predict the ordinal level (e.g., high, medium, low number) of emotional reactions that will be triggered once a news post is published. We focus on five standard emotional reactions (*love, surprise, joy, sadness, anger*) triggered by posts published by New York Times Facebook page[1]. The decision about using Facebook is that it allows users to select one of these five emotional reactions with regards to a post which can be used as the ground truth. Figure 5.1 shows an example of a Facebook news post published by New York Times with all its different features, such as comments and reactions it has triggered.

To address the problem, first, we propose different pre-publication information such as similarity, reactions entropy and semantics and we measure their effectiveness. In addition, we propose two different groups of features extracted from users' comments. These features capture two different aspects of information: the commenting activity (e.g., when the first comment is published) and the content of the comments (e.g., relevance to the post). Next, we propose combining textual and features extracted from users' comments to effectively predict the triggered emotional reactions. Finally, we analyse and discuss the contribution of terms and of the proposed commenting features in estimating the number of reactions on each of the emotional reactions.

In this chapter, we address the following general research question:

RQ3 How can we predict how many people will react with a specific emotion when a news post is published?

The general research question can be further analysed to the following more specific research questions:

RQ3.1 Can we improve the effectiveness of baseline classifiers by adding additional pre-publication information based on news post content?

RQ3.2 Can we improve the effectiveness of baseline classifiers by adding additional post-publication information extracted from users' comments?

RQ3.3 How does a model that combines textual and early commenting features perform?

[1]https://www.facebook.com/nytimes/

Figure 5.1. Example of a Facebook post published by New York Times.

RQ3.4 What is the added value of the commenting features in terms of effectiveness in the task of emotional reactions prediction?

We proceed with a definition of the emotional reactions prediction task in Section 5.2. Section 5.3 introduces the pre- and post-publication features which are proposed for the prediction. We present our experimental setup in Section 5.4 and our results and analysis in Section 5.5. Finally, the conclusions are presented in Section 5.6.

5.2 Task Definition

In this chapter, we focus on the problem of emotional reactions prediction of news posts published on a social network. The problem is defined for two different settings: pre-publication and post-publication prediction. Regarding the pre-publication prediction the problem can be stated as: *Given a news article post, predict the qualitative ordinal level of emotional reactions that the post will trigger*. Regarding the post-publication prediction the problem can be stated as: *Given a news article post and data about users' comments published regarding the post, predict the qualitative ordinal level of emotional reactions that the post will trigger*.

The main aim is to classify a news post with regards to the volume of the emotional reactions it will trigger per emotion. We focus on the following five standard emotions: *love*, *surprise*, *joy*, *sadness*, *anger*. We address the problem as both a 3-class and a 5-class ordinal classification task to capture the different levels of the reactions. Hence, given a news post we assign to it one of these labels: low, medium, high for the 3-class task and one of these labels very low, low, medium, high, very high for the 5-class task per emotion.

5.3 Modeling Emotional Reactions Prediction

In this section, we present our model for the emotional reactions prediction problem. First, we introduce different textual and semantic features that can be extracted before the post is published online. Those features are important to understand why a specific news post triggered massive emotions. Next, we propose features that are extracted from users' comments published shortly after the post's publication. These features are important to investigate if there is any pattern in commenting that can be useful for predicting the emotional reactions' ordinal level.

5.3.1 Pre-Publication Features

Here, we propose features that can be extracted before the publication of the news post.

Publication Date

The date of publication has been widely studied for popularity prediction [15, 152]. In a similar way, we study if the date of publication affects the number of emotional reactions that the post will trigger. We explore the following features: Day of month (1-31), month of publication (1-12), hour of the day (0-23), day of the week (1-7), week of the month.

Term Frequencies

Terms is one of the most important feature for news articles' popularity prediction [6, 152] and sentiment analysis as well as for similar information retrieval tasks [8]. We use the classic term frequency-inverse document frequency (TF-IDF) approach [136] that considers how important is a term in a corpus to represent the content of the post. On the contrary to other studies [152] that used only a small percentage of the vocabulary to represent textual features, we are using all the terms that appear in the collection after stopwords removal. In the rest of the paper, we use the word terms to refer to the TF-IDF representation of the post's content.

Similarities

In addition to terms, we explore four different similarity functions [155]. To this end, for each news post we compute its content similarity with the posts that triggered a large number of each emotion. More formally, let d be a document (i.e., news post) that has to be classified into one of k classes (e.g., low, medium, high). Also, let H be a hyper-document (i.e., aggregation of several posts) of the documents that attracted a large number of a specific emotional reaction e. Then, we can calculate different similarity measures between d and H as described below.

Jaccard similarity. This measure computes the Jaccard similarity between a document d and the hyper-document H as:

$$Jaccard(d,H) = \frac{|W_d \cap W_H|}{|W_d \cup W_H|}$$

In other words, the similarity is calculated using the set of common terms that appear in document d and in the hyper-document H.

Cosine similarity. Cosine similarity is a well known measure for calculating the similarity between documents. The cosine similarity between a document d and a hyper-document H can be estimated as:

$$cosine(d,H) = \frac{\sum_{w \in d} P(w|d)P(w|H)}{\sqrt{\sum_{w \in d} P(w|d)^2 \sum_{w \in H} P(w|H)^2}}$$

where $P(w|d)$ and $P(w|H)$ are the probabilities of a word w occurring in d and hyper-document H respectively.

Symmetric Kullback-Leibler Divergence. The Symmetric Kullback-Leibler Divergence (Symmetric-KLD) computes the similarity between the document d and the hyper-document H based on the distance function known as KL-divergence. We consider the symmetric version of the KL-divergence to compensate for terms that do not appear in any of the distributions. This measure is calculated as:

$$\delta(H,d) = \frac{1}{2}[\ KLD(d|H) + KLD(H|d)]$$

where

$$KLD(d|H) = \sum_{w \in d} P(w|d) \cdot \log \frac{P(w|d)}{P(w|H)}$$

Normalised Kullback-Leibler Divergence. The last similarity measure that we consider is the normalised version of the KL-divergence (Normalised-KLD). Let D be the background collection, then Normalised-KLD is calculated as follows:

$$KLD(d|H) = \sum_{w \in d} P(w|d) \cdot \log \frac{P(w|H)}{P(w|D)}$$

where $P(w|D)$ is estimated based on the background model of the collection.

Reactions Entropy

Although the frequency of terms is a strong feature, it doesn't capture the discriminative power of terms. Therefore, we propose the reactions entropy that can measure how well a term separates documents that attract a high number of emotional reactions from those with a lower number. For example, a term that occurs only in documents with a high number of *angry* reactions has a high entropy compared to another term that appears also in documents with a low

number of *angry* reactions. This measure is inspired by the temporal entropy that was used to determine the time-stamp of a document [82]. To calculate the entropy of a word, first we divide the documents based on the number of reactions they received given a specific emotion. Let $V = \{v_1, v_2, ..., v_N\}$ be a set of partitions where v_1 and v_N contain the documents with the highest and lowest number of reactions respectively. Then the entropy of a word w is defined as follows:

$$VE(w) = 1 + \frac{1}{\log N_V} \sum_{v \in V} P(v|w) \cdot \log P(v|w)$$

where N_V is the number of partitions in the collection and $P(v|w)$ is the probability of the word w to occur in the partition v and is calculated as:

$$P(v_j|w) = \frac{tf(w, v_j)}{\sum_{m=1}^{N_V} tf(w, v_k)}$$

where $tf(w, v_j)$ is the frequency of w in v_j.

Following, we can calculate a score between each document and each volume partition. Given a document d and a partition v this score can be calculated as:

$$Score(d, v) = \sum_{w \in d} VE(w) \cdot P(w|d) \cdot \log \frac{P(w|v)}{P(w|D)}$$

Thus, a term appearing only in the documents that have a high value for a specific reaction will have a high entropy.

Semantics

In general, semantics can capture the similarity of documents that do not share similar terms. The reason for introducing these features is that they can also be very useful for certain entities and concepts which trigger specific reactions and specific volume of such reactions (e.g., music groups). We believe that this can be useful for improving the performance of our task, although this may not be the case for all concepts and entities (e.g., politicians and actors are both represented by the entity persons).

There are several open APIs that can be used to extract entities and concepts from text, including AlchemyAPI, DBPedia Spotlight, and Zemanta. We decided to use Alchemy API[2] to extract entities and concepts. The reason behind this decision is that AlchemyAPI proved to perform better than other APIs on entity extraction from news articles [131] and tweets [134].

[2] `https://console.bluemix.net/`

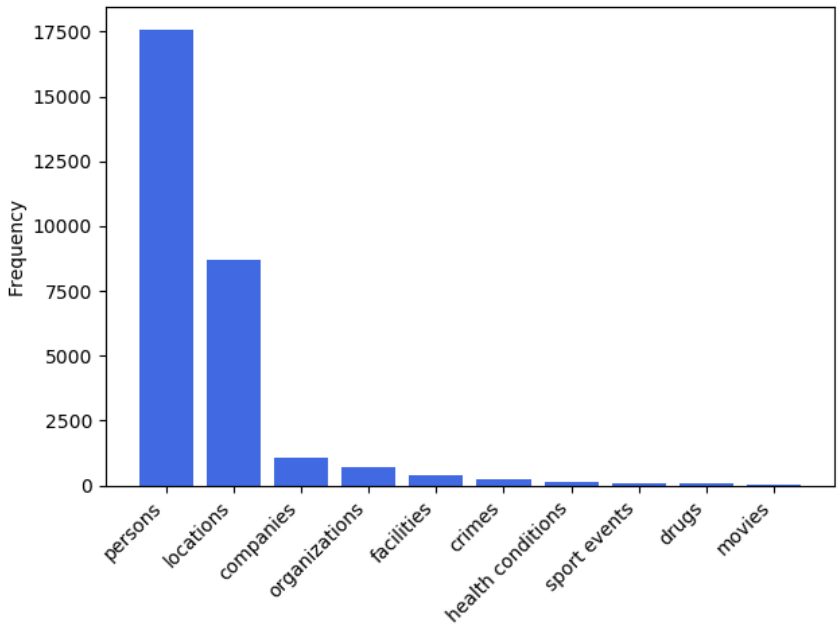

Figure 5.2. Frequency of occurrences in posts of the most frequent extracted entities.

Entities. We apply named entity recognition to extract names of persons, companies, organizations, locations, facilities, crimes, drugs, sport events, movies, and health conditions. Then, we count how many times each entity category appears in each post (e.g., number of persons). Figure 5.2 shows the number of occurrences for each of the most frequently extracted entities from our collection. It can be observed that the most frequent entity is person, while the next two most frequent entities are location and company.

Semantic Concepts. The second group of semantic features are the concepts of the post. These concepts include categories such as art and entertainment, technology and computing, food and drink, etc. We use the primary concepts extracted from every post as additional features for predicting the amount of emotional reactions. The primary concept is determined from the Alchemy API. Figure 5.3 shows the frequency of posts versus the most frequent primary concepts. It can be observed that the most frequent concept is law, government and politics.

Sentiment and Emotion. We use the EmoLex lexicon [109] to compute the sentiment (positive and negative) and the emotion (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) scores for each news post. Even if there is no available lexicon for the reaction *love*, we believe that words that convey *joy*

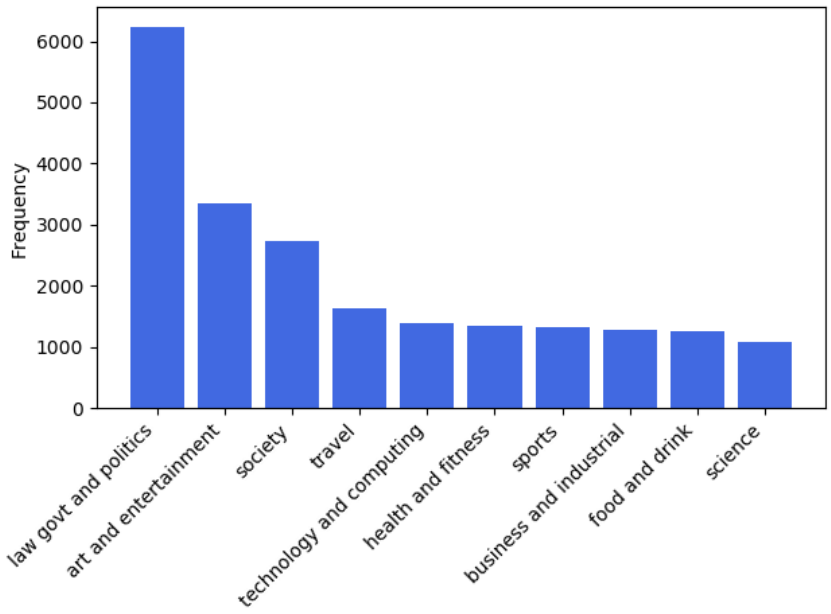

Figure 5.3. Frequency of posts versus extracted concepts.

will be also useful for *love*. The main reason of using EmoLex is the lack of training data. Obtaining training data via Crowdsourcing is a very costly process and still may contain low quality annotations especially due to the difficulty and subjectivity of the sentiment analysis task. We calculate a score for all the eight emotions provided in EmoLex, because we believe that there is some correlation among the emotions. For example, posts that express *love* are not likely to express *disgust*. For generating the scores, we count the number of predefined words provided by EmoLex in each post (for example, to compute the *sadness* score for a post, we count the words that, in the lexicon, are annotated to indicate *sadness*, and which appear in the post).

5.3.2 Post-Publication Features

In this section, we propose features that are extracted from users' comments. We propose two groups of features extracted from users' comments. The first group represents the commenting activity and includes features such as how fast the users publish a comment. For the second group we extract features from the comments' content such as their relevance to the news post.

Here, we should note that the activity of emotional reactions (e.g., number of *sadness* reactions in the first ten minutes) can also be very useful. However, we

do not have access to these data. Therefore, we use features from comments to capture early patterns in users' comments. To extract the commenting features we use three different time ranges: 10, 20, and 30 minutes after the publication time of the news post to explore how useful the different time ranges are and if there is any improvement in performance when a wider time range is used. Also, we do not differentiate between comments and replies to comments.

Early Commenting Activity. The early commenting activity features aim to capture the patterns in the activity of publishing comments below the news post. We explore the following features:

1. *First comment.* Time difference in seconds between publication date of the post and the first comment (if the first comment is published within the specified time range).

2. *Number of comments.* Number of comments published within the specified time range.

3. *Commenting ratio.* Mean time of commenting for those published within the specified time range.

4. *Unique authors.* Number of unique authors for the comments published within the specified time range. This feature can partially capture the discussion activity in the comments since a certain author will probably post more than one comments when there is a discussion.

Early Comments' Content. In this section we propose features that are extracted from the content of the comments published by users. The comments are published below the news posts. These features can reveal if there is any pattern in the content of the comments that are posted about the news post and the emotional reactions it triggers. We propose the three following features:

1. *Length of comments.* This feature is calculated based on the average length of the comments published within the specified time range. The length of a comment is represented by the number of words it contains. This feature is useful because users might tend to post shorter or longer comments regarding the news posts that trigger specific emotional reactions. In addition, longer comments might express stronger emotional reactions that may relate to the reactions triggered regarding the news post.

2. *Relevance to the post.* This feture represents the average relevance of the comments published within the specified time range to the post. This feature is important since there may be comments not related to the post. To

calculate the relevance, we use the word2vec model that is an embedding model proposed by Mikolov et al. [103]. This model learns word vectors via a neural network with a single hidden layer. First, we calculate the average vector for all words in the comment and the post and then we use cosine similarity between the vectors to calculate the similarity score. We use the pre-trained word embeddings that is publicly available and which is generated from news articles to generate the word vectors[3].

3. *Sentiment in comments.* We also measure the sentiment expressed in the comments published within the specified time range. In particular, we calculate the positive, neutral and negative sentiment ratio in the comments. We use an opinion lexicon [74] to calculate the sentiment expressed in a comment. More formally, let $N_t(d,s)$ be the number of comments that express a sentiment s towards the news post d posted during a particular time period t and $N_t(d)$ the number of total comments posted regarding d at t. Then, we define the ratio of comments that share a common sentiment s as:

$$r_t(d,s) = \frac{N_t(d,s)}{N_t(d)}$$

We calculate the ratio for all the three sentiment polarities: positive, neutral and negative.

5.4 Experimental Setup

In this section, we describe the dataset, the experimental settings and the evaluation process we followed to measure the effectiveness of our model on the emotional reactions prediction task.

5.4.1 Dataset

Our dataset consists of news posts collected from The New York Times page[4] in Facebook together with the 5 standard emotional reactions: *love*, *surprise*, *joy*, *sadness*, and *anger* for each post. We use Facebook API[5] to collect the posts, reactions, and comments. Facebook allows users to select an emotional reaction with regards to a post. This information is used to determine how many reactions

[3] https://code.google.com/p/word2vec/

[4] `https://www.facebook.com/nytimes/`

[5] `https://developers.facebook.com/`

each post has triggered. Other types of posts, such as tweets, do not contain information about emotional reactions, and therefore, they need to be manually annotated, a process that is very costly in time and resources.

The collection consists of 26,560 news posts that span from April 2016 to September 2017. We use a 10-fold cross validation to run the experiments. We keep training and test sets always separate. As an example, Figures 5.4 and 5.5 show the distribution of the posts with regards to the emotional reaction *love*. More specifically, Figure 5.4 shows the number of posts versus the number of the *love* reactions they triggered. For clarity reasons, we show only the first part of the distribution and cut the long tail after 1,000 *love* reactions. Figure 5.5 shows the number of *love* reactions per post versus the number of posts with that number of *love* reactions. The other emotional reactions follow similar distributions. From the figures, we can observe that the number of reactions per post follow a long-tail distribution. In other words, few posts collect a high number of reactions, while the majority of posts get very few.

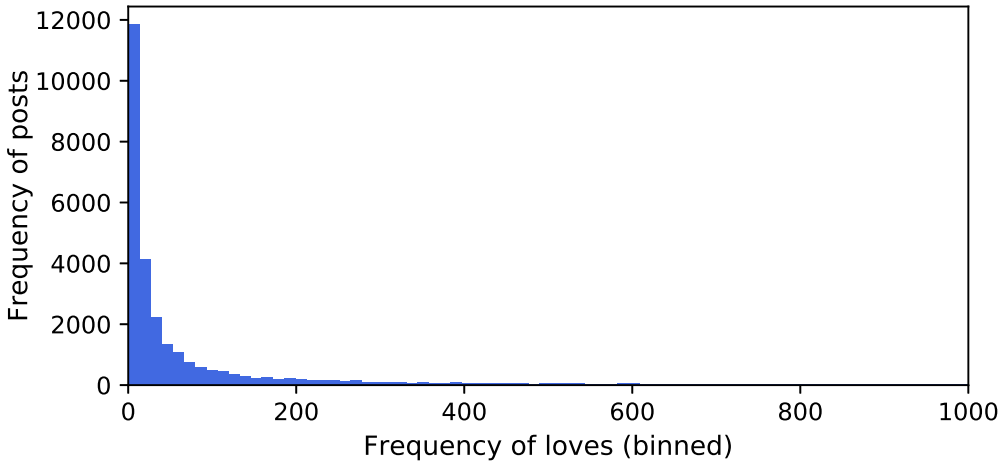

Figure 5.4. Frequency of posts versus number of the emotional reaction love (binned).

5.4.2 Experimental Settings

We perform both a 3-class and 5-class emotional reaction ordinal classification task. For those tasks, we divide the collection into 3 and 5 respectively, balanced classes with regards to the number of each emotional reaction. A balanced classification formulation has also been chosen by several prior studies on popularity

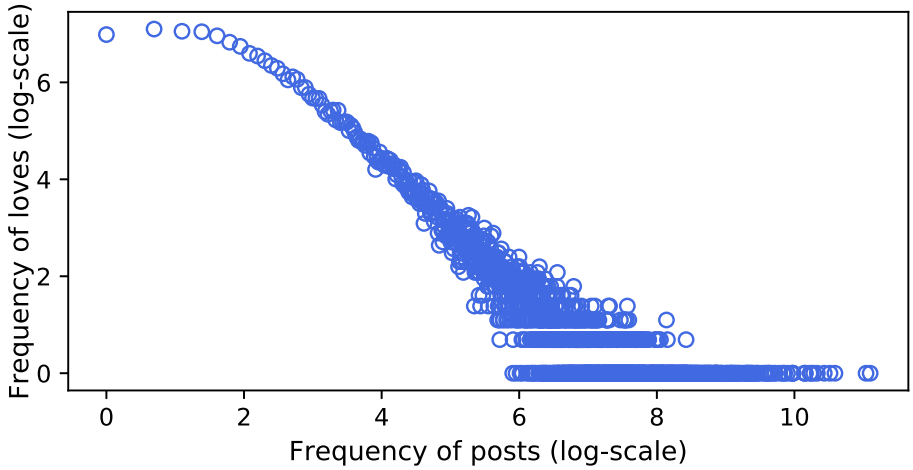

Figure 5.5. Number of love reactions per post versus number of posts with that number of love reactions (log-scale).

prediction [139, 31]. For the 3-class task a news post can get one of the following labels: *low, medium, high*, while for the 5-class one of the *very low, low, medium, high, very high*. We predict the number of the following five different emotional reactions: *love*, *surprise*, *joy*, *sadness*, and *anger*. The emotional reactions, that are available on Facebook, are addressed individually.

Table 5.1 shows the boundaries of the different classes. We observe that the range of the *high* and *very high* classes of the 3-class and 5-class classification respectively is wide. For example, the class *very high* of the 5-class classification contains posts that received from 122 to 67K *love* reactions. This is due to the long-tail distribution of the data and the balanced classes setting.

For the ordinal classification of the emotional reactions, we use Random Forest [26], a decision tree meta classifier[6]. For all the expirements, we use the open source machine learning toolkit scikit-learn[7]. To extract the entities and the concepts from each post we use Alchemy API[8]. We use the EmoLex lexicon [109] to calculate the sentiment and emotion scores of the posts and the opinion lexicon decribed in [74] to calculate the sentiment expressed in a comment. All the similarity measures are calculated between a post and a hyper-document which included 10% of the posts appeared in the training set that collected the highest

[6]We use Random Forest because it obtained the best results on the run trained on terms among the various classifiers that we tried including SVM and Logistic Regression

[7]`http://scikit-learn.org/`

[8]`https://console.bluemix.net/`

Table 5.1. Boundaries of the different classes.

3-class					
	Love	Surprise	Joy	Sadness	Anger
Low	0-9	0-8	0-3	0-2	0-2
Medium	10-47	9-39	4-21	3-31	3-35
High	48-67K	40-23K	22-27K	32-50K	36-67K
5-class					
	Love	Surprise	Joy	Sadness	Anger
Very Low	0-5	0-4	0-1	0-0	0-0
Low	6-13	5-11	2-4	1-4	1-3
Medium	14-33	12-28	5-13	5-18	4-17
High	34-121	29-89	15-63	19-110	18-134
Very high	122-67K	90-23K	64-27K	111-50K	135-67K

number of the reaction. To generate the word vectors we use publicly available pre-trained word embeddings[9]. Pre-processing of the posts involves stop-words removal and stemming with Porter stemmer [127].

5.4.3 Evaluation

The Mean Absolute Error (MAE) is reported for both 3-class and 5-class tasks per emotional reaction. The MAE is appropriate since it handles the misclassified instances in a different way (i.e., an instance *high* classified as *very high* contains lower error value compared to an instance *high* classified as *very low*). MAE has been already defined in Section 4.5.1.

We use one weak and one strong baseline. The weak baseline refers to the publication date features (i.e. hour, day of the week, month, day of the month, week of the month). The same baseline was used by Tsagkias et al. [152]. The strong baseline is the run trained on terms. Significance is measured with the non-parametric Wilcoxon signed-rank test that is appropriate for the ordinal classification.

[9] `https://code.google.com/p/word2vec/`

5.5 Results and Analysis

In this section, first we report the performance of the pre-publication prediction. Next, we report the results using the early commenting features and the performance of combining the post's terms with the early commenting features. Finally, we perform a feature analysis to investigate the importance of the features in predicting the emotional reactions.

5.5.1 Pre-Publication Prediction

Table 5.2 shows the results for the 3-class classification (i.e., *low, medium, high*) for each emotional reaction (*love, surprise, joy, sadness, anger*) using the different pre-publication features with regards to MAE (i.e, the lower the MAE value, the better the run performs). From the results, we observe that the most important feature is the post's terms (i.e., the strong baseline). Indeed, the run trained on terms outperforms all the rest of the runs, for all the emotional reactions. Terms were also proved to be very important for news articles popularity prediction [152]. However, in Tsagkias et al. [152], the rest of the features perform slightly worse than terms, whereas in our results we observe large negative differences (e.g., $\Delta = 24.91\%$ for terms over date for the classification regarding *love*). One explanation for this, is that in Tsagkias et al., the textual feature refers to the top 100 terms ranked by log-likelihood, whereas in our experiments we use the whole vocabulary (after removing stopwords) representing more than 20,000 unique terms.

We observe that the rest of the features manage to obtain similar performance and in most of the cases they significantly outperform the weak baseline (i.e., date). The run trained on similarities performs better than others. This result confirms the importance of content-based features for predicting the number of emotional reactions. Also, we observe that date has little predictive power. That means that the publication date is not important for predicting the number of emotional reactions that a news post will trigger. This result is consistent with previous studies that focused on popularity prediction of news articles [15, 152].

Table 5.2 also shows the results when all the pre-publication features are combined. Given the large difference in performance between using terms and the rest of the features, we combine all the features for two settings: all features without terms (All (- terms)) and all features with terms (All (+ terms)). Surprisingly, the model that is trained only on terms performs better compared to the combination of the features.

Table 5.3 shows the results for the 5-class classification for each emotional

Table 5.2. Performance results (MAE) for the 3-class pre-publication prediction. Scores with ∗ indicate statistically significant improvements with respect to date approach. Scores in italics indicate the best performance per emotional reaction (i.e., per column).

	Love	Surprise	Joy	Sadness	Anger
Date	0.808	0.836	0.852	0.830	0.821
Post's terms	*0.629*∗	*0.649*∗	*0.542*∗	*0.565*∗	*0.503*∗
Similarities	0.743∗	0.726∗	0.628∗	0.661∗	0.601∗
Entropy	0.747∗	0.733∗	0.670∗	0.678∗	0.631∗
Entities	0.856	0.866	0.728 ∗	0.826∗	0.760∗
Concepts	0.895	0.848	0.820∗	0.767∗	0.727∗
Sentiment	0.835	0.823∗	0.780∗	0.756∗	0.743∗
All (-terms)	0.665∗	0.667∗	0.598∗	0.588∗	0.543∗
All (+terms)	0.651∗	0.659∗	0.582∗	0.582∗	0.530∗

Table 5.3. Performance results (MAE) for the 5-class pre-publication prediction. Scores with ∗ indicate statistically significant improvements with respect to date approach. Scores in italics indicate the best performance per emotional reaction (i.e. per column).

	Love	Surprise	Joy	Sadness	Anger
Date	1.513	1.556	1.587	1.492	1.468
Post's terms	*1.232*∗	*1.269*∗	*1.101*∗	*1.059*∗	*0.982*∗
Similarities	1.392∗	1.342∗	1.220∗	1.203∗	1.117∗
Entropy	1.434	1.378∗	1.305∗	1.282∗	1.217
Entities	1.762	1.687	1.502∗	1.537	1.424
Concepts	1.738	1.712	1.666	1.372	1.480
Sentiment	1.582	1.551∗	1.504∗	1.403∗	1.383
All (-terms)	1.302∗	1.299∗	1.193∗	1.113∗	1.057∗
All (+terms)	1.293∗	1.280∗	1.163∗	1.083∗	1.022∗

Table 5.4. Performance results (MAE) for the 3-class ordinal classification using early commenting features. Scores with $*$ indicate statistically significant improvements with respect to terms approach.

	Love	Surprise	Joy	Sadness	Anger
Post's terms	0.629	0.649	0.542	0.565	0.503
$\text{activity}_{t=10}$	0.743	0.631	0.517$*$	0.730	0.596
$\text{activity}_{t=20}$	0.732	0.616	0.504$*$	0.699	0.560
$\text{activity}_{t=30}$	0.724	0.602	0.493$*$	0.690	0.544
$\text{content}_{t=10}$	0.697	0.655	0.556	0.633	0.507
$\text{content}_{t=20}$	0.686	0.660	0.583	0.618	0.507
$\text{content}_{t=30}$	0.683	0.664	0.590	0.609	0.505
$\text{activity+content}_{t=10}$	0.612$*$	0.568$*$	0.448$*$	0.586	0.442$*$
$\text{activity+content}_{t=20}$	0.581$*$	0.539$*$	0.426$*$	0.551$*$	0.408$*$
$\text{activity+content}_{t=30}$	0.555$*$	0.534$*$	0.413$*$	0.539$*$	0.388$*$

reaction for the pre-publication prediction. This task is more difficult compared to the 3-class classification, and therefore, the MAE values are higher. We notice that the performance across the features is consistent to the 3-class classification with terms outperforming the rest of the features. In addition, we observe that entities and concepts do not contain much predictive power and in some cases they even perform worse than date (e.g., on *love* classification). Similar to the 3-class classification, the model that is trained only on terms performs better compared to the combination of the features.

5.5.2 Post-Publication Prediction

Tables 5.4 and 5.5 show the results using the early commenting features on emotional reactions prediction task for the 3-class and 5-class ordinal classification respectively. The tables show the MAE scores for three different groups of features: the commenting activity features (activity), the comments' content features (content) and their combination (activity+content). We use the strong baseline to compare the results (i.e., the run based on post's terms).

From the results we observe that post's terms are better predictors compared to using only the early commenting activity or the comments' content in the

Table 5.5. Performance results (MAE) for the 5-class ordinal classification using early commenting features. Scores with $*$ indicate statistically significant improvements with respect to terms approach.

	Love	Surprise	Joy	Sadness	Anger
Post's terms	1.232	1.269	1.101	1.059	0.982
activity$_{t=10}$	1.396	1.195$*$	1.009$*$	1.334	1.122
activity$_{t=20}$	1.377	1.161$*$	0.989$*$	1.300	1.070
activity$_{t=30}$	1.362	1.142$*$	0.956$*$	1.275	1.044
content$_{t=10}$	1.334	1.249 $*$	1.078$*$	1.175	0.989
content$_{t=20}$	1.311	1.250$*$	1.114	1.151	0.972$*$
content$_{t=30}$	1.298	1.256$*$	1.125	1.124	0.960$*$
activity+content$_{t=10}$	1.177$*$	1.093$*$	0.895$*$	1.103	0.857$*$
activity+content$_{t=20}$	1.112$*$	1.039$*$	0.846$*$	1.042$*$	0.794$*$
activity+content$_{t=30}$	1.074$*$	1.021$*$	0.822$*$	1.014$*$	0.766$*$

cases of *love*, *sadness* and *anger*. However, in case of *surprise* and *joy* the early commenting activity runs perform better compared to terms and in fact in some cases the difference is statistically significant better (e.g., 5-class classification of *surprise* and *joy*). Also, we observe, that in general the runs that use the comments' content features obtain a lower performance compared to terms. Two exceptions are the cases of *surprise* and *joy* on the 5-class task where there are runs that perform statistically better to terms (e.g., content$_{t=10}$ run).

Regarding the performance between the runs that are based only on the activity and those based only on the comments' content, the emotional reactions perform in a different way. More specifically, activity leads to better performance compared to comments' content in case of *surprise* and *joy*, whereas for *love*, *sadness* and *anger*, the comments' content features are better predictors compared to activity. This result suggests that users' may follow different patterns in commenting regarding the different emotional reactions and they probably tend to write more useful comments regarding *love*, *sadness* and *anger*.

More importantly, the majority of runs that use all the early commenting features (i.e., activity+content) perform statistically better compared to the ones trained on the terms of the post. The only exception is the case of *sadness* in the activity+content$_{t=10}$ run. This suggests that in case of *sadness* the terms from the

post are stronger predictors compared to commenting activity. However, the results also prove that for most of the reactions the features that are extracted from the users' commenting activity shortly after the post is published can effectively predict the number of emotional reactions.

Tables 5.6 and 5.7 show the performance of models that are trained on combining the terms extracted from the news post with the early commenting features (activity+content$_{t=10}$) for the 3-class and 5-class classification respectively. We use features from the first ten minutes (i.e., $t = 10$) because we believe that they are very important for the prediction while keeping the advantage of quick access after the post is published.

Table 5.6. Performance results (MAE) for the 3-class classification on combining terms with early commenting features. Scores with $*$ and † indicate statistically significant improvements with respect to terms and activity+content$_{t=10}$ respectively.

	Love	Surprise	Joy	Sadness	Anger
Post's terms	0.629	0.649	0.542	0.565	0.503
activity+content$_{t=10}$	0.612	0.568	0.448	0.586	0.442
terms+activity+content$_{t=10}$	0.540$*$†	0.510$*$†	0.405$*$†	0.499$*$†	0.403$*$†

From the results, we observe that the performance, after combining terms with early commenting features, is significantly improved over both terms and activity+content$_{t=10}$ runs. However, this improvement is not consistent across the different emotions. For example, the least improvements over terms are observed for the reaction *sadness* (e.g., regarding the 3-class classification, the improvement of terms+activity+content$_{t=10}$ over terms is 12.41%) whereas the largest improvements are observed for *joy* (the respective improvement is 28.93%).

One possible explanation for this inconsistency could be that in case of news that trigger large amounts of *anger* and *sadness*, the textual features are very important predictors regardless if they are extracted from the news post or the comments' content. To investigate if there are any different patterns in commenting across the different reactions, we display the boxplot of the number of comments published in the first ten minutes for each class and for each emotional reaction in Figure 5.6. The figure suggests that there is a difference in the distributions of *sadness* compared to *joy* and *surprise*. Therefore, we also calculate the statistical differences in the number of comments published in the first ten minutes for the posts that triggered a high number of *sadness* compared to *surprise* and *joy*. The

Table 5.7. Performance results (MAE) for the 5-class classification on combining terms with early commenting features. Scores with $*$ and † indicate statistically significant improvements with respect to terms and activity+content$_{t=10}$ respectively.

	Love	Surprise	Joy	Sadness	Anger
Post's terms	1.232	1.269	1.101	1.059	0.982
activity+content$_{t=10}$	1.177	1.093	0.895	1.103	0.857
terms+activity+content$_{t=10}$	1.078$*$†	1.012$*$†	0.830$*$†	0.949$*$†	0.789$*$†

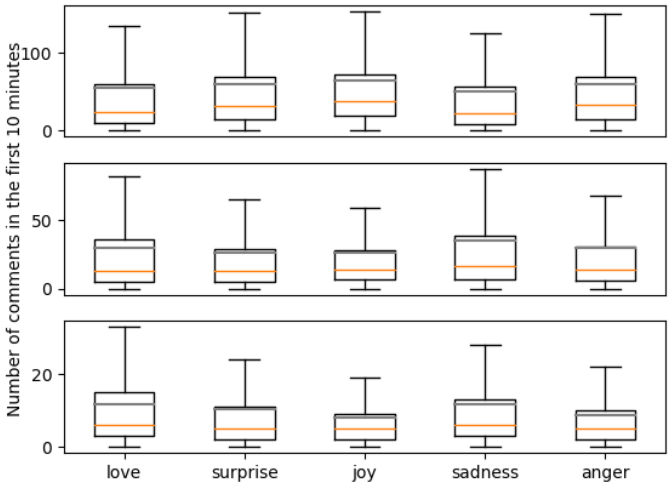

Figure 5.6. Boxplot showing the number of comments published in the first ten minutes for the five emotional reactions and the classes low, medium, high. The yellow and black line refer to median and mean number of comments respectively.

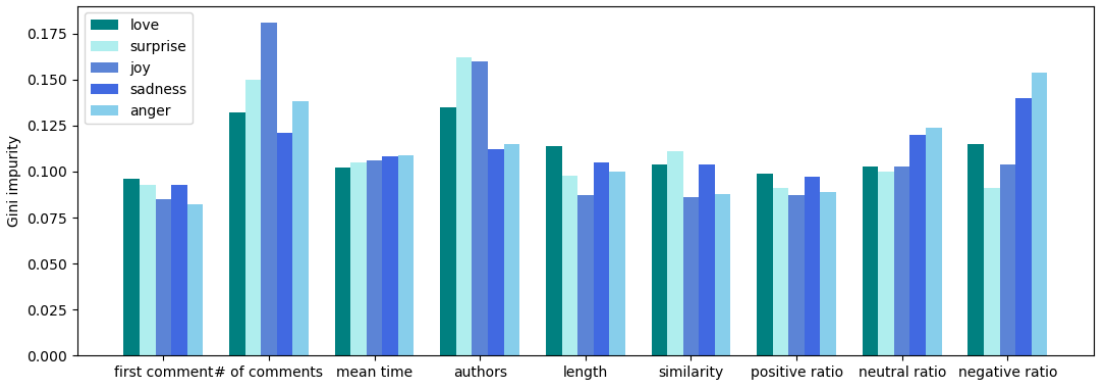

Figure 5.7. Gini importance score for the activity+content$_{t=10}$ run for the 3-class classification per each emotional reaction.

results showed that there is a statistical difference between *sadness* and *surprise* (2-sample t-test, p-value < 0.001) as well as *sadness* and *joy* (2-sample t-test, p-value < 0.001). This suggests that users may have different reaction patterns on news posts that trigger *sadness* compared to those that trigger *surprise* or *joy*.

5.5.3 Feature Analysis

In this section, we analyse the contribution of the early commenting and terms features on the emotional reactions prediction task.

Analysis on early commenting features. To understand the contribution of each feature extracted from the comments on the prediction, we calculated the Gini impurity scores as described in [27]. Figure 5.7 shows the Gini impurity score for each feature in the activity+content$_{t=10}$ run for the 3-class classification per each emotional reaction. From the figure we observe that the number of comments that have been published in the first ten minutes are good predictors for all the five emotional reactions. Indeed for the reaction *joy*, the number of comments is the best predictor. Similar, the number of unique authors feature is important for the reactions *joy* and *surprise*.

An interesting observation is that in case of *sadness* and *anger*, the negative ratio has the highest Gini impurity score. This result suggests that users tend to express their feelings in the comments regarding the posts that trigger *sadness* or *anger*.

Analysis on terms. In addition, we performed further analysis to explore which terms are the most informative for the prediction. As example, we present the top 20 terms that are the most informative for the 3-class classification of

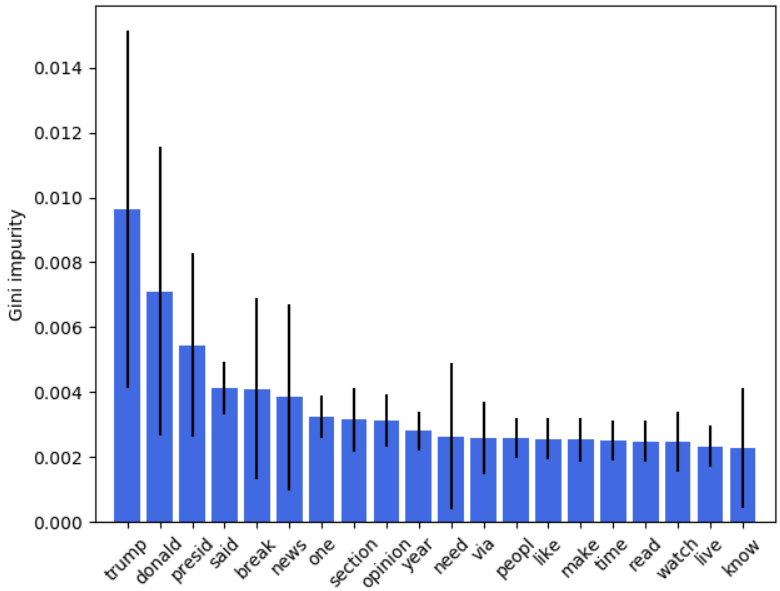

Figure 5.8. Top 20 most important terms for the 3-class classification for surprise.

the emotional reactions *surprise* and *sadness*. The importance of the features is based on Gini impurity as described in [27]. Figures 5.8 and 5.9 show the most informative terms sorted by their importance for the reactions *surprise*, and *sadness* respectively. We observe that in both cases the most informative terms are *donald*, *trump* and *president*. We believe that this happens because of the time range of our collection that contains a lot of articles referring to US Elections 2016. In addition, we observe that there are also some terms that convey sentiment, such as the terms *kill* and *attack* that are informative for the emotional reaction *sadness*.

What is important to mention is that there are some words that are informative for both emotions (e.g., breaking, Donald, Trump, president). This observation suggests that there are terms that in general trigger either a large or a low number of emotional reactions regardless the emotion. In addition to those terms, there are also terms that convey emotion (e.g., excited, attack, etc.) and are important only for a specific emotion (e.g., *sadness*).

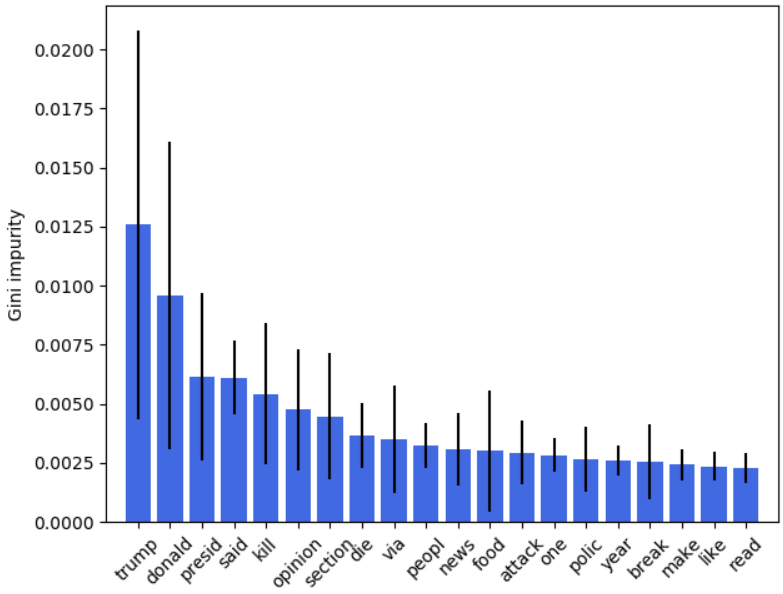

Figure 5.9. Top 20 most important terms for the 3-class classification for sadness.

5.6 Conclusions

In this chapter, we proposed an effective approach for predicting the emotional reactions triggered by news posts on users. We focused on the following five emotional reactions: *love*, *surprise*, *joy*, *sadness*, and *anger*. We proposed features extracted from the news post content to perform a pre-publication prediction and features extracted from users' comments for a post-publication prediction. In addition, we studied the effectiveness of combining early commenting features with news posts' terms on predicting the volume of emotional reactions.

We found that the most important feature for the pre-publication prediction is the terms of the post. The rest of the pre-publication features (e.g., similarities, reactions entropy, entities, concepts, sentiment) managed in most of the cases to outperform the weak baseline (i.e., date), but not the strong baseline (i.e., terms). Surprisingly, a model trained only on terms outperformed even the combination of all the pre-publication features. Also, we showed that features extracted from users' comments are very important for the emotional prediction task. More importantly, we showed that the commenting features contain more predictive power compared to terms for all the emotions except for *sadness*. In addition, our results suggested that the most effective models to predict which posts trigger a high number of emotional reactions are those trained on both

posts' terms and users' comments.

Finally, our analysis showed that the different features extracted from comments are not equally important for the different emotions because there are probably different commenting patterns across the different emotional reactions. For example, one interesting observation from the feature analysis is that the negative ratio is the most important feature for *sadness* and *anger*, whereas the most important predictor for the reaction *joy* is the number of comments.

Chapter 6

Sentiment Propagation for Reputation Polarity

Estimating the impact of what is posted online on the reputation of an entity is an important aspect of tracking sentiment over time. Starting from the observation that similar tweets share the same reputation polarity, in this chapter we explore the effectiveness of propagating sentiment signals to tweets that are about the same topic. We consider augmenting the sentiment lexicons with terms that indicate reputation polarity even if they do not convey sentiment polarity and direct propagation of sentiment signals to texts with similar content. We explore different approaches of estimating the similarity of tweets and of selecting the set of similar tweets. Our results show that sentiment lexicons can be augmented to create reputation polarity lexicons, and that the domain level is a cost-effective level of granularity for doing so. In addition, we show that weakly supervised annotation of reputation polarity is feasible, which is a promising result as such methods are less dependent on the availability of training data. Finally, regarding the different ways of learning a threshold, we do not find any differences among the approaches.

6.1 Introduction

One important aspect of tracking public opinion on social media is to estimate its impact on the entity's reputation. Reputation management analysts are becoming highly interested in tracking and monitoring what is posted about the interested parties in social media. The posts that are published online can be disseminated very fast with major consequences on the reputation of an entity. Hence, one of the core tasks of Online Reputation Monitoring is determining if a

post that is about an entity (e.g., company, person) is going to have a positive or negative impact on the entity's reputation. This task is known as *reputation polarity analysis* and is very important since it gives the opportunity to the interested parties to act promptly when their reputation could be damaged.

Determining the impact of a post on an entity's reputation is a challenging task. One of the key challenges is that several posts are short and contain informal language with several peculiarities. For example, tweets contain a lot of slang language or emoticons and are very different from conventional text. The task of reputation polarity analysis has several similarities with sentiment analysis and therefore, prior work on reputation polarity has evolved from sentiment analysis [122]. However, the two tasks are not the same. For example, one of the state-of-art approaches in sentiment analysis is the lexicon based approach which is based on sentiment lexicons that contain terms that express sentiment. However, sentiment lexicons are general and not effective for reputation polarity. Hence, they need to be expanded with new terms that indicate reputation polarity.

Reputation polarity is more challenging compared to sentiment analysis. One challenge with estimating reputation polarity refers to posts that do not explicitly express a sentiment but have an impact on the entity's reputation. These posts (e.g., tweets) are known as *polar facts*. One example of a polar fact is the tweet *RBS becomes first UK bank to back Visa V.me wallet* which has a positive impact on *RBS* but does not express any sentiment. Therefore, an effective reputation polarity approach should also consider how to identify polar facts and how to estimate their reputation polarity.

In this chapter, we focus on estimating the impact of what is posted online on an entity's reputation. More specifically, we propose sentiment signals propagation to estimate reputation polarity of tweets. We consider two ways of propagating sentiment signals: (i) augmenting the sentiment lexicons with terms that indicate reputation polarity even if they do not convey sentiment polarity; and (ii) direct propagation of sentiment signals to texts with similar content. To address the challenge of polar facts, we propose a polar fact filter that can differentiate between polar facts and reputation-neutral tweets. In addition, we hypothesize that tweets that are about a specific topic tend to have the same reputation polarity. In this way, if there are many tweets about a specific topic, then some of those tweets will explicitly express some sentiment towards the topic and can be used to annotate the polar facts.

Table 6.1 shows some examples of tweets relevant to the entity *HSBC* that are about the *accusations* topic. Table 6.1 also shows the actual (manually annotated) reputation polarity of each tweet, and the sentiment polarity as assigned

by a state-of-the-art lexicon based approach. From the table, we observe that there are some tweets (i.e., *t3*) that do not convey sentiment (the sentiment by lexicon is neutral) but they have a negative reputation polarity, whereas other tweets have an explicit sentiment indicator (i.e., *t1, t2*). Hence, we can use the information from the sentiment of tweets *t1* and *t2* to learn the reputation polarity of *t3*.

Table 6.1. Examples of annotated tweets in the RepLab 2013 training dataset.

Oracle Topic	id	Tweet	Reputation Polarity	Sentiment by Lexicon
accusations	t1	When I wake up I want to find these trending: Barclays, HSBC, executive arrests, fraud & Tory party. NOT Justin Bieber	Negative	Negative
accusations	t2	THE CORPORATE POLITICIANS: 20 years of failure for Britain as they skimmed the system. #cnn, #times, #cnbc, #hsbc	Negative	Negative
accusations	t3	@PoliticalPryers he's ceo of one of the banks involved . He high but not the top! By this time next week RBS, Llyods, HSBC will get same	Negative	Neutral

In this chapter, we focus on the following research question:

RQ4 How can we estimate the impact of posts on an entity? Can we use sentiment signals propagation to estimate the impact of posts on an entity's reputation?

This research question can be further analysed into the following more specific research questions:

RQ4.1 Can we use training material to detect terms with reputation polarity and use them to augment a general sentiment lexicon?

RQ4.2 What is the right level of generalization for a reputation lexicon?

RQ4.3 Can we propagate sentiment to text that is similar in terms of content to improve reputation polarity?

RQ4.4 What is the best way to select the set of pairwise similar tweets that can be used to learn the sentiment that will be propagated?

We evaluate our model on the RepLab 2013 collection [13], that, to the best of our knowledge, is the largest Twitter collection for reputation monitoring. The results show that sentiment signals can be propagated to similar tweets to effectively address the problem of reputation polarity.

The remainder of the chapter is structured as follows. Section 6.2 presents the lexicon augmentation approaches. Section 6.3 explains the direct sentiment propagation approach and the polar fact filter. The experimental setup is described in Section 6.4. Section 6.5 presents the results followed by an analysis and discussion in Section 6.6. Finally, Section 6.7 concludes the chapter.

6.2 Lexicon Augmentation

The first step for sentiment propagation refers to estimating sentiment in tweets. Reputation analysis shares similarities to sentiment analysis and therefore it is a good starting point. In this section, first we present a simple lexicon based approach that can be used to estimate the sentiment expressed in a tweet. Then, we present different approaches that can be used to expand the sentiment lexicons which can be then used to estimate the reputation polarity of a tweet.

6.2.1 Lexicon Based Approach

Our approach starts with a simple way of estimating the sentiment of short texts. This approach does not need any training data and is based on manually created sentiment word lists that are known as sentiment lexicons.

Let *polarity(d)* be the reputation polarity of a tweet d. The reputation polarity can take one of the values $\{1,-1,0\}$ referring to a positive, negative, or neutral reputation polarity respectively. Also, we consider $S(d)$ to be the sentiment score of a document d. Then according to the lexicon based approach, the reputation polarity of a tweet d can be calculated as:

$$polarity(d) = \begin{cases} 1, & \text{if } S_d > 0 \\ -1, & \text{if } S_d < 0 \\ 0, & otherwise \end{cases}$$

where the sentiment score $S(d)$ of a tweet $d = \{t_1, ..., t_i, ..., t_N\}$ is calculated as follows:

$$S(d) = \sum_{t \in d} opinion(t)$$

where the *opinion(t)* refers to the opinion score of the term according to an opinion lexicon. We refer to this methodology as *lexicon based approach*.

6.2.2 Simple Lexicon Augmentation

One simple approach to expand a sentiment lexicon and learn new words is to use tweets that are already labelled to expand negative and positive lists of words. This can be applied either in an unsupervised or in a supervised way. Regarding the unsupervised scenario, the lexicon based approach can be used to create the initial list of labelled tweets, whereas for the supervised scenario, the manually annotated tweets (i.e., the training data provided with the collection) can be used. Then the lexicon based approach can be applied as described in Section 6.2.1 on the expanded positive and negative lists to estimate the reputation polarity of the tweets. We refer to these approaches as *sup-lexAugm* and *unsup-lexAugm* respectively. In addition, we explore the effectiveness of the *unsup-lexAugm* and the *sup-lexAugm* approaches on different granularity levels, namely independent, domain-dependent and entity-dependent.

6.2.3 PMI Based Lexicon Expansion

In this section, we describe an alternative way to expand sentiment lexicons and learn words that indicate reputation polarity. One limitation of the simple lexicon augmentation is that it considers only positive and negative word lists. The simple lexicon augmentation approach considers the absence of positive and negative words as an indicator of a neutral sentiment. However, it is also important to learn reputation-neutral words for an effective estimation of the reputation polarity. To this end, we use a method based on the Pointwise Mutual Information (PMI) that was originally proposed by Church and Hanks [33]. In particular, the positive reputation score for a tweet d is calculated as follows:

$$PMI(d, positive) = \sum_{t \in d} PMI(t, positive)$$

$$PMI(t, positive) = \log_2 \frac{c(t, positive) * N}{c(t) * c(positive)}$$

where *c(t, positive)* is the frequency of the term t in the positive tweets, N is the total number of words in the corpus, $c(t)$ is the frequency of the term in the corpus and *c(positive)* is the number of terms in the positive tweets. A similar process is followed to calculate the PMI of the terms for the negative and neutral classes. The PMI scores are then used to predict the polarity of the test documents as follows:

$$polarity(d) = \begin{cases} 1, & \text{if } max\{PMI(d,positive), PMI(d,neutral), PMI(d,negative)\} = PMI(d,positive) \\ -1, & \text{if } max\{PMI(d,positive), PMI(d,neutral), PMI(d,negative)\} = PMI(d,negative) \\ 0, & otherwise \end{cases}$$

This approach is applied on both supervised and unsupervised settings. For the supervised setting, the training data that are already provided in the collection are used. To apply the method in an unsupervised setting, we use the lexicon based approach to estimate the initial sentiment polarities of the tweets. We refer to these approaches as *sup-PMI* and *unsup-PMI* respectively. In addition, we explore the effectiveness of the *unsup-PMI* and the *sup-PMI* approaches on different granularity levels, namely independent, domain-dependent and entity-dependent.

6.3 Direct Sentiment Propagation

In this section, we present the direct sentiment propagation approach. For this approach, apart from estimating the initial sentiment of tweets, we also measure the similarity of tweets. In addition, we differentiate the sentiment-neutral tweets into reputation-bearing and reputation-neutral tweets. The reputation-bearing tweets are tweets that have either positive or negative impact on the entity's reputation, whereas the reputation-neutral tweets have no impact on the entity's reputation. In the rest of this section, first we present the different approaches we applied to estimate the similarity of the tweets in Section 6.3.1. Then, in Section 6.3.2 we present the different approaches for selecting the set of similar tweets and for learning the propagated polarity. Finally in Section 6.3.3 we describe our approach to build the polar fact filter.

6.3.1 Similarity of Tweets

In this section we describe the different approaches we applied to estimate the similarity of tweets. First, we describe an approach that is based on clustering

with similarity functions and that is proposed by Spina et al. [143]. Next, we explain an approach that calculates the pairwise similarity of the tweets and we propose the *maxDelta* approach that is used to select the set of similar tweets.

Clustering with Similarity Functions

The first approach we consider to estimate the similarity of the tweets is based on Hierarchical Agglomerative Clustering (HAC) [99]. This approach was proposed by Spina et al. [143] and was applied on RepLab 2013 collection for the topic modeling task.

The first step of the approach is to find a classification function that takes as input two tweets and decides if the tweets belong to the same topic. Spina et al. used 13 different signals for this task that belong to four different groups: term, semantic, metadata and time-aware features. Then, a linear kernel SVM model is built for the pairwise classification. Finally, they applied HAC [99] on the previously annotated tweets to create the topic clusters. A detailed description of the approach is presented in [143].

We use this approach to create clusters of similar tweets before we propagate the sentiment to tweets that belong to the same cluster. We refer to this approach as *cluster*.

Pairwise Similarity

The next similarity approach we use is based on the pairwise similarity of tweets. To this end, we calculate the cosine similarity of all the different pairs of tweets. We represent the tweets as vectors where each vector is represented by the word probabilities [99]. The cosine similarity between a tweet d_i and a tweets d_j is calculated as follows:

$$cosine(d_i, d_j) = \frac{\sum_{w \in d_i} P(w|d_i)P(w|d_j)}{\sqrt{\sum_{w \in d_i} P(w|d_i)^2}\sqrt{\sum_{w \in d_j} P(w|d_j)^2}}$$

where $P(w|d_i)$ and $P(w|d_j)$ are the probabilities of a word w occurring in d_i and d_j.

One important challenge regarding the pairwise similarity approach is to define the set of tweets that will be used to determine the polarity that will be propagated. We consider three alternatives to define the set of pairwise similar tweets: i) all the tweets with a similarity score greater that 0, ii) learning the best threshold using the training data, and iii) the maximum similarity difference (i.e, *maxDelta*). The first two alternatives are easy to understand from the

above description. The assumption behind maxDelta is that a tweet will have a high similarity score to a particular set of tweets and a very low similarity to the rest of tweets. MaxDelta determines the reputation polarity based on the set of tweets that have a similarity score to the tweet with the unknown reputation polarity.

To present the maxDelta more formally, we introduce some notation. Let d be a tweet and $l = \{d_1, ..., d_i, ..., d_k\}$ a list of k tweets. In addition, let $cosines = \{cosine(d, d_1), ..., cosine(d, d_i), ..., cosine(d, d_k)\}$ be the ordered list of pairwise similarities between tweet d and the tweets that belong to the list l. Also, let $cosinesDelta = \{\delta_1, ..., \delta_i, ..., \delta_{k-1}\}$ where $\delta_1 = cosine(d, d_1) - cosine(d, d_2)$ and $\delta_k = cosine(d, d_{k-1}) - cosine(d, d_k)$ be a list with the cosine differences of the adjacent tweets. Then the threshold that is applied according to *maxDelta* is estimated as:

$$maxDelta = max\{\delta_1, ..., \delta_i, ..., \delta_{k-1}\}$$

Here, we should note that the first two approaches are based on a fixed threshold. For the first approach the threshold is 0 whereas for the second we learn the threshold using the training data. However, the same threshold is applied to all tweets. On the contrary, the *maxDelta* estimates a different threshold for every tweet and the selected set tweets is not based on a fixed threshold.

6.3.2 Direct Sentiment Propagation

To determine the sentiment signal that will be propagated, we propose two approaches: i) the *frequency* and, ii) the *avgCosine*. The first approach is based on the frequency of each sentiment in the set of similar tweets, whereas the avgCosine approach is based on the average cosine of each sentiment. More formally, let's assume we have a list $l = \{d_1, ..., d_i, ..., d_k\}$ of k tweets that are similar to tweet d and for which we know the sentiment polarities. The sentiment polarities can be defined as $S = \{polarity(d_1), ..., polarity(d_i), ..., polarity(d_k)\}$. The reputation polarity of a tweet d is estimated as:

$$polarity(d) = \begin{cases} 1, & \text{if } max\{C(l, positive), C(l, neutral), C(l, negative)\} = C(l, positive) \\ -1, & \text{if } max\{C(l, positive), C(l, neutral), C(l, negative)\} = C(l, negative) \\ 0, & otherwise \end{cases}$$

The frequency approach estimates the polarity scores as:

$$C(l, positive) = \sum_{i=1}^{k} [polarity(d_i) = 1]$$

where *C(l,positive)* is the number of positive tweets in the list l. The same approach is used for *C(l,neutral)* and *C(l,negative)*.

The avgCosine approach of estimating *C(l,positive)* is based on the average cosine similarity of the tweet d to the similar tweets that we consider per polarity defined as l. According to avgCosine approach the *C(l,positive)* is defined as:

$$C(l, positive) = \frac{\sum_{i=1}^{k} cosine(d, d_i) \forall d_i, polarity(d_i) = 1}{\sum_{i=1}^{k} [polarity(d_i) = 1]}$$

where *cosine (d, d_i)* is the similarity cosine score between d and d_i. The same approach is used for *C(l,neutral)* and *C(l,negative)*.

6.3.3 Polar Fact Filter

One limitation of propagating sentiment is that it tends to over-estimate the number of tweets annotated with reputation polarity. In particular, the methods may propagate the sentiment to both polar facts and reputation-neutral tweets. Therefore, in this section we present a *polar fact filter* that decides whether a tweet is a polar fact or not. To build the polar fact filter we start with tweets that do not contain sentiment words and which represent *polar facts candidates*. The polar facts candidates contain both reputation-bearing tweets and reputation-neutral tweets.

We address the polar fact filter classification as a binary classification task. The polar fact filter is trained on a linear kernel Support Vector Machine (SVM) classifier [30] that is a state-of-art learning algorithm. The classifier is trained to discriminate between two different classes, $y_i \in \{-1, +1\}$, where N is the number of the labelled training data. The training examples are $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N), \mathbf{x} \in R^k$ where k is the number of features.

To build the polar fact filter, we explore a number of different features that have proved to be effective for sentiment classification [87]. The features can be grouped in three classes as follows:

- n-grams: including character grams, unigrams, bigrams, trigrams and 4-grams.

- stylistic: including the number of capitalised words, number of elongated words, number of emoticons, number of exclamation and number of question marks.

- lexicons: including manual and automatic lexicons. In particular, we explore Liu's lexicon [74], NRC emotion lexicon [108], MPQA lexicon [165] and Hashtag Sentiment Lexicon [87].

We explore the effectiveness of the polar fact filter on three different granularity levels: independent, domain-dependent and entity- dependent.

6.4 Experimental Design

In this section we describe the experimental setup for evaluating the effectiveness of the proposed models. First, we give a description of the dataset that was used in this study. Second, we give details regarding the experimental settings. Then, we describe the implementation of the polar fact filter followed by a description of the runs. Finally, we explain the evaluation metrics we used to measure the effectiveness of the approaches.

6.4.1 Collection

We use the RepLab 2013 collection [13] that contains Twitter data in English and Spanish. To the best of our knowledge, RepLab 2013 collection is the largest Twitter collection for reputation monitoring. The tweets are about 61 different entities that belong to four domains: automotive, banking, universities, and music. The collection contains training and test data. The training data (45,679) were collected three months before the test data (96,848). For our experiments we use only the tweets that are annotated as related to the entities that refer to 34,882 tweets for the training set and 75,470 tweets for the test set. The rest of the crawled tweets (1,038,060 tweets) represent the background dataset and refer to tweets published between the training and test set.

6.4.2 Experimental Settings

We use publicly available word lexicons in English [112] and in Spanish [123] to identify the words that indicate positive or negative sentiment. We use information from tweets' metadata to identify the language of the tweet. We use the same tokeniser for English and Spanish tweets. For the results that are reported, we considered the tweets that are relevant to an entity (tweets manually annotated as *related*) from the test set.

6.4.3 Polar Fact Filter

To build the polar fact filter, we first apply the lexicon based approach on the collection. Then, we take the tweets that are annotated as neutral (19,932 sentiment-neutral tweets in the training set) and we divide them into two classes as follows: (i) tweets that are reputation-neutral, (ii) tweets that are reputation-bearing. We do not differentiate between positive and negative polarity at this stage. We follow the same process for the test data (43,748 sentiment-neutral tweets). The evaluation is estimated in three different scenarios: independent, domain-dependent, and entity-dependent.

6.4.4 Runs

We test the classification performance of the polar fact filter using the features presented in Section 6.3.3. We test the performance when the features are used alone and when they are combined. More specifically, regarding the polar fact filter we use runs with the following combinations of features:

n-grams:	all n-grams and character features
stylistic:	all stylistic features
lexicons:	all lexicon features
n-grams+styl:	all n-grams and stylistic features combined
n-grams+lex:	all n-grams and lexicon features combined
styl+lex:	all stylistic and lexicon features combined
n-grams+styl+lex:	all n-grams, stylistic and lexicon features combined

Regarding the propagation approaches that are based on lexicon augmentation we have the following runs:

sup-lexAugm:	supervised simple lexicon augmentation
unsup-lexAugm:	unsupervised simple lexicon augmentation
sup-PMI:	supervised PMI lexicon augmentation
unsup-PMI:	unsupervised PMI lexicon augmentation

Regarding the direct propagation approach we have the following runs to estimate the similarity of tweets:

cluster:	clustering tweets as described in [143]
pairwise similarity:	similarity based on cosine similarity of tweets

Regarding the direct propagation approach we have the following runs to estimate the polarity that will be propagated:

frequency: based on the frequency of each sentiment in the set of similar tweets
avgCosine: based on average cosine similarity to each sentiment

Regarding the different approaches for determining the set of tweets that are used to learn the propagated sentiment we have the following runs:

no threshold: based on all the similar tweets
best threshold: based on best threshold learned from the training data
maxDelta: based on the maximum delta difference in the similarity of tweets

Here, we should note that both the noThershold and maxDelta approaches are weakly supervised since they use the polar fact filter to differentiate between reputation-bearing and reputation-neutral tweets. However, the best threshold approach is supervised since in addition to the polar fact it also uses the training data to learn the best threshold.

6.4.5 Evaluation Metrics

We report F-scores for the evaluation of the polar fact filter and the reputation polarity approaches. Regarding the reputation polarity approaches, the performance is measured on the overall output of the classification experiments (both English and Spanish results) using the annotated data provided with the RepLab 2013 collection. We use the McNemar test [100] to measure the statistical difference, that is appropriate for comparisons of nominal data.

6.5 Results

In this section we present the results of our study. First, we present the results on sentiment propagation using the lexicon augmentation approaches. We present separately results for unsupervised and supervised settings. Next, we present the results regarding the effectiveness of the polar fact filter on differentiating polar fact tweets and reputation-neutral tweets. Finally, we present the performance of the direct sentiment propagation approach on the reputation polarity prediction task.

Table 6.2. Performance results of the approaches when trained on a unsupervised setting. A star(∗) indicates statistically significant improvement over the lexicon based approach.

Method	Independent	Domain-Dep.	Entity-Dep.
lexicon-based	0.368	0.368	0.368
unsup-lexAugm	0.371 (+0.82%)	0.392∗ (+6.52%)	0.394∗ (+7.07%)
unsup-PMI	0.296 (-19.57%)	0.278 (-24.46%)	0.305 (-17.12%)

6.5.1 Lexicon Augmentation

Our first research question was: *Can we use training material to detect terms with reputation polarity and use them to augment a general sentiment lexicon?* We investigate this question by calculating the effectiveness of the lexAugm and PMI approaches on both unsupervised and supervised setting. For the unsupervised scenario, we used the lexicon based approach to obtain initial tweets annotated by sentiment. Regarding the supervised scenario, we learned the reputation polarity lexicons using the annotated training data.

Table 6.2 shows the performance results of learning the reputation lexicons in an unsupervised way. The first approach (unsup-lexAugm) updates the initial sentiment lexicon with the new terms that are present in the training data, whereas the unsup-PMI learns the reputation polarity lexicons by estimating the pointwise mutual information of the terms. The results show that augmenting the lexicon in an unsupervised setting improves the performance of the reputation polarity task compared to the baseline. In particular, the unsup-lexAugm approach performs statistically better compared to baseline on the domain and entity dependent settings. However, the unsupervised PMI approach (unsup-PMI) achieves a lower performance compared to the baseline for all the three different granularity levels.

Here, we should note that a key difference between the lexAugm and PMI approaches is that the first learns words for positive and negative polarities, whereas PMI also considers the neutral class. A further analysis on the results showed that part of the improvement achieved by unsup-lexAugm is due to the fact that the positive augmented lexicon is much larger compared to the negative augmented lexicon (i.e., in the independent scenario there are 20,187 positive terms versus 10,750 negative terms) and to the fact that the dataset is imbalanced with the positive class being the largest one. However, unsup-PMI also considers tweets annotated as neutral by the lexicon, a large part of which are positive or

Table 6.3. Performance results of the approaches when trained on a supervised setting. A star($*$) indicates statistically significant improvement over the lexicon based approach. The Δ difference is given over the lexicon based approach.

Method	Independent	Domain-Dep.	Entity-Dep.
lexicon-based	0.368	0.368	0.368
sup-lexAugm	0.431$*$ (+17%)	0.455$*$ (+24%)	0.460$*$ (+25%)
sup-PMI	0.547$*$ (+49%)	0.572$*$ (+55%)	0.586$*$ (+59%)

negative (i.e., the polar facts), and this has an effect on its performance.

Table 6.3 shows the performance results of learning the reputation lexicons in a supervised way. The results suggest that there is a substantial improvement compared to the unsupervised scenario. This improvement occurs for all the different levels of granularity. We also observe that the best performance is achieved on the entity-dependent training scenario for both approaches. More importantly, all the improvements are statistically better than the baseline. Finally, the results show that the sup-PMI expansion approaches are more efficient compared to the sup-lexAugm approaches. This suggests that considering the neutral class for augmenting the lexicon in a supervised way can be very useful for the estimation of the reputation polarity.

Given the high performance of the sup-PMI entity-dependent approach, we are interested to investigate if it is robust across the entities. Therefore, as a further analysis we calculated the standard deviation in the F-measure across the entities regarding the sup-PMI entity-dependent approach and we found that it is very low (i.e., 0.063). This suggests that it is indeed robust across the entities.

Regarding the second research question: *What is the right level of generalization for a reputation lexicon?* We observe from the presented results that the performance improves when the lexicons become more specific. In particular, we observe that the approaches perform better on a domain-dependent compared to independent lexicon augmentation and similarly they perform better on an entity-dependent to domain-dependent setting. This observation is valid for both the augmentation approaches and for both the unsupervised and supervised settings. Regarding the lexAugm approach, we observe that even the entity-dependent lexicons perform better than the domain-dependent lexicon, the difference is very small (25% versus 24% improvement over the baseline, respectively). Therefore, regarding the lexAugm approach, the conclusion is that the domain level is a cost-effective level of granularity for augmenting a sentiment

lexicon to create a reputation lexicon. However, regarding the PMI approach, the performance difference between the domain and the entity-dependent lexicons is considerable. Regarding PMI, the results suggest that the entity-dependent expansion is the most effective level of granularity.

6.5.2 Polar Fact Filter

As already mentioned, one limitation of the lexicon based approach is that it estimates a large number of reputation bearing (positive or negative reputation polarity) tweets as neutral which are known as *polar facts*. Figure 6.1 shows the distribution of the tweets that are annotated as polarized (positive or negative sentiment) by the lexicon based approach over the total number of tweets per entity in the training set. The tweets that do not contain sentiment words represent the *polar facts candidates*. The polar facts candidates contain both polar fact tweets and reputation-neutral tweets. From the figure we observe that in general the amount of polar facts candidates is larger than the amount of polarized tweets. Also, we observe that for some entities (e.g., E001, E002) the difference in the amount between the polarized and polar fact candidate tweets is large whereas for other entities (e.g., E097, E206) the difference is small.

Figure 6.2 shows the distribution of polar fact tweets over the reputation-neutral tweets in the polar fact candidates set. Here, again we observe that the difference between the amount of polar fact tweets and reputation-neutral tweets is not consistent for all the entities. We observe that there are entities for which the majority of the polar facts candidates are actually polar facts (e.g., E055, E056) whereas there are entities which contain more reputation-neutral than polar fact tweets. This suggests that the sentiment words of the lexicons are not equally useful to label the reputation across the entities and that there are entities for which the polar facts are a very common phenomenon.

Table 6.4 shows the results of the polar fact filter for the different n-grams, stylistic, and lexicon features and their combinations. The different columns compare the results based on the independent, domain-dependent and entity-dependent training scenarios. From the results we observe that for most of the features the entity-dependent scenario yields a better F-measure compared to the independent and domain-dependent scenario.

In addition, we observe that the best performance using the stylistic features is obtained for the domain-dependent training scenario. More specifically, the performance on the domain dependent scenario is higher compared to the entity-dependent scenario for the exclamation marks and the elongated words. This difference in the performance is 12.3% for exclamation marks and 5.47% for

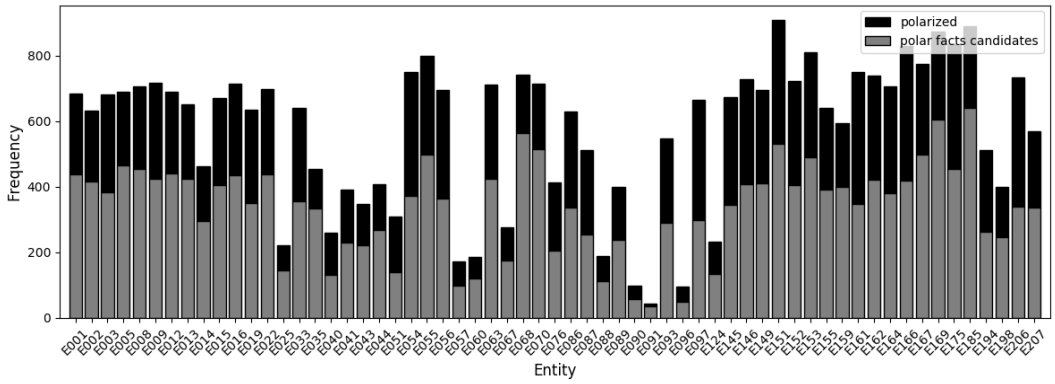

Figure 6.1. Distribution of polar fact candidate tweets per entity in the RepLab 2013.

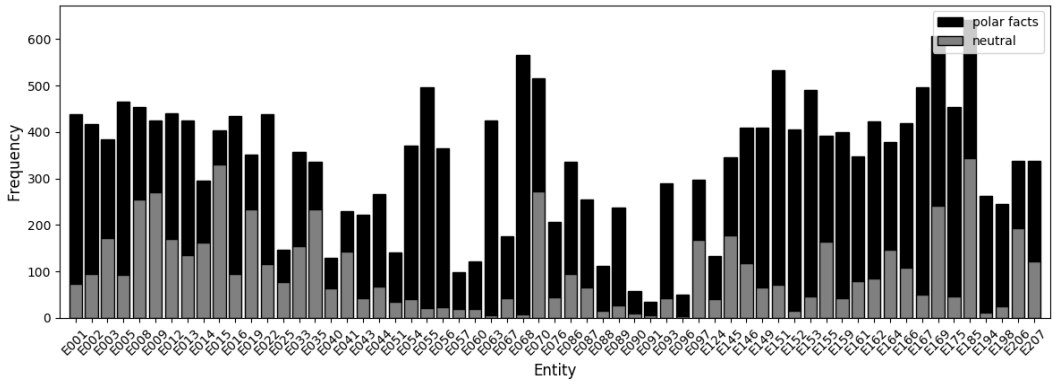

Figure 6.2. Distribution of labeled training polar fact tweets per entity in the RepLab 2013.

Table 6.4. Performance results (F-measure) of the polar fact filter classification when trained on independent (indep.), domain dependent (domain) and entity dependent (entity) data.

		Indep.	Domain	Entity
n-grams	unigrams	0.597	0.620	0.673
	2-grams	0.560	0.594	0.672
	3-grams	0.532	0.567	0.625
	4-grams	0.518	0.555	0.585
	char	0.623	0.651	0.688
stylistic (styl.)	emoticons	0.495	0.496	0.501
	exclamation marks	0.324	0.569	0.499
	question marks	0.496	0.496	0.525
	elongated	0.510	0.494	0.467
	capitals	0.472	0.467	0.518
lexicons (lex.)	Liu	0.322	0.410	0.462
	NRC	0.440	0.426	0.477
	MPQA	0.492	0.416	0.457
	HSL	0.534	0.514	0.523
groups	n-grams	0.633	0.654	**0.692**
	stylistic	0.503	0.540	0.527
	lexicons	0.532	0.513	0.565
combinations	n-grams + styl.	0.635	0.655	0.691
	n-grams + lex.	0.649	0.657	0.663
	styl. + lex.	0.539	0.531	0.579
	n-grams + styl. + lex.	**0.654**	**0.660**	0.668

elongated words. Although the differences are not very large, this result suggests that these features may be used with different patterns depending more on the domain rather than on the entity.

Regarding the independent and the domain-dependent training scenarios, the best performance is obtained when all the features are combined. This result suggests that all the different feature groups contain some predictive power. However, in case of the entity-dependent scenario, the best performance is obtained with the n-grams feature group that contains the n-grams and character grams. This leads to the conclusion that terms are good indicators of polar fact tweets when the model is trained on the entity-dependent scenario.

For the rest of the experiments we use the model trained on n-grams, and character grams feature group on the entity-dependent scenario to differentiate between the reputation-neutral and polar facts tweets since this model obtained the best performance.

6.5.3 Direct Sentiment Propagation

In this section, we present the results of propagating sentiment signals directly to tweets that discuss about the same topic to estimate reputation polarity. First, we present the results before applying any threshold regarding the pairwise similarity and then we present the results after we have learned the best thresholds.

Regarding the research questions, we try to address the third and the fourth questions. For the third question: *Can we propagate sentiment to texts that are similar in terms of content to improve reputation polarity?* we apply the direct sentiment propagation approach. To address the last research question: *What is the best way to select the set of pairwise similar tweets before propagating the sentiment?* we explore propagating sentiment using three alternatives: without any threshold, learning the best threshold and with the *maxDelta* approach.

First, we apply the direct propagation approach without using any threshold[1]. Table 6.5 shows the results of the sentiment propagation approaches on the reputation polarity task using both the cluster and pairwise similarity. Overall, the results suggest that sentiment can be effectively propagated topically to annotate tweets with reputation. In particular, we observe that all the different combinations lead to statistically better results compared to the baseline. From the results, we observe that the approach that is based on the pairwise similarity performs better than the one based on clustering. Also, using the maximum frequency of tweets with a specific polarity is more efficient than using the aver-

[1] the threshold is 0

Table 6.5. Performance results (F-measure) of the sentiment propagation approaches using no threshold. A star(∗) indicates statistically significant improvement over the lexicon based approach. The Δ difference is given over the lexicon based approach.

method	frequency	avgCosine
lexicon-based (unsupervised)	0.368	0.368
cluster	0.472∗ (+28%)	0.457∗ (+24%)
pairwise similarity	0.526∗ (+43%)	0.495∗ (+35%)

Table 6.6. Performance results (F-measure) of the sentiment propagation approaches using no threshold, best threshold and maxDelta approaches. A star(∗) indicates statistically significant improvement over the lexicon based approach. The Δ difference is given over the lexicon based approach.

method	frequency	avgCosine
lexicon based (unsupervised)	0.368	0.368
pairwise similarity - no threshold	0.526∗ (+43%)	0.495∗ (+35%)
pairwise similarity - best threshold	0.529∗ (+44%)	0.499∗ (+36%)
pairwise similarity - maxDelta	0.524∗ (+42%)	0.488∗ (+33%)

age cosine similarity. The results suggest that the best approach is the one based on pairwise similarity and maximum frequency which improves the baseline by +43%. The results confirm the hypothesis that tweets that are about a similar topic tend to share the same reputation polarity.

In addition, we learn the best threshold regarding how many similar tweets need to be considered to determine the polarity that will be propagated. More specifically, we learn two different thresholds, one for the frequency and one for the avgCosine approach. We used the provided training data to learn the best threshold. Figure 6.3 shows the performance results on the training data after we applied different thresholds on the propagation approaches. The thresholds refer to the similarity score. Although the similarity values are normalised, the majority of similarity scores are below the value of 0.5. Therefore, we applied different thresholds between $0 - 0.5$ with a step of 0.05. From the figure we observe that the best threshold for both approaches is 0.05. In addition, we observe that the performance decreases for thresholds that are greater than 0.05.

Table 6.6 shows the results of the sentiment propagation approaches on the

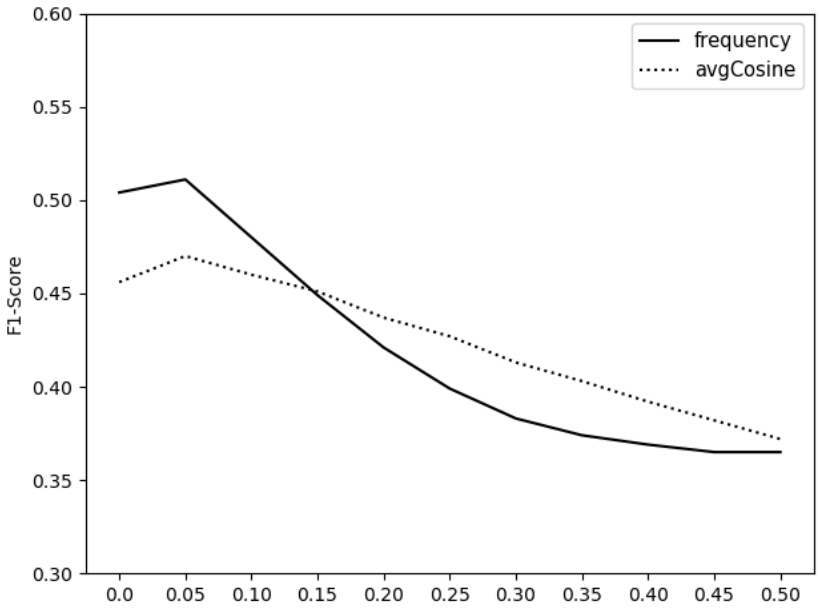

Figure 6.3. Performance scores using different thresholds on the training data for frequency and avgCosine approaches.

reputation polarity task using the different approaches on selecting the tweets to be considered for the propagation. The results show that all the different approaches managed to significantly improve the performance of the baseline. We observe that the best performance is obtained when we learned the best threshold using the training data. This result occurs for both the frequency and avgCosine approaches. However, the differences among the three different alternatives (no threshold, best threshold, maxDelta) is very small.

Table 6.7 compares the best results published until now for reputation polarity on the RepLab 2013 dataset (SVM trained on message and reception features and on an entity-dependent scenario) [122] with our best supervised and weakly-supervised approaches in terms of F-measure. The supervised approach based on PMI outperforms the best result of [122] with a 5.6% relative improvement in terms of F-measure (0.586 vs 0.553). This indicates that it is not necessary to use many features to get competitive results in reputation polarity. We also see that fully supervised approaches outperform weakly supervised ones. Our best weakly supervised approach is the one based on propagation to similar tweets using max combination and no threshold[2], however, is only 5% worse than [122]

[2]the best threshold approach is supervised

Table 6.7. Comparison with the state-of-the-art results.

Method	F-Measure
Peetz et al. 2016 (Best published result)	0.553
Supervised - PMI & Entity Dependent	0.586
Weakly Supervised - Propagation (pairwise similarity & frequency)	0.526

(0.526 vs 0.553). This small difference indicates that weakly supervised annotation of reputation polarity is feasible, which is a promising result as such methods are less dependent on the availability of training data.

6.6 Analysis and Discussion

In this section we perform a more detailed analysis of the results. First, we perform a post-evaluation analysis regarding the number of tweets per reputation polarity and per topic. Then we focus on the polar fact filter and we try to detect the cases in which the polar fact filter failed to differentiate between reputation-bearing and reputation-neutral tweets.

6.6.1 Reputation Polarity per Topic

First, we perform a post-evaluation analysis regarding the reputation polarity of tweets per topic. We use manually annotated topics provided with the training data of the collection. Figure 6.4 shows the distribution of reputation polarity per topic for a 100 randomly selected topics in the training set. The figure suggests that the majority of tweets that discuss the same topics tend to have the same reputation polarity. This happens for all the three reputation polarities (positive, neutral and negative). More importantly, we observe that there are topics for which all the tweets share the same reputation polarity (e.g., topic 20). As further analysis, Figure 6.5 shows the frequency of topics over the percentage of tweets that belong to the same reputation polarity. Regarding the reputation polarity, we consider the ones that contain the largest number of tweets. For example, if a topic contains {25, 12, 3} positive, neutral and negative tweets respectively, then the percentage is calculated as $max(25, 12, 3)/sum(25, 12, 3) = 62.5\%$. From the figure we observe that more than 2,500 topics contain tweets that belong to a single reputation polarity (i.e. percentage is 1.0). This observation supports our initial assumption that tweets that discuss the same topic tend to share the

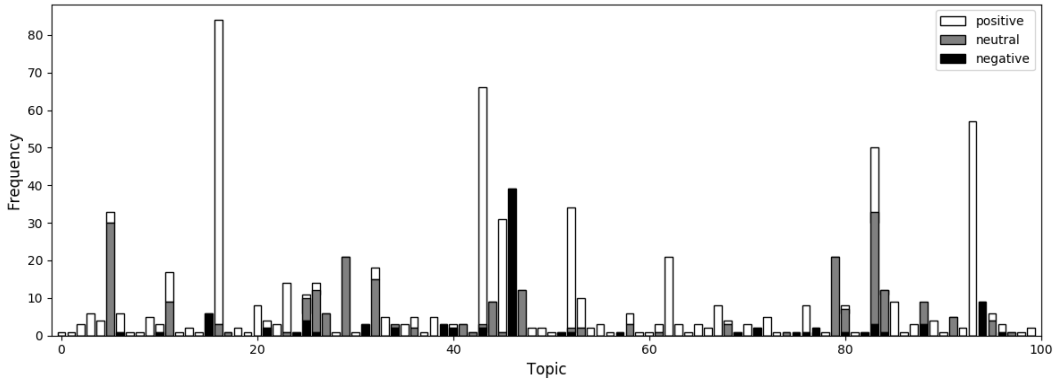

Figure 6.4. Frequency of tweets per reputation polarity and per topic for 100 randomly selected topics.

same reputation polarity.

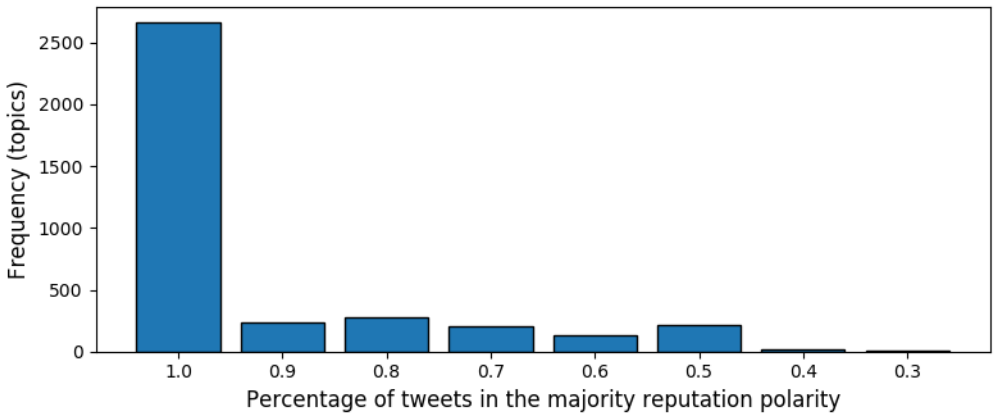

Figure 6.5. Frequency of topics per the percentage of tweets that belong to the most popular reputation polarity.

6.6.2 Polar Fact Filter Failure Analysis

In this section, we do a more detailed analysis on the results of the polar fact filter to understand the cases that the polar fact did not manage to differentiate between the polar facts and the reputation-neutral tweets. A manual analysis showed that the top five lowest performances are obtained for entities from the automotive domain (i.e., *Mazda*, *Yamaha*, *Subaru*, *Honda* and Ford). In addition, the entity *Whitney Houston* from the music domain achieved a low performance

in predicting the polar facts since the majority of polar facts were classified as neutral.

A manual analysis showed that the difficult cases are the tweets that have spelling mistakes and use non-conventional language. For example, the tweet *in my subaru mixing with my car speakers looool* was misclassified as neutral because the only sentiment indicator was written in emphatic lengthening (i.e., looool). Also some polar facts that were wrongly classified as neutral are very short tweets such as *#My100Wishes. Get a Mazda RX 8 one day* or *#Top10GreatestVoices Whitney Houston.*. If these words do not appear in the training data then it is very difficult to predict the reputation polarity. On the other hand, the polar facts that were correctly predicted seem to contain a better structure, such as *Stanford moves ahead with plans to radically change humanities doctoral education* or *Whitney Houston's voice is pure*.

6.7 Conclusions

In this chapter we have presented an effective methodology to estimate the reputation polarity of tweets. Starting from the observation that similar tweets share the same reputation polarity, we explored the effectiveness of propagating sentiment signals to tweets that are about the same topic. We consider two different lexicon augmentation approaches that learn reputation polarity words. The lexicon augmentation approaches were evaluated on unsupervised and supervised settings and on three different granularity levels, independent, domain-dependent and entity-dependent. In addition, we proposed a direct sentiment propagation approach. We explored different approaches of estimating the similarity of tweets and of selecting the set of similar tweets. Also, we explored two different ways of learning the propagated sentiment. Finally, we explored the effectiveness of a polar fact filter that is able to differentiate between polar facts and reputation-neutral tweets.

The results of our experiments strongly support our initial hypothesis: sentiment signals can be used to annotate reputation polarity, starting with sentiment-bearing texts and propagating sentiment to sentiment-neutral similar texts. Augmenting the sentiment lexicon improves results up to 25% if we generate a specific lexicon for each entity of interest. But, remarkably, generating domain-dependent lexicons (which requires less training material) gives very similar results (24% improvement over the original sentiment lexicon). The conclusion is that sentiment lexicons can be augmented to create reputation polarity lexicons, and that the domain level is a cost-effective level of granularity for doing so.

If we use a fully supervised approach to learn reputation polarity words (based on PMI scores), performance is 5.6% better than the best published result on the dataset so far [122]. This indicates that learning PMI values to predict reputation polarity is very effective.

Direct propagation of sentiment is also effective. In all conditions, the improvement is above 20% with respect to the no propagation baseline and for the best weakly-supervised setting (propagating to similar tweets using the max approach) the improvement is +43%. This approach is weakly supervised, because both the initial sentiment annotation and the propagation are unsupervised; the only supervised mechanism is the polar fact filter that prevents propagation to truly neutral tweets. Results, however, are only 5% worse than [122] (0.526 vs 0.553), which is a fully supervised approach. This small difference indicates that weakly supervised annotation of reputation polarity is feasible, which is a promising result as such methods are less dependent on the availability of training data. Finally, regarding the different ways of learning a threshold, we did not find any differences among the three approaches that we explored. In particular, learning the best threshold performs better than no threshold and maxDelta but the improvement is very small.

Chapter 7

Conclusions

7.1 Summary

The main motivation of this thesis was to address the problem of tracking opinion on social media towards a topic. To address this problem, we focused on posts that are published on Facebook and Twitter. The first step of tracking opinion on social media is to understand if there is an opinion expressed in a post and if the expressed opinion is about the topic of interest. To address this aspect, we focused on the posts that are published in Twitter and which are known as tweets. We first analysed the topical distribution of tweets to understand if they are about a single topic. Then, we focused on Twitter opinion retrieval and we proposed a topic specific stylistic approach to retrieve tweets that are relevant to a topic and also express opinion about it. Next, we applied state-of-the-art time series to find patterns and trends in opinion and we explored their effectiveness in forecasting opinion in the future. In addition, we developed a methodology for extracting sentiment spikes and we proposed a methodology to extract and rank the likely reasons that may have caused the sentiment spikes. To estimate the number of people that support a certain opinion, we proposed and evaluated pre- and post-publication information signals. Finally, it is also important to know the impact of the posts that are published online. Therefore, we proposed to track sentiment signals propagation to measure the impact of public posts and opinion on an entity's reputation.

The rest of this chapter is organised as follows. In Section 7.2 we revisit our research questions and provide answers to each of them. In Section 7.3 we list the future research directions following from this thesis.

7.2 Answers to Research Questions

In Chapter 3 we focused on the task of finding tweets that are relevant to a topic and also express opinion about it. We addressed the following general question:

RQ1 How can we find documents that are opinionated and express opinion about a topic in a microblogging collection? Can we make use of the textual peculiarities that are present in posts such as tweets to improve Twitter opinion retrieval?

This general research question led to the following detailed questions:

RQ1.1 How many topics are discussed in a single tweet?

To address this question, we applied an LDA topic model on a collection of tweets and blogs and compared the percentage of documents with one topic. We found that the majority of tweets is about a single topic in contrast to blogs that are usually about more than one topics. This implies that if a tweet is opinionated then it is very likely that it will be opinionated for this topic.

RQ1.2 What is the most effective combination of stylistic variations regarding topic-specific Twitter opinion retrieval?

We explored the effectiveness of different combinations of stylistic variations. We found that all the examined combinations of the stylistic variations (*emoticons*, *exclamation marks*, *emphatic lengthening* and *opinionated hashtags*) managed to perform significantly better on both the relevance and opinion baselines. Also, we found that the best performance is achieved with *emoticons*, *exclamation marks* and *emphatic lengthening*.

RQ1.3 Is the importance of stylistic variations in indicating opinion topic dependent?

To address this question, we calculated the effectiveness of the topic specific stylistic model under two settings: (i) without information from tweets' topics and (ii) using information from tweets' topics. We explored different combinations in calculating frequencies and inverse frequencies. We found that most of the combinations perform statistically better under the topic-based settings compared to the non topic-based settings. This shows that

stylistic variations are indeed topic-specific and the amount of the opinion information they hold depends on the topic of the tweet. We also found that there is no statistical difference between the different combinations of stylistic variations when they are compared under the same settings.

In Chapter 4 we focused on tracking opinion over time and we explored the following research question:

RQ2 How can we model opinion evolution and identify the important causes of opinion change?

This general research question led to the following detailed questions:

RQ2.1 Can conventional time series methods be applied to track sentiment evolution over time and forecast sentiment in the future?

We applied state-of-the-art time series tools including frequency analysis and data decomposition to track sentiment evolution over time. We found that conventional time series methods can be applied to track sentiment evolution over time. More specifically, they are appropriate to plot signals that show a topic's popularity and sentiment evolution towards the topic under examination. In addition, useful observations can be obtained by plotting the seasonality and sentiment velocity or acceleration. Regarding forecasting sentiment, we compared the effectiveness of naïve, mean and ARIMA forecasting tools and we showed that in some cases the naïve that is a simple forecasting approach can outperform ARIMA that is a more sophisticated approach.

RQ2.2 Can outlier detection be applied to identify sentiment spikes?

We applied outlier detection on positive and negative sentiment evolution and we found that outlier detection is a very useful tool that can be used to help us extract the sentiment spikes. We used the outlier detection tool to extract several sentiment spikes regarding different entities.

RQ2.3 How does an approach based on a combination of topic model with KL-divergence perform in extracting the likely reasons that caused a sentiment spike?

We found that combining topic model with KL-divergence can be used to

effectively extract the likely reasons that caused a sentiment spike. In addition, we explored and discussed challenging cases that the topic model with KL-divergence approach did not agree with the annotators and this contributed to a better understanding of the data.

In Chapter 5 we focused on estimating how many people support a specific opinion and we explored the following research question:

RQ3 How can we predict how many people will react with a specific emotion when a news post is published?

This general research question led to the following detailed questions:

RQ3.1 Can we improve the effectiveness of baseline classifiers by adding additional pre-publication information based on news post content?

We explored the effectiveness of various pre-publication features extracted from the content of the news post on predicting the emotional reactions. We compared the results with two baselines: a weak that is based on the date and a strong that is based on the post's terms. We found that pre-publication features based on the news posts' content could reliably improve the results of the weak baseline, but not of the strong baseline. Also, we showed that the most effective pre-publication feature for predicting the triggered emotional reactions is the terms of the post.

RQ3.2 Can we improve the effectiveness of baseline classifiers by adding additional post-publication information extracted from users' comments?

We explored the effectiveness of various post-publication features extracted from the users' comments that are published below the post. We tried different time ranges regarding the publication of the comment. We found that the majority of runs that use all the early commenting features (i.e., post-publication information) perform statistically better compared to the ones trained on the terms of the post (i.e., strong baseline).

RQ3.3 How does a model that combines textual and early commenting features perform?

We combined textual and early commenting features and we showed that this combination can effectively predict the number of triggered emotional

reactions of users. We found that the combination of all the features leads to statistically better results compared to the baselines. Also, we found that the improvement is not consistent for all the reactions and this may have to do with the terms that are used in the posts and which convey emotion.

RQ3.4 What is the added value of the commenting features in terms of effectiveness in the task of emotional reactions prediction?

We analysed the effectiveness of the commenting features for the task of emotional reactions prediction. We found that added value of the commenting features is not consistent across the reactions. Regarding sadness and anger most added value came from negative ratio, whereas in case of love, surprise and joy it came from number of comments and number of unique authors.

In Chapter 6 we wanted to understand the impact of posts that are published online on an entity's reputation. We asked the following research question:

RQ4 How can we estimate the impact of posts on an entity? Can we use sentiment signals propagation to estimate the impact of posts on an entity's reputation?

The last general research question led to the following detailed questions:

RQ4.1 Can we use training material to detect terms with reputation polarity and use them to augment a general sentiment lexicon?

We explored two alternatives for augmenting lexicon approaches: (i) using an unsupervised way where the initial sentiment is obtained using a state-of-art sentiment lexicon; and (ii) using data that were manually annotated by reputation experts. We found that augmenting the sentiment lexicon in a supervised setting significantly improves results compared to the baseline for all the different levels of granurality. In addition, using the provided training data is more effective compared to unsupervised setting. The conclusion is that sentiment lexicons can be augmented to create reputation polarity lexicons. If we use the PMI approach to learn reputation polarity words, performance is 5.6% better than the best published result on the dataset so far [122]. This indicates that learning PMI values to predict reputation polarity is very effective.

RQ4.2 What is the right level of generalization for a reputation lexicon?

To answer this question we explored three alternatives: (i) building a general purpose lexicon with all available training material; (ii) building domain-specific lexicons with training material for entities in a given domain (e.g., banking, automotive); (iii) building entity-specific lexicons with separated training material for each entity. We found that augmenting the sentiment lexicon improves results up to 25% if we generate a specific lexicon for each entity of interest. Remarkably, generating domain-dependent lexicons (which requires less training material) gives very similar results to entity-dependent lexicons (24% improvement over the original sentiment lexicon). Therefore, regarding the lexAugm approach, the conclusion is that the domain level is a cost-effective level of granularity for augmenting a sentiment lexicon to create a reputation lexicon. However, regarding the PMI approach, the performance difference between the domain and the entity-dependent lexicons is considerable. Regarding PMI, the results suggest that the entity-dependent expansion is the most effective level of granularity.

RQ4.3 Can we propagate sentiment to text that is similar in terms of content to improve reputation polarity?

In order to answer this question we considered two alternatives: (i) first perform text clustering to detect topics, and then propagate sentiment to tweets that belong to each topic; (ii) directly propagate sentiment from a sentiment-bearing text to other texts that are pairwise similar. In addition, we also experimented with the use of a polar fact filter to avoid over-propagation to polarity-wise neutral texts. We found that direct propagation of sentiment signals is effective. In all conditions, the improvement is above 20% with respect to the no propagation baseline, and for the best weakly supervised setting (i.e., propagating to similar tweets using the max approach) the improvement is +43%. The best approach is the one that is based on learning the best threshold (supervised) that outperforms the baseline by +44%.

RQ4.4 What is the best way to select the set of pairwise similar tweets that can be used to learn the sentiment that will be propagated?

To answer this question, we explored the following alternatives: (i) using all the pairwise similar tweets; (ii) learning a threshold; and (iii) using the

maximum difference of the pairwise similarities. We also explored the effectiveness of two different approaches regarding learning the propagated polarity, the first is based on the frequencies of sentiment and the second on the average pairwise similarity. We found that the best approach to decide the set of tweets that will be used for estimating the propagated polarity is the one based on the best threshold. However, the difference among the three explored alternatives is very small.

7.3 Future Research Directions

This thesis has resulted in several lessons for tracking public opinion in social media. In the following we lay out future research directions, in particular on use of deep learning, multimodal sentiment analysis, emotional reactions of users, temporal sentiment propagation for reputation analysis, forecast sentiment spikes and multidisciplinary research.

Use of deep learning. The majority of prior work on opinion and sentiment analysis uses state-of-the art machine learning approaches. One of the most important limitations of machine-learning approaches is that their effectiveness depends on the set of selected features that are usually hand-crafted.

Recently, deep learning approaches have emerged and have been applied to many state-of-the-art NLP and IR problems. Deep learning is the application of artificial neural networks to learning tasks using networks of multiple layers. A number of studies tried to address sentiment analysis and related challenges with deep learning approaches [92]. For example, deep learning has been applied for estimating the emotion intensity in a text [65] or aspect extraction [126]. One limitation of the deep learning approaches is that the results are not easy to explain in contrast to the state-of-the art machine learning approaches.

There are still open questions that need to be addressed regarding the application of deep learning on sentiment analysis. One interesting direction would be to examine the effectiveness of the convolutional or recursive neural network algorithms on opinion analysis and negation handling in tweets. In addition, it is worth to explore the effectiveness of deep learning approaches on challenging tasks such as hate speech and irony detection.

Multimodal sentiment analysis. The increasing popularity of video sharing social media platforms such as YouTube suggests that different modalities of information (e.g., text, images, audio, video) should be combined to effectively

address opinion analysis. Multimodal sentiment analysis is an emerging field at the intersection of natural language processing, computer vision, and speech processing. Affective traces, such as facial and vocal expressions can be captured from videos, and be leveraged in addition to the textual content to address sentiment analysis.

The application of multimodal sentiment analysis approach for humour or irony detection is also still an open problem. Irony and humour detection is an emerging field that has attracted attention [130, 71]. However, these problems are still a challenge because they are culture-specific. Therefore, multimodal information can be useful in identifying the sarcastic comments by taking advantage of vocal and facial expressions.

Emotional reactions of users. Future models on emotional reactions prediction should be able to predict the actual number of the emotional reactions that are triggered by a post. For example, a regression approach can be used to predict the number of emotional reactions. Another direction is to modify existing deep learning approaches such as convolutional neural network and investigate their effectiveness on the specific problem. These approaches should take into account the temporal evolution of words and emotional reactions. Emotional reactions can be then used as an additional information for other problems such as fake news detection.

Temporal sentiment propagation for reputation analysis. The set of terms that have a reputation impact changes over time and depends on the topics that are being discussed. For example, the tweet *"Brad Pitt playing around with his oldest child. So adorable"* has a positive impact on Brad Pitt whereas the tweet *"I can't believe that Brad Pitt is accused of child abuse"* has a negative impact. To this end, we can assume that the word *child* is an additional indicator of positive reputation polarity for tweets that are temporally adjacent to the first example. For tweets that are temporally adjacent to the second tweet, the word *child* is an additional indicator of negative reputation. One possible direction, is to explore the effectiveness of incorporating a temporal aspect in augmenting sentiment lexicons and propagating sentiment signals based on temporal windows. Another interesting direction is to study the effect of using the sentiment spikes as an indicator for creating temporal windows.

Forecast sentiment spikes. An interesting extension to our work is forecasting sentiment spikes in future. One way to achieve this is to track how sentiment towards an entity changes over time and detect indicators that are related to

sentiment spikes. Some examples of these indicators are specific terms, number of tweets, emotion expressed in tweets or users who post messages. For example, since some of these spikes are related to news events, it would be interesting to examine if information from news events (e.g., type of event, popularity of event) can be useful to predict sentiment spikes that may occur in a short period of time. For instance, natural disasters (e.g., fires, earthquakes) that are discussed a lot in news have a high probability to cause sentiment spikes that are related to political parties. Another useful information is the seasonality that is related to entities that have a clear seasonal component (e.g., TV series). In such cases information from historical data is worth to be examined since they can be proved to be important in predicting sentiment spikes.

An additional way to predict the sentiment spikes is to explore if there are any users whose posts are related to the spikes. In social media, there are some users that have many followers or users who post a lot of messages with regards to a specific entity. Hence, it would be interesting to estimate the probability of a sentiment spike given the tweets of some users given the content of these tweets.

Multidisciplinary research. The combination of research from different fields is still under-explored. Applying sentiment analysis methods on economics research or on human and social science domains can yield interesting results. For example, it is possible to explore how geographic places or meteorological variables and events influence the level of happiness within a society [104]. Additionally, sentiment analysis could be applied on a marketing domain to predict the success of a product or of movies [16]. Another interesting direction would be to apply sentiment analysis on health domains to explore how emotions correlate with well being or with health in general.

Bibliography

[1] Differences between social media and social networking. https://www.socialmediatoday.com/content/5-differences-between-social-media-and-social-networking, 2010.

[2] Statista 2018. Number of social network users worldwide from 2010 to 2021 (in billions). https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, 2018.

[3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, 2011.

[4] D. Agarwal, B.-C. Chen, and X. Wang. Multi-faceted ranking of news articles using post-read actions. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 694–703, 2012.

[5] C. C. Aggarwal. *An Introduction to Social Network Data Analytics*, pages 1–15. Springer, 2011. ISBN 978-1-4419-8462-3.

[6] F. Alam, F. Celli, E. A. Stepanov, A. Ghosh, and G. Riccardi. The social mood of news: Self-reported annotations to design automatic mood detection systems. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, PEOPLES '16, pages 143–152, 2016.

[7] J. C. D. Albornoz, I. Chugur, and E. Amigó. Using an emotion-based model and sentiment analysis techniques to classify polarity for reputation. In *Proceedings of CLEF 2012: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '12, 2012.

[8] M. Aliannejadi and F. Crestani. Venue suggestion using social-centric scores. *CoRR*, abs/1803.08354, 2018.

[9] M. Aliannejadi, S. A. Bahrainian, A. Giachanou, and F. Crestani. University of Lugano at TREC 2015: Contextual suggestion and temporal summarization tracks. In *Proceedings of the 24th Text REtrieval Conference*, TREC 2015, 2015.

[10] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. Automatic construction of an opinion-term vocabulary for ad hoc retrieval. In *Proceedings of the 30th European Conference on Information Retrieval Research*, ECIR '08, pages 89–100, 2008.

[11] G. Amati, M. Bianchi, and G. Marcone. Sentiment estimation on twitter. In *Proceedings of the 5th Italian Information Retrieval Workshop*, IIR '14, pages 39–50, 2014.

[12] E. Amigó, A. Corujo, J. Gonzalo, E. Meij, and M. D. Rijke. Overview of RepLab 2012: Evaluating Online Reputation Management Systems. In *Proceedings of CLEF 2012: Conference on Multilingual and Multimodal Information Access Evaluation*, 2012.

[13] E. Amigó, J. C. D. Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Mart, E. Meij, M. D. Rijke, and D. Spina. Overview of RepLab 2013: Evaluating online reputation monitoring systems. In *Proceedings of CLEF 2013: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '13, pages 333–352, 2013.

[14] X. An, R. A. Ganguly, Y. Fang, B. S. Scyphers, M. A. Hunter, and G. J. Dy. Tracking climate change opinions from twitter data. In *Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining: Workshop on Data Science for Social Good,* KDD '14, 2014.

[15] I. Arapakis, B. B. Cambazoglu, and M. Lalmas. On the feasibility of predicting popular news at cold start. *Journal of the Association for Information Science and Technology*, 68(5):1149–1164, 2017.

[16] S. Asur and B. A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, 2010.

[17] S. A. Bahrainian, I. Mele, and F. Crestani. Modeling discrete dynamic topics. In *Proceedings of the 32nd Symposium on Applied Computing*, SAC '17, pages 858–865, 2017.

[18] A. Balahur and H. Tanev. Detecting entity-related events and sentiments from tweets using multilingual resources. In *Proceedings of CLEF 2012: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '12, 2012.

[19] K. Balog, G. Mishne, and M. de Rijke. Why are they excited?: Identifying and explaining spikes in blog mood levels. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstration*, EACL '06, pages 207–210, 2006.

[20] R. Bandari, S. Asur, and B. A. Huberman. The pulse of news in social media: Forecasting popularity. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, ICWSM '12, pages 26–33, 2012.

[21] G. Barbier and H. Liu. *Data Mining in Social Media*, pages 327–352. Springer, 2011. ISBN 978-1-4419-8462-3.

[22] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6): 1554–1563, 1966.

[23] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning research*, 3:993–1022, 2003.

[24] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, 2006.

[25] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[26] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[27] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and regression trees*. Wadsworth, 1984. ISBN 9780412048418.

[28] S. Brody and N. Diakopoulos. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 562–570, 2011.

[29] A. Castellanos, J. Cigarrán, and A. García-serrano. Modelling techniques for twitter contents: A step beyond classification based approaches. In *Proceedings of CLEF 2013: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '13, 2013.

[30] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. ISSN 2157-6904.

[31] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International World Wide Web Conference*, WWW '14, pages 925–936, 2014.

[32] J. M. Chenlo, J. Atserias, C. Rodriguez, and R. Blanco. FBM-Yahoo! at RepLab 2012. In *Proceedings of CLEF 2012: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '12, 2012.

[33] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[34] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[35] J. Clos, A. Bandhakavi, N. Wiratunga, and G. Cabanac. Predicting emotional reaction in social networks. In *ECIR '17*, pages 527–533, 2017.

[36] J.-V. Cossu, B. Bigot, L. Bonnefoy, M. Morchid, X. Bost, G. Senay, R. Dufour, V. Bouvier, J.-M. Torres-Moreno, and M. El-Bèze. LIA@RepLab 2013. In *Proceedings of CLEF 2013: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '13, 2013.

[37] W. B. Croft, D. Metzler, and T. Strohman. *Search engines: Information retrieval in practice*, volume 283. Addison-Wesley Reading, 2010.

[38] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, 2010.

[39] L. De Vries, S. Gensler, and P. S. Leeflang. Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of Interactive Marketing*, 26(2):83–91, 2012.

[40] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 231–240, 2008.

[41] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL '14, pages 49–54, 2014.

[42] S. Edosomwan, S. K. Prakasan, D. Kouame, J. Watson, and T. Seymour. The history of social media and its impact on business. *Journal of Applied Management and Entrepreneurship*, 16(3):79–91, 07 2011.

[43] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3/4): 169–200, 1992.

[44] J. Filgueiras and S. Amir. POPSTAR at RepLab 2013: Polarity for reputation classification. In *Proceedings of CLEF 2013: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '13, 2013.

[45] P. Fu, C. Wu, and Y. Cho. What makes users share content on facebook? Compatibility among psychological incentive, social capital focus, and content type. *Computers in Human Behavior*, 67:23–32, 2017.

[46] N. Fuhr, A. Giachanou, G. Grefenstette, I. Gurevych, A. Hanselowski, K. Jarvelin, R. Jones, Y. Liu, J. Mothe, W. Nejdl, et al. An information nutritional label for online documents. *ACM SIGIR Forum*, 51(3):46–66, 2018.

[47] W. Gao and F. Sebastiani. Tweet sentiment: From classification to quantification. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 97–104, 2015.

[48] S. Gerani, M. Carman, and F. Crestani. Aggregation methods for proximity-based opinion retrieval. *ACM Transactions on Information Systems (TOIS)*, 30(4):1–36, 2012.

[49] A. Giachanou and F. Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):28, 2016.

[50] A. Giachanou and F. Crestani. Opinion retrieval in twitter: Is proximity effective? In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, SAC '16, pages 1146–1151, 2016.

[51] A. Giachanou and F. Crestani. Opinion retrieval in twitter using stylistic variations. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, SAC '16, pages 1077–1079, 2016.

[52] A. Giachanou and F. Crestani. Tracking sentiment by time series analysis. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1037–1040, 2016.

[53] A. Giachanou, I. Markov, and F. Crestani. Opinions in federated search: University of Lugano at TREC 2014 federated web search track. In *Proceedings of the 23rd Text REtrieval Conference*, TREC 2014, 2014.

[54] A. Giachanou, M. Harvey, and F. Crestani. Topic-specific stylistic variations for opinion retrieval on twitter. In *Proceedings of the 38th European Conference on Information Retrieval Research*, ECIR '16, pages 466–478, 2016.

[55] A. Giachanou, I. Mele, and F. Crestani. Explaining sentiment spikes in twitter. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2263–2268, 2016.

[56] A. Giachanou, J. Gonzalo, I. Mele, and F. Crestani. Sentiment propagation for predicting reputation polarity. In *Proceedings of the 39th European Conference on Information Retrieval Research*, ECIR '17, pages 226–238, 2017.

[57] A. Giachanou, I. Mele, and F. Crestani. USI participation at SMERP 2017 text retrieval task. In *Proceedings of the 1st Exploitation of Social Media for Emergency Relief and Preparedness Workshop (Data Challenge Track)*, SMERP@ECIR '17, 2017.

[58] A. Giachanou, I. Mele, and F. Crestani. USI participation at SMERP 2017 text summarization task. In *Proceedings of the 1st Exploitation of Social Media for Emergency Relief and Preparedness Workshop (Data Challenge Track)*, SMERP@ECIR '17, 2017.

[59] A. Giachanou, I. Mele, and F. Crestani. A collection for detecting triggers of sentiment spikes. In *Proceedings of the 40th International ACM SIGIR*

Conference on Research and Development in Information Retrieval, pages 1249–1252, 2017.

[60] A. Giachanou, F. Rangel, F. Crestani, and P. Rosso. Emerging sentiment language model for emotion detection. In *Proceedings of the 4th Italian Conference on Computational Linguistics*, CLiC-it '17, 2017.

[61] A. Giachanou, P. Rosso, I. Mele, and F. Crestani. Emotional influence prediction of news posts. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, ICWSM '18, pages 592–596, 2018.

[62] A. Giachanou, P. Rosso, I. Mele, and F. Crestani. Emotional reactions prediction of news posts. In *Proceedings of the 9th Italian Information Retrieval Workshop*, IIR '18, 2018.

[63] A. Giachanou, P. Rosso, I. Mele, and F. Crestani. Early commenting features for emotional reactions prediction. In *Proceedings of the 25th International Symposium on String Processing and Information Retrieval*, SPIRE '18, 2018.

[64] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. Technical report, Standford, 2009.

[65] P. Goel, D. Kulshreshtha, P. Jain, and K. K. Shukla. Prayas at EmoInt 2017: An ensemble of deep neural architectures for emotion intensity prediction in tweets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, WASSA@EMNLP 2017, pages 58–65, 2017.

[66] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[67] M. A. Greenwood, N. Aswani, and K. Bontcheva. Reputation profiling with GATE. In *Proceedings of CLEF 2012: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '12, 2012.

[68] J. Han, M. Kamber, and J. Pei. In *Data Mining: Concepts and Techniques*, chapter 3: Data Preprocessing. Morgan Kaufmann Publishers Inc., 2011. ISBN 0123814790, 9780123814791.

[69] V. Hangya and R. Farkas. Filtering and polarity detection for reputation management on tweets. In *Proceedings of CLEF 2013: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '13, 2013.

[70] Y. He, C. Lin, W. Gao, and K.-F. Wong. Dynamic joint sentiment-topic model. *ACM Transactions on Intelligent Systems and Technology*, 5(1):1–21, 2013.

[71] D. I. Hernańdez Farías, V. Patti, and P. Rosso. Irony detection in twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT)*, 16(3):19, 2016.

[72] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR'99*, pages 50–57, 1999.

[73] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *SIGKDD Workshop on SMA*, pages 80–88, 2010. ISBN 9781450302173.

[74] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, 2004.

[75] X. Huang and W. B. Croft. A unified relevance model for opinion retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 947–956, 2009.

[76] X. Huang, M. C. Smith, M. J. Paul, D. Ryzhkov, S. C. Quinn, D. A. Broniatowski, and M. Dredze. Examining patterns of influenza vaccination in social media. In *Proceedings of the AAAI Joint Workshop on Health Intelligence*, W3PHIAI, pages 4–5, 2017.

[77] D. J. Hughes, M. Rowe, M. Batey, and A. Lee. A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage. *Computers in Human Behavior*, 28(2):561–569, 2012.

[78] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the Association for Information Science and Technology*, 60(11):2169–2188, 2009.

[79] H. Jeong and H. Lee. Using feature selection metrics for polarity analysis in RepLab 2012. In *Proceedings of CLEF 2012: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '12, 2012.

[80] Y. Jiang, W. Meng, and C. Yu. Topic sentiment change analysis. In *Proceedings of the 7th International Conference on Machine Learning and Data Mining in Pattern Recognition*, MLDM '11, pages 443–457, 2011.

[81] G. J. Kang, S. R. Ewing-Nelson, L. Mackey, J. T. Schlitt, A. Marathe, K. M. Abbas, and S. Swarup. Semantic network analysis of vaccine sentiment in online social media. *Vaccine*, 35(29):3621–3638, 2017.

[82] N. Kanhabua and K. Nørvåg. Improving temporal language models for determining time of non-timestamped documents. In *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '08, pages 358–370, 2008.

[83] J. M. Kapp, C. Peters, and D. P. Oliver. Research recruitment using facebook advertising: Big potential, big challenges. *Journal of Cancer Education*, 28 (1):134–137, 2013.

[84] R. Kaptein. Learning to analyze relevancy and polarity of tweets. In *Proceedings of CLEF 2012: Conference on Multilingual and Multimodal Information Access Evaluation*, 2012.

[85] J. Karlgren and L. Ericsson. Semantic space models for profiling reputation of corporate entities notebook for RepLab at CLEF 2013. In *Proceedings of CLEF 2013: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '13, 2013.

[86] J. Karlgren, M. Sahlgren, F. Olsson, F. Espinoza, and O. Hamfors. Profiling reputation of corporate entities in semantic space. In *Proceedings of CLEF 2012: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '12, 2012.

[87] S. Kiritchenko, X. Zhu, and S. M. Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.

[88] E. Kouloumpis, T. Wilson, and J. Moore. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 538–541, 2011.

[89] R. Krestel, P. Fankhauser, and W. Nejdl. Latent dirichlet allocation for tag recommendation. In *RecSys'09*, pages 61–68, 2009.

[90] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 03 1951.

[91] K. Kurniawati, G. G. Shanks, and N. Bekmamedova. The business impact of social media analytics. In *Proceedings of the 21st European Conference on Information Systems*, ECIS '13, 2013.

[92] Z. Lei, W. Shuai, and L. Bing. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.

[93] B. Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[94] A. Lomi, A. Giachanou, F. Crestani, and S. Angelopoulos. Table for two: Explaining variations in the evaluation of authenticity by restaurant critics. In *Proceedings of the 12th Organization Studies Workshop*, 2018.

[95] Z. Luo, M. Osborne, and T. Wang. An effective approach to tweets opinion retrieval. *World Wide Web*, 18(3):545–566, 2015.

[96] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 142–150, 2011.

[97] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 blog track. In *Proceedings of The 16th Text REtrieval Conference*, TREC 2007, 2007.

[98] W. G. Mangold and D. J. Faulds. Social media: The new hybrid element of the promotion mix. *Business Horizons*, 52(4):357–365, 2009.

[99] C. D. Manning, P. Raghavan, and H. Schütze. Scoring, term weighting and the vector space model. In *Introduction to Information Retrieval*. Cambridge University Press, 2008. ISBN 0521865719, 9780521865715.

[100] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[101] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 171–180, 2007.

[102] I. Mele, S. A. Bahrainian, and F. Crestani. Linking news across multiple streams for timeliness analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 767–776, 2017.

[103] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, NIPS '13, pages 3111–3119, 2013.

[104] L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, 8(5): e64417, 2013.

[105] S. Mohammad, S. Kiritchenko, and X. Zhu. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the 7th International Workshop on Semantic Evaluation*, SemEval@NAACL-HLT 2013, pages 321–327, 2013.

[106] S. M. Mohammad. #Emotional tweets. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation*, SemEval '12, pages 246–255, 2012.

[107] S. M. Mohammad and F. Bravo-Marquez. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, WASSA@EMNLP 2017, 2017.

[108] S. M. Mohammad and P. D. Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 26–34, 2010.

[109] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[110] C. S. Montero, H. Haddad, M. Mozgovoy, and C. B. Ali. Detecting the likely causes behind the emotion spikes of influential twitter users. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics*, CICLing '16, 2016.

[111] A. Mosquera, J. Fernández, J. M. Gómez, P. Martínez-Barco, and P. Moreda. DLSI-Volvam at RepLab 2013: Polarity classification on twitter data. In *Proceedings of CLEF 2013: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '13, 2013.

[112] F. Å. Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis of microblogs. In *ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, pages 93–98, 2011.

[113] I. Ounis, C. Macdonald, M. D. Rijke, G. Mishne, and I. Soboroff. Overview of the TREC 2006 blog track. In *Proceedings of the 15th Text REtrieval Conference*, TREC 2006, 2006.

[114] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC-2008 blog track. In *Proceedings of the 17th Text REtrieval Conference*, TREC 2008, 2008.

[115] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the TREC-2011 microblog track. In *Proceedings of the 20th Text REtrieval Conference*, TREC 2011, 2011.

[116] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi. Sentiment analysis of twitter data for predicting stock market movements. In *Proceedings of the 2016 International Conference on Signal Processing, Communication, Power and Embedded System*, SCOPES '16, pages 1345–1350, 2016.

[117] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th on International Language Resources and Evaluation Conference*, LREC '10, pages 1320–1326, 2010.

[118] G. Paltoglou and K. Buckley. Subjectivity annotation of the microblog 2011 realtime adhoc relevance judgments. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 344–355, 2013.

[119] G. Paltoglou and A. Giachanou. *Opinion Retrieval: Searching for Opinions in Social Media*, pages 193–214. Springer International Publishing, 2014. ISBN 978-3-319-12511-4.

[120] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[121] M.-H. Peetz, M. D. Rijke, and A. Schuth. From sentiment to reputation ILPS at RepLab 2012. In *Proceedings of CLEF 2012: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '12, 2012.

[122] M. H. Peetz, M. De Rijke, and R. Kaptein. Estimating reputation polarity on microblog posts. *Information Processing and Management*, 52(2):193–216, 2015.

[123] V. Perez-Rosas, C. Banea, and R. Mihalcea. Learning sentiment lexicons in spanish. In *Proceedings of the 8th International Language Resources and Evaluation Conference*, LREC '12, 2012.

[124] R. Plutchik. Emotion: Theory, research, and experience: Vol. 1. theories of emotion. In *Approaches to Emotion*, pages 3–33. Academic press, 1980.

[125] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, 1998.

[126] S. Poria, E. Cambria, and A. Gelbukh. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108: 42 – 49, 2016. ISSN 0950-7051.

[127] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[128] M. Rafla, N. J. Carson, and S. M. DeJong. Adolescents and the internet: What mental health clinicians need to know. *Current Psychiatry Reports*, 16(9), 2014.

[129] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. ICWSM'10, pages 1–8, 2010.

[130] A. Reyes, P. Rosso, and T. Veale. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268, 2013.

[131] G. Rizzo and R. Troncy. NERD: Evaluating named entity recognition tools in the web of data. In *Proceedings of the Workshop on Web Scale Knowledge Extraction*, WEKEX' 11, 2011.

[132] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu. Empatweet: Annotating and detecting emotions on twitter. In *Proceedings of the 8th International Language Resources and Evaluation Conference*, LREC '12, pages 3806–3813, 2012.

[133] J. Saias. In search of reputation assessment: Experiences with polarity classification in RepLab 2013. In *Proceedings of CLEF 2013: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '13, 2013.

[134] H. Saif, Y. He, and H. Alani. Semantic sentiment analysis of twitter. In *Proceedings of the 11th international conference on The Semantic Web - Volume Part I*, ISWC '12, pages 508–524, 2012.

[135] G. Salton. *Automatic Information Organization and Retrieval*. McGraw Hill Text, 1968. ISBN 0070544859.

[136] G. Salton. *Automatic text processing: The transformation,analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN 0-201-12227-8.

[137] M. Sanicas. Scientists can vaccinate us against fake news. `https://www.weforum.org/agenda/2017/08/scientists-can-vaccinate-against-the-post-truth-era`, 2017.

[138] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1): 22–36, Sept. 2017.

[139] B. Shulman, A. Sharma, and D. Cosley. Predictability of popularity: Gaps between prediction and understanding. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*, ICWSM '16, pages 348–357, 2016.

[140] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a H1N1 pandemic. *PLOS ONE*, 6(5):1–10, 05 2011.

[141] A. N. Smith, E. Fischer, and C. Yongjian. How does brand-related user-generated content differ across youtube, facebook, and twitter? *Journal of Interactive Marketing*, 26(2):102–113, 2012.

[142] D. Spina, J. Carrillo-de Albornoz, T. Martin, E. Amigó, J. Gonzalo, and F. Giner. UNED online reputation monitoring team at RepLab 2013. In *Proceedings of CLEF 2013: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '13, 2013.

[143] D. Spina, J. Gonzalo, and E. Amigó. Learning similarity functions for topic detection in online reputation monitoring. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 527–536, 2014.

[144] W. Strunk. *The elements of style*. Penguin, 2007.

[145] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, Aug. 2010. ISSN 0001-0782.

[146] J. T. and N. R. M. The thesaurus approach to information retrieval. *American Documentation*, 9(3):192–197.

[147] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[148] S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, and X. He. Interpreting the public sentiment variations on twitter. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1158–1170, 2014.

[149] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14, pages 1555–1565, 2014.

[150] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[151] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[152] M. Tsagkias, W. Weerkamp, and M. De Rijke. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1765–1768, 2009.

[153] M. Tsytsarau and T. Palpanas. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514, 2012.

[154] P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 417–424, 2002.

[155] C. J. van Rijsbergen. *Information retrieval*. Butterworth, 1979. ISBN 0-408-70929-4.

[156] K. D. Varathan, A. Giachanou, and F. Crestani. Temporal analysis of comparative opinion mining. In *Proceedings of the 18th International Conference on Asian Digital Libraries*, ICADL '16, pages 311–322, 2016.

[157] K. D. Varathan, A. Giachanou, and F. Crestani. Comparative opinion mining: A review. *Journal of the Association for Information Science and Technology*, 68(4):811–829, 2017.

[158] C. L. Ventola. Social media and health care professionals: Benefits, risks, and best practices. *Pharmacy and Therapeutics*, 39(7):491, 2014.

[159] E. Villatoro-Tello, C. Rodríguez-Lucatero, C. Sánchez-Sánchez, and A. P. López-Monroy. UAMCLyR at RepLab 2013: Profiling task. In *Proceedings of CLEF 2013: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '13, 2013.

[160] J. Villena-Román, S. Lana-Serrano, C. Moreno-García, J. García-Morera, and J. C. González-Cristóbal. DAEDALUS at RepLab 2012 : Polarity classification and filtering on twitter data. In *Proceedings of CLEF 2012: Conference on Multilingual and Multimodal Information Access Evaluation*, CLEF '12, 2012.

[161] H. Waheed, M. Anjum, M. Rehman, and A. Khawaja. Investigation of user behavior on social networking sites. *PLOS ONE*, 12(2):1–19, 02 2017.

[162] C. Wang, D. Blei, and D. Heckerman. Continuous time dynamic topic models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, UAI '08, pages 579–586, 2008.

[163] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 424–433, 2006.

[164] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[165] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, 2005.

[166] C. Yang, S. Bhattacharya, and P. Srinivasan. Lexical and machine learning approaches toward online reputation management. In *Proceedings of CLEF 2012: Conference on Multilingual and Multimodal Information Access Evaluation*, 2012.

[167] K. Yang, N. Yu, A. Valerio, H. Zhang, and W. Ke. Fusion approach to finding opinions in blogosphere. pages 1–8, 2007.

[168] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *SIGKDD'09*, pages 937–946, 2009.

[169] N. Zagorski. Using many social media platforms linked with depression, anxiety risk. *Psychiatric News*, 52, 2017.

[170] M. Zhang and X. Ye. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 411–418, 2008.

[171] W. X. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 338–349, 2011.