Understanding the Topics and Opinions from Social Media Content

by

Yiwei Zhou

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Supervised by: Dr. Alexandra I. Cristea

Department of Computer Science

July 2017

Contents

List of Tables

List of Figures

Acknowledgments

This thesis would not be possible without the support of many people. First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Alexandra I. Cristea. She provided me invaluable advice and feedback to guide me through the study. Her continuous support and encouragement give me strength to explore the challenges and help me through the difficulties.

I would like to extend my gratitude to my committee members, Prof. Rob Procter and Prof. Leslie Carr, for their time and insightful comments. Special thanks to my advisors, Dr. Maria Liakata and Prof. Rob Procter, for providing me helpful suggestions; my co-authors, Dr. Elena Demidova and Dr. Nattiya Kanhabua, for their inspirations and directions.

I spent one summer in the R&D Laboratory at the Financial Intelligence Unit in HSBC. I am grateful to Dr. Ilya Zheludev for his supervision and help during this enjoyable experience. I acquired a significant amount of knowledge that directly contributes to my study and future career.

I must thank the previous and current members of the Intelligent and Adaptive Systems group and the Human Centred Computing division in the Computer Science department of University of Warwick. These brilliant people helped me a lot both in my PhD study and in my personal life.

I was very lucky to spend some time in two prestigious research institutes during my PhD study, the L3S research centre and the Alan Turing Institute. I am very grateful to have the chances to meet many top minds in my research area, which have greatly broadened my view and inspired my ideas. I would like to thank the colleagues in these two research institutes for offering me such wonderful experiences.

I also want to thank the Warwick Graduate School for offering me the Chancellor's International Scholarship to make my PhD study possible.

I am fortunate to have a bunch of good friends. We shared many happy memories together, but now we are separated on different continents. I wish to thank them for helping me get through many hard times and bringing so much joy to my life.

I want to thank my cousin, Qu Zhihong, for her support, understanding and care.

Finally, I would like to thank my amazing parents, Liu Xiuli and Zhou Chunlin, for everything.

List of Publications

This is a list of my publications that have been published, based on my work on this thesis. Below, I show for each of them which chapter reflects the work related to the paper.

- R. Townsend, A. Tsakalidis, Y. Zhou, B. Wang, M. Liakata, A. Zubiaga, A. I. Cristea, and R. Procter. Warwickdcs: From phrase-based to target-specific sentiment recognition. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 657–663. ACL, 2015 (Chapter 7)

- Y. Zhou, E. Demidova, and A. I. Cristea. Analysing entity context in multilingual wikipedia to support entity-centric retrieval applications. In *Proceedings of the 1st International KEYSTONE (semantic keyword-based search on structured data sources) Conference*, pages 197–208. Springer, 2015 (Chapter 4)

- Y. Zhou, A. I. Cristea, and Z. Roberts. Is wikipedia really neutral? A sentiment perspective study of war-related wikipedia articles since 1945. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation.* ACL, 2015 (Chapter 6)

- Y. Zhou, E. Demidova, and A. I. Cristea. Who likes me more?: analysing entity-centric language-specific bias in multilingual wikipedia. In *Proceedings of the 31st ACM Symposium on Applied Computing*, pages 750–757. ACM, 2016 (Chapter 6)

- Y. Zhou and A. I. Cristea. Towards detection of influential sentences affecting reputation in wikipedia. In *Proceedings of the 8th International ACM Web*

Science Conference, pages 244–248. ACM, 2016 (Chapter 7)

- Y. Zhou, N. Kanhabua, and A. I. Cristea. Real-time timeline summarisation for high-impact events in twitter. In *Proceedings of the 22nd European Conference on Artificial Intelligence*, volume 285, pages 1158–1166. IOS Press, 2016 (Chapter 5)

- Y. Zhou, E. Demidova, and A. I. Cristea. What's new? analysing language-specific wikipedia entity contexts to support entity-centric news retrieval. *Transactions on Computational Collective Intelligence*, 26:210–231, 2017 (Chapter 4)

- Y. Zhou, A. I. Cristea, and L. Shi. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *Proceedings of the 18th International Conference on Web Information Systems Engineering*, pages 18–32. Springer, 2017 (Chapter 8)

Abstract

Social media has become one indispensable part of people's daily life, as it records and reflects people's opinions and events of interest, as well as influences people's perceptions. As the most commonly employed and easily accessed data format on social media, a great deal of the social media textual content is not only factual and objective, but also rich in opinionated information. Thus, besides the topics Internet users are talking about in social media textual content, it is also of great importance to understand the opinions they are expressing. In this thesis, I present my broadly applicable text mining approaches, in order to understand the topics and opinions of user-generated texts on social media, to provide insights about the thoughts of Internet users on entities, events, etc. Specifically, I develop approaches to understand the semantic differences between language-specific editions of Wikipedia, when discussing certain entities from the related topical aspects perspective and the aggregated sentiment bias perspective. Moreover, I employ effective features to detect the reputation-influential sentences for person and company entities in Wikipedia articles, which lead to the detected sentiment bias. Furthermore, I propose neural network models with different levels of attention mechanism, to detect the stances of tweets towards any given target. I also introduce an online timeline generation approach, to detect and summarise the relevant sub-topics in the tweet stream, in order to provide Internet users with some insights about the evolution of major events they are interested in.

Chapter 1

Introduction

1.1 Overview

Social media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allows the creation and exchange of user-generated content [138]. Social media has become one indispensable part of people's daily lives, as it records and reflects people's thoughts, ideas, opinions and events of interest [60], as well as influencing people's perceptions [86]. According to the Office for National Statistics in UK, the percentage of social media usage in Internet activities of adults continues to grow, rising to 63% in 2016[1]. This was an increase from 61% in 2015 and 45% in 2011. A similar trend can be observed worldwide. Such frequent usage of social media sites makes them essential information sources to understand the human world. However, the increase of user-generated data on social media sites lies far beyond the capability of human beings to decipher and analyse, even disregarding the bias human beings may introduce.

Textual data is the most commonly employed and easily accessed data format on social media [123]. It is thus a problem of the age to process and make sense of the social media textual content [130]. To understand the textual content on social media, one major problem is to decipher the topics discussed by Internet users [164, 186, 225]. A great deal of the social media textual content is not only factual and objective [15, 115], but also rich in opinionated information [97, 206]. Thus, besides the topics Internet users are talking about on social media, it is also

[1] https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2016

of great importance to understand the opinions they are expressing [175]. In this thesis, I focus on the above two instrumental areas of social media text mining: *Topic Analysis* and *Opinion Mining*.

The overall research goal of this thesis can be summarised as: *developing automatic text mining approaches to understand the topics and opinions in huge volume user-generated texts on social media.*

Compared with traditional text mining [88], there are some inherent additional challenges that need to be addressed when mining the textual content on social media, regardless of the purpose:

- **Huge Volume:** Looking at the example of Twitter, 672 million tweets were sent in relation to the 2014 World Cup; with a record of 618,725 tweets per minute when Germany won the World Cup Final[2]. Such high data volume requires the proposed approaches to be efficient and scalable. They should be capable of compressing reduplicative information, as well as differentiating and extracting useful information from the massive amount of unstructured textual data. Moreover, the data on social media sites emerges dynamically [265]. For some cases, it is essential to provide dynamic and real-time result updates, thus the ability to process data streams can also be an essential factor [7].

- **Informality:** Most of the natural language texts used on social media sites are informal and ungrammatical. Besides the widely occurring misspellings, the specific syntax newly generated for social media has increased the interpretation difficulty [130]. For example, emoticons are often used to express feelings; abbreviations are widely employed for usage convenience; culture-dependent Internet slangs have become popularised. It is important for the proposed approaches to consider and employ the informality of the textual content on social media.

- **Additional Information:** The textual content on social media sites is not standalone; on the contrary, it is often associated with additional information [123]. For example, each tweet is associated with a number of retweets, a number of "favourite"s, sometimes with the geo-tag; each Wikipedia article is associated with rich link information to demonstrate multiple kinds of relationships with other articles; each Facebook post is associated with a number

[2] `https://blog.twitter.com/2014/insights-into-the-worldcup-conversation-on-twitter-india`

of “like”s and a number of “share”s. The associated information can provide new opportunities to perform social media textual data mining, and some thoughtful measures need to be taken to absorb this external information into traditional approaches.

- **Diverse Information Needs:** The characteristics of textual data on different social media sites can differ greatly; even when facing the same social media site, there are various requirements for knowledge discovery in different application scenarios [123]. For textual content mining on social media, there is no solution that can fit for all application scenarios, thus different approaches need to be developed, conditional on the format and volume of the textual data, as well as the information needs of the potential users.

Concretely, I propose *automatic and effective* text mining approaches to understand the topics and opinions of user-generated content from Wikipedia[3] and Twitter[4], two social media sites that are rich in publicly available textual content, to provide insights about the thoughts of Internet users on named entities, policies, movements, as well as real-world events. In order to make the proposed approaches applicable on texts from various domains, the domain-specific features considered in these approaches are kept to a minimum.

With respect to *Topic Analysis*, I first propose approaches to *construct language-specific topic representations for entities in multilingual Wikipedia*, which can be applied to *provide language-specific results for entity-centric information retrieval.* Then I develop approaches to *detect and summarise fine-grained topics (sub-events) of high-impact events of interest in Twitter to form real-time timelines.*

Sentiment or opinion expressions, either explicitly or implicitly, are inevitable for user-generated content. In the area of *Opinion Mining*, I explore the *aggregated entity-centric sentiment bias across multilingual Wikipedia*, and propose an algorithm to *extract reputation-influential sentences that lead to sentiment bias.* I also develop ways to *detect target-specific stances in tweets* to deal with scenarios when the target is explicitly mentioned, implicitly mentioned, or not mentioned at all in the tweet.

[3] `https://www.wikipedia.org/`

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

1.2 Research Questions and Objectives

The main research questions and contributions of this thesis can be summarised as follows:

Topic representations of influential entities, such as celebrities and multinational corporations on the web, can vary across languages, reflecting language-specific topical aspects related to these entities. An important source of multilingual background knowledge about influential entities is Wikipedia — an online community-created encyclopaedia, containing more than 280 language editions. Such language-specific topic representations for entities could be further applied to provide context information when users simply utilise the entity names for relevant documents. Thus, in Chapter 4, I focus on the research question **RQ1. Can language-specific topic representations be constructed for entities employing knowledge from multilingual Wikipedia?** This research question is addressed by the following objectives:

O1.1. Create contexts to derive language-specific topic representations for the entities;

O1.2. analyse the similarities and differences in each entity's language-specific topic representations;

O1.3. propose an approach to improve the performance of IR applications on entity queries.

Twitter has become a valuable source of event-related information, namely, breaking news and local event reports. Due to its capability of transmitting information in real-time, I exploit the tweet stream for timeline summarisation of high-impact events, such as protests, accidents, natural disasters or disease outbreaks. Such summaries can serve as important event digests, where users urgently need information, especially if they are directly affected by the events. In Chapter 5, I study the research question **RQ2. Can timelines of high-impact events be generated automatically from the tweet stream?** Whilst RQ1 aims at generating an overview of topical aspects related to the entity, in RQ2, more fine-grained sub-topics are detected and the temporal dimension is further introduced to demonstrate the evolvement. RQ2 is addressed by the following objectives:

O2.1. Detect sub-topics relevant to the major event of interest in the tweet stream;

O2.2. summarise the detected sub-topics, to generate a timeline reflecting the evolution of the major event.

Sentiment and opinion expressions, either explicit or implicit, may be inevitable for user-generated content, even for Wikipedia, which sets the "Neutral Point of View" (NPOV) as its core policy. I perform extensive experiments to prove that the language and culture backgrounds of Wikipedia contributors make the NPOV policy of Wikipedia vary across its language editions, building linguistic points of view (LPOV). In Chapter 6, I tackle the research question **RQ3. Is there a language-specific sentiment bias in the multilingual Wikipedia, when talking about certain entities?** This research question is addressed by the following objectives:

O3.1. Propose an approach to quantify the entity-centric sentiment bias in multilingual Wikipedia at the corpus level;

O3.2. analyse the existence and extent of the entity-centric sentiment bias in multilingual Wikipedia.

Wikipedia has become the most frequently viewed online encyclopaedia website and an essential information source that influences people's perception towards entities. Some sentences in Wikipedia articles convey the contributors' opinions implicitly and have direct and obvious impact on people's opinions towards the mentioned named entities. In Chapter 7, I define and explore the research question **RQ4. Can the positive or negative reputation-influential information in Wikipedia be identified?** This research question is addressed by the following objectives:

O4.1. Annotate a dataset consisting of positive reputation-influential sentences, negative reputation-influential sentences and reputation non-influential sentences from Wikipedia articles;

O4.2. detect the reputation-influential sentences in Wikipedia articles from various domains, as well as the directions of these sentences that would influence the mentioned entities' reputation.

Besides real-word events reporting tweets, there are also a large number of tweets demonstrating Internet users' stances towards entities, policies, movements, events, etc. The stance of a tweet is determined not only by its content, but also by the given target. It remains a challenge to detect the stances of tweets with respect to a specific target, especially when the target is only implicitly mentioned, or not mentioned at all in the tweet, because it is necessary to infer the relationship between the topic discussed in the tweets and the given target. In Chapter 8, I work on the research question **RQ5. Can the performance of target-specific stance detection in tweets be improved, and if so, how?** This research question is addressed by the following objectives:

O5.1. Model the interaction between the tweet and the given target, construct the tweet's vector representation conditional on the target;

O5.2. detect the stance of the tweet towards the given target based on its target-specific vector representation.

1.3 Conclusion and Thesis Outline

In this chapter, I have overviewed the background and challenges of mining the textual content on social media. Then I have summarised the research questions of this thesis. The research questions are interconnected: RQ1 and RQ3 analyse the semantic differences between language-specific editions of Wikipedia, from the related topical aspects perspective and the aggregated sentiment bias perspective, respectively; RQ1 generates an overview of an entity's related topical aspects, RQ2 represents a further step, towards analysing not just topical aspects, but also the evolvement of sub-topics for the high-impact event, which is more fine-grained than RQ1 and considers an additional temporal dimension; RQ4 tries to detect the sentences that lead to the aggregated sentiment bias in multilingual Wikipedia, which is explored in RQ3; while in RQ4, all sentences mention the target entity of interest explicitly by name, RQ5 explores to solve the problem when the target may or may not be explicitly mentioned.

In Chapter 2, I will present an overview of the related work. Chapter 3 introduces some classical text mining techniques employed in this thesis. Chapter 4 and Chapter 5 present works relevant to Topic Analysis: Chapter 4 describes the work on creating language-specific topic representations of entities to support entity-centric, language-specific information retrieval applications (RQ1); Chapter 5 depicts the work on detecting and summarising sub-topics for high-impact events from the tweet stream (RQ2). Chapter 6, Chapter 7 and Chapter 8 present works relevant to Opinion Mining: Chapter 6 introduces the work on understanding the existence and extent of entity-centric language-specific sentiment bias in multilingual Wikipedia (RQ3); Chapter 7 discusses the work on detecting reputation-influential sentences in Wikipedia articles (RQ4); Chapter 8 provides details about the work on target-specific stance detection in tweets (RQ5). The thesis is concluded in Chapter 9, with the contributions summarised and some ideas about future directions provided.

Chapter 2

Related Work

In this chapter, I start by introducing the general categories of Social Media Mining. Then I elaborate on recent works on Topic Analysis and Opinion Mining of textual content on social media, which are the main focus of this thesis. Specifically, I review the main lines of research of these two areas and demonstrate how the work in this thesis identifies existing gaps and proposes new solutions.

2.1 Social Media Mining

Social Media Mining is the process of representing, analysing, and extracting actionable patterns from social media data [309]. With the development of information technology, various social media sites have emerged, such as Facebook, Twitter, Wikipedia, Youtube, Flickr and LinkedIn, to serve people's interaction and communication needs in different scenarios. Depending on the social media sites, the generated social media data typically takes diverse forms, which include text, image, audio, video, network structure, etc.

Various studies have been conducted with respect to different social media data forms and social media sites, aiming at understanding human behaviour and building applications to benefit people's daily life. Following [149], the area of Social Media Mining can be roughly categorised into three areas, based on the mining objects: Social Media Content Mining, Social Network Structure Mining and Social Media Usage Mining.

The content on social media sites ranges from structured data in databases to multimedia data. In [19, 157, 248], researchers employed textual data on Twitter to detect influenza epidemics. In [328], researchers analysed conversations on Twitter to understand how Internet users spread, support or deny rumours. In [204,290], researchers developed an automatic human age estimator, based on the images crawled from Flickr; in [290], researchers proposed a machine learning framework to tackled the automated image tagging task on Flickr. In [196], researchers performed activity recognition on YouTube videos. Researchers have also been using a combination of audio, visual and textual information on YouTube to perform multimodal sentiment analysis [222]. The network structures generated on social media sites from online interactions and explicit relationship links in social media [43] have also attracted a lot of researchers' attention. For example, in [18], researchers identified the influential and susceptible users in the Facebook friendship network; in [297], community detection was performed on Wikipedia, Flickr, Facebook, Google+ and Twitter; in [27], researchers examined the role of Facebook friendship network in online information diffusion; in [265], researchers analysed the information flow of the retweet network that was formed during a political protest against the rise in university tuition fees in England; in [264], researchers identified key users on Twitter, by analysing the dynamic retweet network around specific topics. Internet users' behaviour on social media sites results in a huge volume of access logs, server logs, browser logs, etc. The usage data has helped researchers to understand and predict user behaviour on social media sites. In [266], researchers measured human activity on the web, based upon the numbers of article views for Wikipedia. In [70], researchers analysed the relationship between users' personality traits and the usages of Twitter and Facebook. In [275], researchers mined the server access logs of Flickr to find patterns of user viewing behaviour. In [10], researchers employed the usage features of Yahoo! Answers to perform quality estimation.

Among the various formats of content on social media, textual data is the data format stored in most social media sites [123], which is also the focus of this thesis. Text mining is a process to extract useful information from unstructured textual data through the identification and exploration of interesting patterns [88]. The area of *Mining Textual Content on Social Media* has been active since the generation of social media due to its wide application. Techniques such as Machine Learning, Data Mining, Semantic Web, Natural Language Processing and Information Retrieval have been applied in this area to satisfy the needs of different application scenarios [9, 130]. Besides the aforementioned Disease Surveillance, other sub-areas include: Question Answering [38], Recommendation [110], Topic Analysis [314], Opinion Mining [150], etc. Topic Analysis and Opinion Mining are two basic and instrumental sub-areas of mining social media text, the results of which are often employed by other sub-areas. For this reason, my research focuses on these two areas, which are described in the following in greater detail, with emphasis on the current state-of-the-art, and its unsolved issues and problems.

2.2 Topic Analysis

Topic Analysis determines a text's topic structure, a representation indicating what topics are included in a text and how topics change within the text [169]. There are two lines of research under the category of Topic Analysis: Topic Categorisation [232, 241, 281] and Topic Detection [57, 217, 226]. *Topic Categorisation* classifies the documents into *pre-known* categories; while *Topic Detection* aims at detecting *unknown* topics.

Constructing *Topic Representations* for textual content is an indispensable step for both lines of research. While the general approaches for *Content Representation*, which will be discussed later in Section 3.2, can all be applied on constructing topic representations, researchers have been developing some approaches especially useful for the purpose of analysing the topics. In [173], researchers performed n-gram feature selection and feature weighting based on their uniqueness with respect to topics. Their work was further developed in [111], with n-gram features replaced by relation features, themes, inter-related concepts, etc. In [65], researchers weighted unigram features by the possibilities of their occurrences in topic-specific summaries. Besides n-grams, extracted named entities [154], multi-word segments and phrases [125, 176] have also been employed as features for topic representation. In [250], researchers used word-clusters, to replace single words as features. Other information, such as temporal information of the text, has also been considered in [8, 151], when constructing topic representations of texts. More effective topic representation approaches than the state-of-the-art, are provided in Chapter 4 and Chapter 5, respectively, in order to perform a comprehensive and accurate analysis of the semantic differences in multilingual Wikipedia when discussing certain entities from the related topical aspects perspective, and detect fine-grained sub-topics in real-time for high-impact events from diverse and ungrammatical tweets.

Information from public knowledge bases, such as Wikipedia, has been exploited by researchers to augment the representations of texts for topic analysis purpose, as in [90, 91, 121, 124, 125, 127, 279, 308], etc. My work on constructing and analysing entity-centric, language-specific topic representations of multilingual Wikipedia is related to these works. More discussion about relevant research and the differences between my work with former works are included in Section 4.6 of Chapter 4.

Researchers have been focusing on various tasks in the Topic Detection research area. For example, *First Story Detection* [217,254], *Emerging Topic Detection* [13, 57,105], *Event Tracking* [48,172], *Timeline Generation* [170,282,301], etc. Most works tackle the above problems from two directions: *Clustering* [57,254] and *Topic Modelling* [13,105]. My work described in Chapter 5 is in line with these works, with more related works presented, and the differences between my work and former research discussed in Section 5.4.

2.3 Opinion Mining

Opinion Mining is the field of study that analyses people's opinions, sentiments, appraisals, attitudes and emotions toward entities and their attributes expressed in the written text [174]. Opinion Mining offers organisations the ability to monitor various social media sites in real time and act accordingly [87]. Opinion Mining has been applied in many areas. In [42,229], researchers applied opinion information in predicting the trend of stock markets. In [273,278], researchers have proved that opinion information was beneficial in forecasting the results of political elections. In [78,79], researchers exploited the customers' sentiment towards products, to perform recommendations. [17,21] demonstrated that the estimation of movies' future box revenues can also benefit from the opinion information on social media sites. The outputs of opinion mining include: graded opinion-related scores [97,206], opinion-related labels [135,150,167], opinion lexicons [137,178], etc. Some common approaches for generating graded scores and labels will be presented in Section 3.3.

According to [87, 174], opinion mining has been carried out mainly at three levels of granularity: document level, sentence level and aspect level. *Document-level* opinion mining assumes that the document contains a coherent opinion on one object expressed by the author of the document [87]. Example works on document-level opinion mining are [198,212]. The fact that even a single document may contain multiple opinions on the same object necessitates the application of sentence-level opinion mining [122, 211]. In *sentence-level* opinion mining, the document is split into sentences first, then the analysis is performed towards the resulting sentences, rather than the whole document. Because of the 140 characters limitation of Twitter (the limitation was relaxed in 2016[1]), there is only one sentence in one tweet for most of the time, thus sentiment classification at the tweet level is generally considered as similar to sentence-level opinion mining. *Aspect-level* opinion mining is often employed to discover people's opinion on certain aspects in one sentence or document [78, 79, 87, 167, 174]. Aspect-level opinion mining often involves some sub-problems, such as aspect extraction [77, 197, 307], association between sentiment expressions and aspects [156, 267], and aspect-centric sentiment summarisation [165, 179].

To analyse the collective sentiment and opinion information on social media sites, researchers have been performing opinion mining at the *corpus level*. Example works include [42, 97, 206]. In Chapter 6, I present my work on analysing entity-centric opinion bias in multilingual Wikipedia corpora, with more relevant works reviewed in Section 6.4, in terms of specific differences to my work.

Besides the above traditional problems on opinion mining, some new challenges have been proposed, relevant to opinions expressed in natural language texts. In [174], researchers pointed out that the opinion information can hide in factual statements. For example, "I bought the mattress a week ago, and a valley has formed in the middle." and "Google has more users than Bing." are both factual sentences, but they imply the authors' opinions implicitly. In [100], researchers proved that the syntactic choices of factual statements can influence readers' perceptions towards the incidents described. Analysis around this kind of *polar facts*, was referred to as implicit sentiment/opinion expression detection in some studies on product reviews [168, 268, 312]; it was referred to as bias analysis in some studies on some texts that were supposed to be written from a neutral point of view [230, 300]; it was also referred to as reputation polarity analysis in some studies focusing on the influence of the *polar facts* on named entities reputation, along with other subjective texts [14, 96]. My work on detecting reputation-influential sentences is detailed in Chapter 7, along with other relevant works to this problem discussed in Section 7.6. In [193], researchers proposed the problem of detecting target-specific stances in tweets. This interesting problem is also a target of this thesis. A detailed explanation of problem can be found in Chapter 8, with a summary of related works included in Section 8.4.

[1] `https://twitter.com/twitter/status/742749353689780224?lang=en`

2.4 Conclusion

In this chapter, I have reviewed recent works on social media mining, with a special focus on mining the textual content on social media to understand the topics and opinions. My work on creating language-specific topic representations for entities in multilingual Wikipedia, described in Chapter 4, builds on top of the most recent works on constructing topic representations for textual content; my work on real-time timeline summarisation for high-impact events in Twitter, presented in Chapter 5, further builds on research on Topic Detection and Timeline Generation; my work on analysing entity-centric sentiment bias in the multilingual Wikipedia, discussed in Chapter 6, further expands the area of analysing aggregated sentiment at the corpus level; my work on detecting reputation-influential sentences in multilingual Wikipedia, described in Chapter 7, follows from the most recent research on detecting sentences with implicit sentiment expression; my work on target-specific stance detection in tweets, presented in Chapter 8, builds on research on stance detection.

The next chapter explores the background of this thesis from a technical and technological point of view, explaining the main theories and techniques used in this thesis, as well as the overall framework relevant for the implementation of my research.

Chapter 3

Technological Background

In this chapter, I start by presenting the general framework used by researchers for mining textual content on social media. Then I explore some classical techniques for content representation and text classification. Since it is not possible to discuss all the baseline approaches in detail, I only include the techniques frequently employed or developed in this thesis. The description of other relevant baseline techniques can be found in following chapters, where they are directly relevant to the research presented there.

Textual Content Mining on Social Media is the main focus of the work in this thesis. Following [123], I present the general framework for mining textual content on social media, which consists of four consecutive phases: Data Extraction, Preprocessing, Content Representation and Knowledge Discovery, as shown in Figure 3.1.

3.1 General Framework for Mining the Textual Content on Social Media

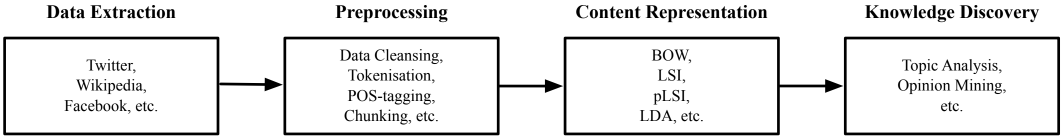

Figure 3.1: Framework for Mining the Textual Content on Social Media [123].

Data Extraction: In Data Extraction phase, I collect textual data from interested social media sites using the provided API. Some regular expressions may be applied to filter out the unnecessary texts.

Preprocessing: The texts directly extracted from the API usually contains some uninformative parts, such as the HTML tags, thus the Data Cleansing step is needed during the Preprocessing phase. After Data Cleansing, the textual data will be tokenised for further process. Depending on the application, other steps may be needed, such as POS-tagging and Chunking.

Content Representation: The unstructured natural language texts will be represented by vectors during the Content Representation phase. In the traditional Bag-of-Words (BOW) model [240], each dimension of the vector representation corresponds to one unique word in the document corpus, which makes the vector representation of very high dimensionality. Some dimension reduction approaches have been developed, such as Latent Semantic Indexing (LSI) [73], Probabilistic Latent Semantic Indexing (PLSI) [118] and Latent Dirichlet Allocation (LDA) [41], to map the texts to lower dimensional space, with each dimension corresponding to one concept or one topic. The approaches employed for Content Representation are critical for the following Knowledge Discovery phase, as vector representations conditional on the objective of Knowledge Discovery often lead to better performance.

Knowledge Discovery: Based on the vector representations of texts, various data mining algorithms can be applied to discover underlying patterns, such as Classification and Clustering. The results can be further processed and interpreted to provide insights from the texts, such as discussed topics and expressed opinions in the original social media textual data.

It should be noted that the four phases can either be executed sequentially after the former phase has terminated, or be executed recursively, to process the data streams under online settings. Moreover, I separate the procedure of mining the textual content on social media into four phases for convenience; in practice, these phases are dependent on each other. The techniques employed in one phase rely on other phases, and may also be directly driven by the objective of Knowledge Discovery. Especially adjacent phases, such as the Content Representation phase and the Knowledge Discovery phase, are often combined to be optimised jointly. For example, in [131, 136, 143, 261], researchers stacked models for classification, such as Support Vector Machine (SVM), Multilayer Perceptron (MLP) or affine layers, which can be seen as MLP without the hidden layer, on top of multiple neural network-based content representation structures; in [155, 227], researchers incorporated the categorical information into the LDA model. After training the resulting models with the labelled dataset, it was able to perform a joint inference on the vector representations and the classes of test documents. In the following, the phases which need most improvements are further analysed in greater detail.

3.2 Content Representation

3.2.1 Bag-of-Words Representation

In the BOW model, the word ordering information is ignored. The dimensionality of a document's vector representation equals to the number of distinct words in the document corpus, denoted by V. The most widely used weighting schemas are *binary*, *term frequency (tf)* and *term frequency — inverse document frequency (tfidf)* [240].

In the *binary-based* BOW vector representation, the weight for each word can only be 1 or 0, simply indicating the presence of a word, not the importance.

In the *tf-based* BOW vector representation, each word is weighted by its number of occurrences in the document. However, the term frequency is not a suitable measurement of word importance. Some stop words, such as 'the', 'us' and 'you', can occur frequently in some documents, but they are not informative of the content in the document.

For document d, where $d \in \{1, \ldots, D\}$, the weight of vth unique word w_v from the document corpus in its *tfidf-based* BOW vector representation is calculated as follows:

$$tfidf_{d,v} = tf_{d,v} \times log\frac{D}{df_v}, \tag{3.1}$$

where $v \in \{1, \ldots, V\}$, $tf_{d,v}$ is the number of occurrences of word w_v in document d; D is the number of documents in the corpus; df_v represents the number of documents in the corpus containing the word w_v.

For the *tfidf* weighting scheme, the importance of a word to a document is not only measured by its number of occurrences, but also dependent on its informativeness. Words appearing in a lot of documents are considered less informative than rarely appearing words. Resulting from that, words tend to have higher weights if they occur many times within a small number of documents in the corpus. Correspondingly, the influence of some uninformative but frequent words will be dampened.

Researchers have also been using n-grams to increase the expressive capability of the BOW vector representation. In this way, each dimension corresponds to one word or one n-gram that occurs in the document corpus.

The BOW model is easy to implement. The resulting vectors can be interpreted effortlessly. However, they are often of high sparsity and dimensionality. Besides, the textual variants of words, the synonymy and the polysemy problem, are not addressed; the syntax and semantic information, or the order of words in the documents is not fully explored. For the above reasons, some improvements are needed, to achieve desirable performance using the BOW model in various applications. However, for some applications, BOW model is a straightforward method, which is especially useful when results need to be explained easily. Thus, in this thesis, I develop improved variations of the BOW model when constructing the entity-centric topic representations of multilingual Wikipedia (Chapter 4); when creating tweets' vector representations to cluster near-duplicate tweets (Chapter 5); when creating vector representations for entity-centric language-specific contexts to analyse the sentiment bias in multilingual Wikipedia (Chapter 6); as well as when creating vector representations for sentences in Wikipedia articles to classify if they express sentiment implicitly (Chapter 7). I make different variations to the original BOW model according to different application scenarios, details of which can be found in corresponding chapters.

3.2.2 Probabilistic Representation via Latent Dirichlet Allocation

LDA is a generative probabilistic model for collections of texts. It is a three-level hierarchical Bayesian model, in which each document in a document corpus is modelled as a finite mixture over an underlying set of topics, and each topic is modelled as an infinite mixture over an underlying set of topic probabilities [41]. In other words, LDA model represents the documents as vectors with each dimension corresponding to one topic, which is a distribution of words, and the values of each dimension represent the degrees to which this topic is referred to in the documents.

Figure 3.2 shows the graphical model representation of LDA, based on [40]. In Figure 3.2, α is the proportion parameter of the Dirichlet prior on the per-document topic distributions; η is the topic parameter of the Dirichlet prior on the per-topic word distributions. The α parameter and the η parameter specify the prior beliefs about topic sparsity in the documents, and the prior beliefs about word sparsity in the topics, respectively. β_k is the word distribution over the vocabulary of topic k, where $k \in \{1, \ldots, K\}$. θ_d is the topic *probabilistic representation* of

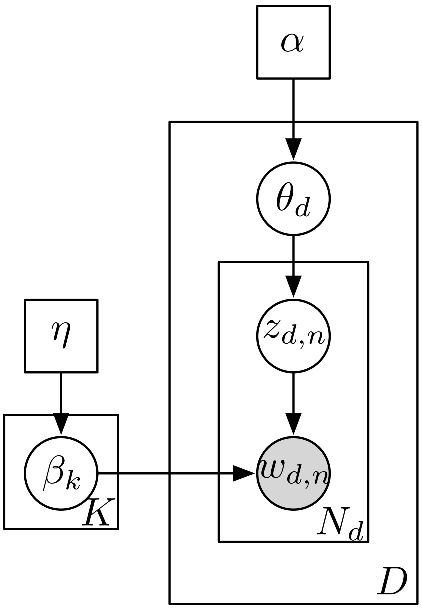

Figure 3.2: Graphical model representation of LDA [40].

document d, where $d \in \{1, \ldots, D\}$. Each document d is represented by a sequence of N_d words, $\mathbf{w_d} = (w_{d,1}, \ldots, w_{d,N_d})$, $w_{d,n}$ is the n^{th} word in document d, where $n \in \{1, \ldots, N_d\}$. I use $\theta_{d,k}$ to denote the topic proportion of topic k in document d, where $\sum_{k=1}^{K} \theta_{d,k} = 1$; $z_{d,n}$ is the topic assignment for the nth word in document d.

The generative process of LDA model is as follows (I use Dir to represent Dirichlet distribution and $Multinomial$ to represent Multinomial Distribution):

- Choose $\theta_d \sim \mathrm{Dir}(\alpha)$.
- Choose $\beta_k \sim \mathrm{Dir}(\eta)$.
- For each of the N_d words in document d:
  - Choose $z_{d,n} \sim \mathrm{Multinomial}(\theta_d)$.
  - Choose $w_{d,n} \sim \mathrm{Multinomial}(\beta_{z_{d,n}})$.

The joint likelihood of all variables can be written as:

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N_d} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \boldsymbol{\beta}). \tag{3.2}$$

The key inferential problem is to compute the conditional distribution of the topic structure given the observed documents, which can be written as:

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}) = \frac{p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w})}{p(\boldsymbol{w})}, \tag{3.3}$$

which is an intractable problem. Some approximate inference algorithms, such as Gibbs Sampling [103] and Variational Inference [41], have been developed to perform parameter estimation and inference.

Except for being used as a dimension reduction method to generate probabilistic representations for documents, the LDA model can also be directly applied for topic analysis purposes, as each dimension in the vector representations corresponds to interpretable latent topics. The LDA model outperforms the LSI model in terms of generating understandable topics and the pLSI model in terms of generative power. The LDA model provides an intuitive probabilistic foundation for dimension reduction, and is easy to be extended and modified for various application scenarios. However, whilst the model learns patterns based on word co-occurrences, the word ordering information is still ignored; the number of topics needs to be fixed empirically ahead of learning, which is an influential factor in the performance; the LDA model tends to learn broad topics, such as “Sports”, “Finance” and “Politics”, thus is not applicable when sharper and more fine-grained topics are needed, which is the case for the fine-grained topic detection proposed in Chapter 5. In this thesis, I apply the LDA model to enrich the vector representations of Wikipedia sentences, in order to determine if they are reputation-influential (Chapter 7).

3.2.3 Distributed Representation via Neural Networks

A distributed representation means a many-to-many relationship between two types of representations, such as concepts and neurons [116]. Researchers have been proposing various algorithms to learn distributed representations of words to capture their syntactic and semantic relationships. Many word embeddings generation approaches have been proposed based on words’ distributional properties in large samples of language data, which include *Skip-gram* [187] and *Glove* [216]. The Skip-gram model tries to learn word embeddings that are useful for predicting the surrounding context words [187], while the Glove model tries to learn word embeddings whose dot products equal to the logarithm probabilities of co-occurrences in defined contexts [216].

Various compositional architectures have been proposed to generate distributed representations for sentences/documents based on word embeddings. For example, the *Paragraph Vector* model [160] optimises sentence representations by predicting the target word, using the concatenation of the sentence vector representation with vector representations of words in the context; the *Skip-thoughts* model [146] learns sentence representations using an encoder-decoder neural network architecture, the encoder encodes word embeddings to a sentence vector representation and the decoder tries to predict the surrounding sentences; the *Deep Averaging Network (DAN)* model [131] takes the average of the vector representations of words from the sentence and passes it through one or more feedforward layers; the *Recursive Neural Tensor Network (RNTN)* model [251] employs a sentence parse tree to iteratively compute vector representations for higher nodes in the tree based on lower nodes using the same composition function; the *Dynamic Convolutional Neural Network (DCNN)* model [136] applies convolutional layers combined with dynamic pooling layers to capture word relations of varying size; the *Document Vector through Corruption (Doc2VecC)* model [55] overcomes one drawback of the *Paragraph Vector* model, which is the complexity of the models grows with the length of the document, by representing each document as an average of the embeddings of words randomly sampled from the document.

Recurrent Neural Network (RNN) [85] and *Convolutional Neural Network (CNN)* [162] are two frequently employed basic structures in various sentence modelling approaches, which I will further describe here, based on [159,161] and [143], respectively.

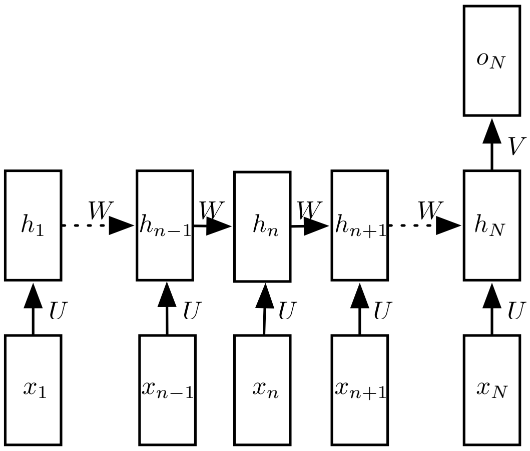

Figure 3.3: A basic recurrent neural network [159].

Figure 3.3 shows a basic RNN structure. RNN makes use of *sequential* information, by "memorising" what has happened so far. Given a sentence/document consisting of a sequence of N words $\mathbf{w} = (w_1, \dots, w_N)$, I first map all the words to the embedding space $\mathbf{x} = [x_1, \dots, x_N]$, where $x_n \in \mathbb{R}^{d_0}$ represents the word embedding of word w_n and $n \in \{1, \dots, N\}$. To reduce the number of parameters needed to be learnt, the basic RNN updates the hidden states h_n based on the same weights U and W for each word embedding x_n, as follows:

$$h_n = \tanh(Ux_n + Wh_{n-1}). \tag{3.4}$$

The output at the last step $o_N = \tanh(Vh_N)$ is generally used as the distributed representation of the sentence/document. Assume $h_N, o_N \in \mathbb{R}^{d_1}$, then $U \in \mathbb{R}^{d_1 \times d_0}$, $W \in \mathbb{R}^{d_1 \times d_1}$ and $V \in \mathbb{R}^{d_1 \times d_1}$ are the weights need to be learnt.

The basic RNN can be seen as a regular deep feedforward neural network, i.e. MLP, as elaborated in Section 3.3.3, with many hidden layers. In practice, the basic RNN cannot capture information many steps ago, because of the *vanishing gradient* and *exploding gradient* problem [213]. Many structures have been proposed to augment the performance of basic RNN in learning long-term dependencies, such as *Gated Recurrent Unit (GRU)* [58], *Long Short-Term Memory (LSTM)* [117] and *Attention Mechanism* [144].

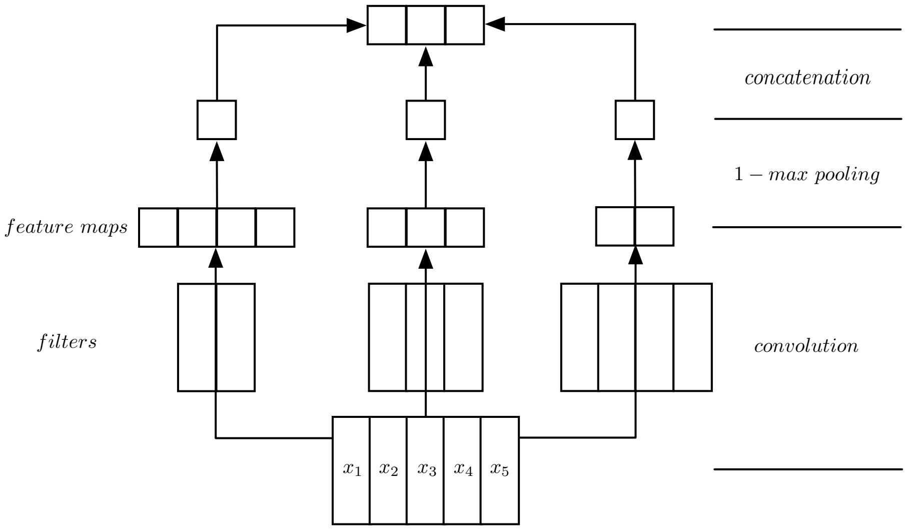

Figure 3.4: Single layer convolutional neural network for sentence modelling.

Figure 3.4 shows a single layer CNN for sentence/document modelling. CNN can be understood as a *hierarchical* architecture, which is good at extracting position invariant and compositional features. The convolution operation applies a filter $w_f \in \mathbb{R}^{kd_0}$ on the concatenation of vector representations of k consecutive words:

$$c_i = f(w_f^T x_{i:i+k-1} + b_f), \tag{3.5}$$

where $i \in \{1, \ldots, N-k+1\}$, f is the rectified linear unit function and $b_f \in \mathbb{R}$ is a bias term. The result feature maps $\mathbf{c} = (c_1, c_2, \ldots, c_{N-k+1}) \in \mathbb{R}^{N-k+1}$. The 1-max pooling operation aims at capturing the most important and salient feature in each feature map and reducing the output dimensionality by taking the maximum value of each feature map:

$$\hat{c} = \max\{\mathbf{c}\}, \tag{3.6}$$

where $\hat{\mathbf{c}} \in \mathbb{R}$.

The equation above illustrates the feature extraction process of one filter. Many filters with varying sliding window size k can be applied to obtain multiple features. The features extracted by different filters are concatenated to form the distributed representation of the sentence.

In [305], researchers pointed out that RNN is better at generating a sentence vector representation that reflects the semantic information of the *whole* sentence, while CNN outperforms RNN in accentuating the *informative parts* in the sentence.

The distributed sentence/document vector representations generated by various neural network structures are continuous, dense and abstractive, which reflect the syntactic and semantic information of the sentences. They have strong expressive powers, which is beneficial for downstream applications. After stacking the content representation structures described in this section with some classification structures described in the following section, the weights in the resulting model can be jointly trained via labelled training data. However, the resulting vector representations are hard to interpret. Moreover, there are usually a large number of weights in the neural network that need training, which results in high demands on the amount of training data, the computation capability of the hardware, as well as the performance of the optimiser. Besides, there are many influential hyper-parameters involved in the neural network structures, the tuning of which is a non-trivial task. In this thesis, I demonstrate that delicately designed neural network structures for certain tasks make the benefits of distributed representations far overweigh the shortcomings. I propose to apply attention mechanisms on top of the traditional neural network models, to generate distributed vector representations of tweets, conditional on the target, in order to perform target-specific stance detection (Chapter 8).

3.3 Classification Methods

3.3.1 Lexicon-based Classifier

In lexicon-based classification, documents are assigned labels by comparing the number of words/n-grams that appear from pre-constructed word/n-gram lists [84]. Lexicon-based classification is mostly used to infer the sentiment polarities of documents with the help of sentiment lexicons. Existing sentiment lexicons can be roughly divided into two categories: *Polarity-based lexicons* and *Valence-based lexicons*. In Polarity-based lexicons, words/n-grams are annotated with the overall sentiment orientations, i.e., *positive* or *negative*, such as in the Opinion Lexicon [122], Macquarie Semantic Orientation Lexicon (MSOL) [192] and the Multi-perspective Question Answering (MPQA) Opinion Lexicon [288]. In Valence-based lexicons, words/n-grams are annotated by the valence scores of the sentiment intensity, such as used by the AFINN Lexicon [205], SentiWordNet Lexicon [25], Sentiment140 Lexicon [194] and the NRC Hashtag Sentiment Lexicon [194].

Three main approaches have been proposed to generate sentiment lexicons: *manual*, *dictionary-based* and *corpus-based* [174].

The AFINN Lexicon [205] and the MPQA Lexicon [288] were constructed through manual approaches: each word/n-gram in these two lexicons was annotated manually by the authors. This was labour-intensive and time-consuming; moreover, the annotation results can be biased, because of the differences in cognition among human beings. The AFINN Lexicon used discrete values ranges from -5 (very negative) to $+5$ (very positive) to denote the sentiment valences. The dictionary-based approach used a small set of sentiment seed words with known positive or negative orientations to bootstrap the collection of positive and negative words, based on the synonym and antonym structure of a dictionary [174].

Dictionary-based approaches are employed in the construction processes of the SentiWordNet Lexicon [25], Opinion Lexicon [122] and MSOL [192]. Concretely, for the Opinion Lexicon, researchers enriched the adjective seed words with their synonyms and antonyms from WordNet [188]. The adjective seed words shared the same sentiment orientation as their synonyms and opposite sentiment orientations as their antonyms. For the SentiWordNet Lexicon, researchers additionally trained multiple classifiers based on the definitions of the enriched sentiment seed words, and applied the classifiers to calculate the positive, negative and objective scores of all the words in WordNet. An extra random-walk step was further performed on the WordNet definiens-definiendum binary relationship graph to adjust the positive and negative scores, out of the intuition that the positivity and negativity can be mapped from the definitions to the words being defined. The final positive and negative valences were determined by applying power law distribution functions to the rankings of the positive and negative scores generated by the random-walk step, and the objective valences were assigned based on the positive and negative intensities. For each word, its positive, negative and objective valences were continuous values, which ranged in the interval $[0.0, 1.0]$ and their sum was 1.0. For the MSOL, researchers initially used eleven affix patterns to expand the seed words set, then they employed the group information from the Macquarie Thesaurus [37] to perform another expansion: if a group had more positive seed words than negative seed words, all the words/n-grams in the group were marked as positive; otherwise, all the words/n-grams in the group were marked as negative; if a word/n-gram occurred in multiple groups, its sentiment orientation was determined by its most common sentiment orientation. The sizes of the lexicons generated by dictionary-based approaches are restricted by the sizes of the dictionaries, which are not adequate to cover the language-usage variations and multi-word expressions in social media texts.

The corpus-based approach also uses a small set of sentiment seed words with known positive or negative orientations to bootstrap, but is based on the syntactic or co-occurrence patterns in a large corpus [175]. Researchers exploited the same corpus-based approach to generate the Sentiment140 Lexicon [194] and the NRC Hashtag Sentiment Lexicon [194]. The construction of these two lexicons was based on the assumption that a coherent sentiment orientation was expressed in all words/n-grams in a tweet. Researchers initially utilised 32 positive hashtags and 36 negative hashtags to annotated a tweet corpus: a tweet was considered positive if it contained one of the 32 positive hash-tagged seed words, and negative if it contained one of the 36 negative hash-tagged seed words. Then the Point-wise Mutual Information (PMI) scores for all words/n-grams were calculated, which indicated their association with positive sentiment orientation if the scores were positive, and their association with negative sentiment orientation if the scores were negative. The PMI scores were further employed as the sentiment valences, which were continuous values ranged in the interval $(-\infty, \infty)$. The corpus-based approach is an efficient and automatic solution to generate domain dependent and context specific sentiment lexicons. However, for some words/n-grams without the sufficient number of occurrences, the reliability of their assigned sentiment orientation/valence is in question; the intra-sentential sentiment coherency assumption can be invalid for some sentences with complex syntactic structures.

The lexicon-based classification is unsupervised and relies on the linguistic heuristics introduced by the researchers. When using a polarity-based lexicon, because only words/n-grams with strong sentiment intensities are included in the lexicon, the sentiment orientation of a document is decided by the differences between the number of positive words/n-grams and the number of negative words/n-grams from the sentiment lexicon that appear in the document, as in [122, 212]. Specifically, when using $+1$ to denote positive sentiment orientation, and -1 to denote negative sentiment orientation, the sentiment orientation of document d, denoted by SO_d, can be decided as follows:

$$SO_d = sgn(\sum_{w_v \in \mathbf{w_d} \cap \mathbf{w_L}} tf_{d,v} \times so_v). \tag{3.7}$$

In the above equation, $\mathbf{w_d}$ represents all the words in the document, $\mathbf{w_L}$ represents all the words in the lexicon, so_v represents the sentiment orientation of word w_v labelled in the lexicon, $tf_{d,v}$ represents the term frequency of word w_v in document d, sgn represents the sign function.

When using a valence-based lexicon, words/n-grams carrying different levels of sentiment information are all included in the lexicon, the sentiment valence of a document is usually calculated as the average sentiment valence of all the words/n-grams from the sentiment lexicon that appear in the document, as in [52, 189, 274]. Specifically, the sentiment valence of document d, denoted by SV_d, can be calculated as follows:

$$SV_d = \frac{\sum_{w_v \in \mathbf{w_d} \cap \mathbf{w_L}} tf_{d,v} \times sv_v}{\sum_{w_v \in \mathbf{w_d} \cap \mathbf{w_L}} tf_{d,v}}. \tag{3.8}$$

In the above equation, sv_v represents the sentiment valence of word w_v annotated in the lexicon.

The lexicon-based classification is unsupervised, it can be easily implemented, interpreted and modified. In particular, the valence-based lexicons can be employed to generate gradable results. However, the coverage and credibility of the lexicon limit its effectiveness, especially when facing texts of great flexibility and variability, such as textual content on social media. The classification result is dependent on the rules introduced by the researchers, which only consider individual words/n-grams in the texts for most of the time, and ignore the syntactic and semantic information. Even though some researchers have proposed additional rules to modify the sentiment orientations and valences of the words/n-grams in the lexicon, based on their contexts [128,256], these rules are not comprehensive enough to cover all the language usage patterns, especially when facing domain-dependency and polysemy scenarios. The high dependency on handcrafted rules also restricts the application of lexicon-based classification to areas where such rules can be easily generalised, such as sentiment polarity classification, but not areas where more sophisticated inference is needed, such as target-specific stance detection. When facing data from various *unknown* domains, lexicon-based classification using existing lexicons is stronger in generality than other approaches. In this thesis, the lexicon-based, *unsupervised* sentiment analysis is employed to *quantify* the *aggregated* sentiment bias in multilingual Wikipedia contexts of the specified entity, which come from various *unknown* domains (Chapter 6).

3.3.2 Support Vector Machine Classifier

SVM [44] is one of the most frequently used text classification algorithms. SVM is a linear two-class classifier, its objective being to find a hyperplane that separates the training examples from two classes, with the maximum margin. The margin refers to the distance from the closest training example(s) to the hyperplane. Let the inputs be some d_1-dimensional input vector $x \in \mathbb{R}^{d_1}$, and the label $y \in \{-1, 1\}$. Assume the training set is separable by a linear hyperplane in the input space, which can be represented by w and b, where $w \in \mathbb{R}^{d_1}$ is the weight vector and $b \in \mathbb{R}$ is the bias term. Based on [202], for the m^{th} training example $(x^{(m)}, y^{(m)})$, where $m \in \{1, \ldots, M\}$, its distance (geometric margin) $s^{(m)}$ to the hyperplane (w, b) can be calculated from its functional margin $\hat{s}^{(m)}$, as:

$$s^{(m)} = \frac{\hat{s}^{(m)}}{\|w\|} = \frac{y^{(m)}(w^T x^{(m)} + b)}{\|w\|}. \tag{3.9}$$

Let $s = \frac{\hat{s}}{\|w\|} = \min_m s^{(m)}$. Then the optimisation problem can be formulated as:

$$\begin{aligned} &\max_{\hat{s},w,b} \frac{\hat{s}}{\|w\|} \\ &\text{subject to: } y^{(m)}(w^T x^{(m)} + b) \geq \hat{s}. \end{aligned} \tag{3.10}$$

Because s is invariant to the scaling of w and b, one could introduce the scaling constraint that $\hat{s} = 1$. Then the optimisation problem changes to:

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 \\ &\text{subject to: } y^{(m)}(w^T x^{(m)} + b) \geq 1. \end{aligned} \tag{3.11}$$

In practice, the training examples are often not linearly separable, or a much larger margin can be achieved if some errors are allowed. To deal with this, researchers in [67] proposed the soft-margin hyperplane, which rewrote the optimisation problem as:

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{m=1}^{M} \xi_m \\ &\text{subject to: } y^{(m)}(w^T x^{(m)} + b) \geq 1 - \xi_m,\ \xi_m \geq 0. \end{aligned} \tag{3.12}$$

In the above equations, ξ_m are slack variables, if $0 \leq \xi_m \leq 1$, then the m^{th} training example is correctly classified; if $\xi_m \geq 1$, then the m^{th}th training example is wrongly classified. $C > 0$ is the *soft-margin parameter*, or *penalty parameter*, which is used to balance the goals of maximising the margin and minimising the amount of misclassifications. The bigger C is, the larger penalty will be assigned to misclassifications, the more sensitive the hyperplane will be to outliers. Soft-margin hyperplanes are employed in Chapter 5 and Chapter 7 to improve the performance of SVM classifiers on the validation datasets.

On the one hand, the constraints $y^{(m)}(w^T x^{(m)} + b) \geq 1 - \xi_m$ and $\xi_m \geq 0$ can be combined to $\xi_m = \max(0, 1 - y^{(m)}(w^T x^{(m)} + b))$ [245], which further transforms the optimisation problem to:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{m=1}^{M} \max(0, 1 - y^{(m)}(w^T x^{(m)} + b)). \tag{3.13}$$

The optimal (w^*, b^*) for above problem can be obtained using the gradient-based methods, such as [245].

On the other hand, the dual form of the primal problem can be obtained using Lagrange multipliers α_m [220]:

$$\max_{\alpha} \sum_{m=1}^{M} \alpha_m - \frac{1}{2} \sum_{m,m'=1}^{M} y^{(m)} y^{(m')} \alpha_m \alpha_{m'} \langle x^{(m)}, x^{(m')} \rangle$$

$$\text{subject to: } 0 \leq \alpha_m \leq C, \ \sum_{m=1}^{M} \alpha_m y^{(m)} = 0. \tag{3.14}$$

The optimal $w^* = \sum_{m=1}^{M} \alpha_m y^{(m)} x^{(m)}$. $\alpha_m > 0$ is obtained only for training examples that have $\hat{s}^{(m)} = 1$, and these training examples, which are called *support vectors*, are the ones closest to the hyperplane. The optimal values for α_m can be resolved through the Sequential Minimal Optimisation (SMO) algorithm [220], which can be applied to resolve w^* and b^*.

For a new input vector x, the prediction label:

$$y = sgn({w^*}^T x + b^*) \tag{3.15}$$

The SVM algorithm can be written entirely in terms of the inner products $\langle x^{(m)}, x^{(m')} \rangle$. Given a feature mapping ϕ, which maps input vectors x into high-dimensional space, it is feasible to get the SVM algorithm to learn in this space without explicitly representing $\phi(x)$, which may be expensive to calculate because of its high dimensionality. One only need to replace the $\langle x^{(m)}, x^{(m')} \rangle$ with $K(x^{(m)}, x^{(m')}) = \langle \phi(x^{(m)}), \phi(x^{(m')}) \rangle$, where $K(\cdot, \cdot)$ represents the kernel function, which is much easier to compute. This is the *kernel trick* mentioned in [44, 114]. The kernel trick is very helpful when it is hard to separate the training set linearly in the original input space, because the transformed input vectors, obtained by the feature mapping ϕ, may be linearly separable in the high-dimensional space. Example kernel functions include Linear, Polynomial, Gaussian/Radial Basis Function (RBF), which is usually selected through cross validation and grid search. Only the RBF kernel $K(x^{(m)}, x^{(m')}) = \exp(-\gamma \parallel x^{(m)} - x^{(m')} \parallel^2)$ is employed in this thesis. the kernel parameter γ controls the width of Gaussian, which further controls the flexibility of the decision boundary [35]. When γ is small, the classification of each training example is influenced by all the support vectors, thus the decision boundary is smooth; when γ increases, the locality of the support vector expansion also increases, and the classification of each training example is mainly influenced by its “close” support vectors, which results in greater curvature and a more flexible decision boundary.

The training of an SVM is a convex quadratic programming problem, which means the solution is guaranteed to be unique and globally optimal. The optimality is only influenced by “difficult points” that are close to the decision boundary. Besides, the overfitting problem can be controlled by tuning the penalty parameter C. However, the SVM classifier is still inefficient in terms of the required number of training examples and adaptable components to represent certain types of function families, compared with multilayer neural network-based classifier [36]. For multi-classification scenarios, multiple SVM classifiers have to be trained and then apply the one-vs-one or the one-vs-all scheme [95], which is not intuitive and adequate to model the interactions of input vectors across different categories. In this thesis, I demonstrate that the SVM classifiers are efficient in classifying if a Wikipedia sentence is reputation-influential (Chapter 7), and if a tweet is discussing a real-world event (Chapter 5), after being equipped with delicate feature engineering and a hierarchical classification strategy.

3.3.3 Multilayer Perceptron Classifier

MLP is a feedforward neural network model consisting of multilayers of perceptrons with non-linear activation functions. For the binary classification scenario, similar to SVM, the *Perceptron* algorithm [237] also tries to find a linear hyperplane in the input space, represented by a weight vector and a bias term (a single perceptron), that can separate the training examples from two classes. However, it simply tries to classify all the training examples correctly by iteratively updating the weight vector and the bias term, according to the mistake-driven learning approach. The Perceptron algorithm will only converge if the training examples are linearly separable and the convergence process can be very slow. Unlike the SVM algorithm, it is unstable to the perturbations of the input vectors. For training examples that are not linearly separable, one can either apply the kernel trick [114] introduced in Section 3.3.2, or use a combination of multiple perceptrons and non-linear activation functions.

The MLP consists of three kinds of layers: the Input layer, the Output layer, and the Hidden layer. Each node/perceptron/neuron in MLP is fully connected to the nodes in adjacent layers [134]. Figure 3.5 depicts a sample MLP with a single hidden layer, as an example. [120] has proven that a single hidden layer is enough to make MLP a universal approximator.

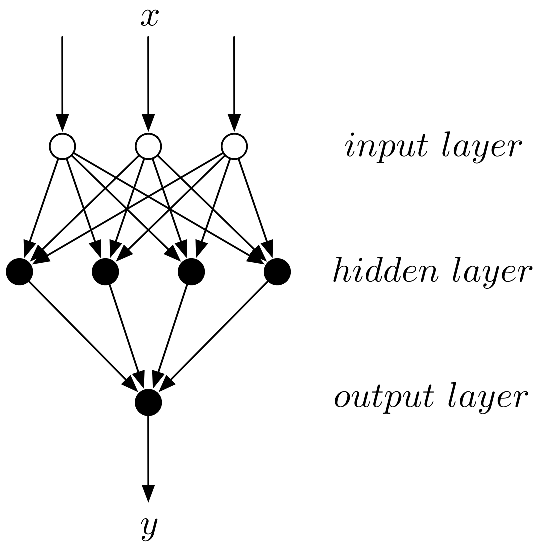

Figure 3.5: Multilayer perceptron with one hidden layer.

In Figure 3.5, the number of nodes in the input layer equals to $d_1 + 1$, where d_1 is the dimension of the input vector $x \in \mathbb{R}^{d_1}$, and 1 represents the extra bias node. The number of nodes in the output layer equals the dimension of the output vector $y \in \mathbb{R}^{d_3}$. Each edge in Figure 3.5 is associated with a weight or bias, and the black nodes represent the operation that calculates the sum of the weighted input from the former layer and applies a non-linear transformation, using the activation function. Following [1], the output y can be calculated as follows:

$$y = g(b_2 + W_2(f(b_1 + W_1 x))). \tag{3.16}$$

Assume the number of nodes in the hidden layer equals to d_2. In the above equation, g and f are non-linear activation functions, which may be selected amongst the threshold function, the piecewise linear function, the sigmoid function, the Gaussian function, the softmax function, etc. [134], according to different application scenarios. $W_1 \in \mathbb{R}^{d_2 \times d_1}$, $b_1 \in \mathbb{R}^{d_2}$, $W_1 \in \mathbb{R}^{d_3 \times d_2}$ and $b_2 \in \mathbb{R}^{d_3}$ are the weights and bias terms to train, respectively. Various optimisation algorithms, which includes *batch training* methods, such as the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm, the Conjugated Gradient (CG) algorithm and the

Gradient Descent (GD) algorithm, as well as *online training* methods, such as the Stochastic Gradient Descent (SGD) algorithm, have been applied to train the neural networks [163, 203]. The batch training methods require computations with respect to the entire training examples to make an update, which is very computationally expensive; the optimisation processes are likely to be stuck at local minima, because of the lack of fluctuation and randomness. The SGD algorithm [45] for online training is more widely used, as it is simple to implement and fast to converge, especially when facing a large number of training samples. With $\theta = \{W_1, W_2, b_1, b_2\}$ denoting the set of parameters to train, the SGD algorithm tries to minimise an objective function $L(\theta)$, by updating θ in the opposite direction of the gradient of $L(\theta)$ with respect to θ for each training example $(x^{(m)}, y^{(m)})$. Commonly employed objective functions are *Sum-of-squares Error* and *Cross-entropy* [39]. The Sum-of-squares Error is applied to measure the squares of the absolute errors between the outputs and the target values; the Cross-entropy is applied when the outputs are the estimated posterior distributions over the classes to measure the distances between the estimated distributions and the true distributions. Studies [98, 147] have proven that the Cross-entropy loss outperformed the Sum-of-squares Error loss in classification scenarios. The updated parameters are given by:

$$\theta = \theta - \eta \cdot \nabla_\theta L(\theta; x^{(m)}, y^{(m)}), \tag{3.17}$$

where η is the learning rate, $\nabla_\theta L(\theta; x^{(m)}, y^{(m)})$ is the gradient of the objective function with respect to θ for the training example $(x^{(m)}, y^{(m)})$, which can be calculated through the back-propagation method [287]. To reduce the variance in parameter update and employ optimised matrix operations, in practice, the SGD algorithm often updates the parameters with respect to a few training examples (mini-batches).

With varying settings, the MLP classifier is more efficient, powerful and flexible than the SVM classifier [36], especially after stacking with various neural network structures for content representation, as discussed in Section 3.2.3. The researchers in [63] have proven that a regularised perceptron with hinge loss as the objective function is equivalent to a linear SVM. Because MLP itself is a universal approximator, various kernel tricks are not necessary to model the interactions between different dimensions of the input vector. An MLP classifier with the *softmax* activation function for the output layer is able to generate the probability distribution of the input belonging to various classes, and thus can be directly employed for multi-classification scenarios. However, the training of an MLP is a non-quadratic, non-convex problem, with many local minima on the surface of the objective function (the training of a single perceptron, however, is a convex problem). Several tricks, such as Momentum, Nesterov Accelerated Gradient (NAG), adaptive learning rates and scientific weight initialisation have been exploited to improve the convergence. In MLP, the optimality is influenced by all the training examples, thus is more prone to overfitting. To alleviate this problem and improve the generalisation performance of neural networks, several techniques have been proposed, such as *Early Stopping* [224], *Weight Decay* [152] and *Dropout* [253]. In this thesis, the state-of-the-art performance is achieved by stacking and jointly training the MLP classifier with the attention-based neural network structure to perform target-specific classification (Chapter 8).

3.4 Conclusion

In this chapter, I have introduced the framework employed by the works in this thesis, to perform text mining on social media. In addition, I have presented various content representation and classification techniques, and summarised their advantages and disadvantages, as well as their suitable application scenarios. In the following chapters, more detailed descriptions about how these techniques are applied in my works will be presented, as well as the developments I propose to achieve better performance to meet the diverse information needs.

Chapter 4

Analysing Entity-centric Topic Representations of Multilingual Wikipedia

In this chapter, I present a novel model, which I have created for applying topic representations for entities of multilingual Wikipedia to provide language-specific results for entity-centric information retrieval, to answer *RQ1. Can language-specific topic representations be constructed for entities employing knowledge from multilingual Wikipedia?* This chapter and Chapter 5 represent the part of the study on topic analysis of social media text. Other studies on Wikipedia can be found in Chapter 6 and Chapter 7. The work in this chapter has been published in [320,322].

4.1 Introduction

Topic representations of entities with worldwide influence, such as celebrities and multinational corporations, can vary greatly on web pages or in other documents originating from various cultures and written in different languages. These various topic representations can reflect language-specific facts and views on the entity in different language-speaking communities. In order to enable better language-specific topic representations for entities to support entity-centric information retrieval applications, methods to systematically identify an entity's topical aspects, i.e. the key talking points related to the entity typical in a specific language, need to be developed.

For example, in English news articles, the entity "Angela Merkel", the Chancellor of Germany, is often associated with US and UK politicians, such as "Barack Obama" and "David Cameron". Also, discussions of European importance, such as the Greek financial situation, are included. In contrast, although the news articles from German media also include European topics, they frequently focus on the domestic political topics, featuring discussions of political parties in Germany, scandals around German politicians, local elections, finances and other country-specific topics. Taking another example, in the case of multinational companies, such as GlaxoSmithKline (a British healthcare company), topics related to its local activities are prevalent in the news articles in specific languages. These topics range from the effectiveness of the various vaccines developed by the company to the sports events sponsored by this company in a specific country.

In this chapter, I focus on the problem of creating entities' language-specific topic representations to support entity-centric, language-specific information retrieval applications.

To create language-specific topic representations for an entity, I need to obtain comprehensive multilingual contexts of this entity. I have chosen Wikipedia as a knowledge base, to obtain such contexts. Over the recent years, Wikipedia has expanded into a large and much-used source of information on the Web (with almost 24 million users, and growing at a rate of 10 edits/sec by editors from all around the world[1]). Wikipedia is currently available in more than 280 different language editions[2] that are being increasingly linked. Different language editions of Wikipedia contain language-specific descriptions of millions of entities and can provide a rich source for cross-cultural analytics. For example, recent studies [30, 49, 113, 181, 235] have discovered the content differences of multilingual Wikipedia articles discussing the same topic. As an entity's descriptions in different Wikipedia language editions can evolve independently, they often include overlapping, as well as language-specific topical aspects. Different ways of creating contexts for an entity using Wikipedia are discussed in Section 4.2, including *Article-based* and *Graph-based* approaches. I propose a similarity measure to analyse the similarities and the differences of the entity's language-specific topic representations derived from its Wikipedia contexts, in a case study using 219 entities of four different entity types from five languages. This work demonstrates the benefits of constructing entity topic representations,

[1] http://en.wikipedia.org/wiki/Wikipedia:Statistics

[2] http://meta.wikimedia.org/wiki/List_of_Wikipedias

illustrating the principles and methods, which can then be transferred and generalised to other languages and other entities. My experiments in Section 4.4 show that the proposed *Graph-based* approach can effectively provide comprehensive, yet accurate, contexts to derive the language-specific topic representations.

Moreover, I propose a *context-based information retrieval model* in Section 4.3, which applies the entities' language-specific topic representations to support entity-centric information retrieval applications. When using entity names as queries, traditional keyword matching-based information retrieval models cannot achieve desirable performance, because there is little contextual information provided in the query. The proposed model augments the query with the semantic knowledge extracted from Wikipedia, i.e., its topic representations, thus enabling the retrieval of the documents that describe information relevant to the entity, even if the entity is not mentioned by name. This information can include relevant events, which are likely to impact on the entity, or are otherwise related to it. At the same time, while using this model, the precision of the retrieved documents is only marginally reduced. I have implemented the proposed model on an information retrieval application, which includes: (i) targeted retrieval of entity-centric information using language-specific topic representations; (ii) an overview of the language-specific topical aspects in each retrieved document that is relevant to the query. I have performed a case study in Section 4.5, to demonstrate the impact of this model in the context of news articles retrieval through the application. The results illustrate that the entity's topic representations, especially the ones derived from the *Graph-based* contexts, can enhance the recall of the information retrieval application, while keeping high precision, providing positive results are news articles that describe current events *relevant* to the entity, without having to mention the entity explicitly (further details explained in Section 4.5.1). I further propose *Language Specificity* to measure the level of language specificity of the proposed information retrieval model when serving users with different language backgrounds. The results show that the context-based information retrieval model is able to provide highly specific language-based results through exploiting language-specific topic representations.

4.2 Creation of an Entity's Language-specific Topic Representations

In this section, I define an entity's *language-specific topic representations*, present a similarity measure of the topic representations and discuss alternative ways to create language-specific contexts for the entity from the multilingual Wikipedia, from which to derive its language-specific topic representations.

4.2.1 Definition of an Entity's Language-specific Topic Representations

Entity-centric topic representations reflect the contextual topical information of the entity, with each dimension corresponding to a topical aspect. I define the entity's language-specific topic vector representation as follows:

Definition 1 *The topic representation* $\mathbf{r_{e,n}}$ *for the entity* e *of the language* l_n *is represented as a vector, where* $n \in \{1, \ldots, N\}$*. Each dimension of the vector corresponds to a topical aspect* a_k *that is relevant to* e*, where* $k \in \{1, \ldots, K\}$*. Concretely, it can be represented as follows:*

$$\mathbf{r_{e,n}} = (r_{e,n,1}, \ldots, r_{e,n,K}). \tag{4.1}$$

In [176], researchers pointed out that individual terms are not effective to represent the semantic information of documents, because of the noise associated with semantic ambiguity; better performance was achieved when representing documents with multi-phrase features. In [122], researchers extracted noun phrases from reviews as aspects associated with each product, for a more fine-grained sentiment analysis. Inspired by the above works, I propose that the entity's relevant topical aspects are *noun phrases* that occur in the entity's contexts in various languages. As demonstrated in [122,176], this method not only tackles the problem of semantic ambiguity by employing the phrasal information, but also diminishes the noise from terms carrying little topical information, such as adjectives and adverbs. Additionally, since only the noun phrases are retained for each context, the dimensionality of its vector representation is reduced to reduce the computational cost.

The weights of the topical aspects are based on two factors: (1) the language-specific *aspect frequency* — the frequency of co-occurrences of the topical aspect and the entity in a language, and (2) the *language frequency* — the number of languages in which the entity contexts contain the topical aspect. The first weighting factor prioritises the topical aspects that frequently co-occur with the entity in a particular language. The second factor assigns higher weights to the language-specific topical aspects of the entity rarely mentioned in other languages.

Inspired by the term frequency–inverse document frequency ($tfidf$), which is discussed in Section 3.2.1, given a multilingual data collection, the weight $r_{e,n,k}$ is calculated as follows:

$$r_{e,n,k} = af_{e,n,k} \times log\frac{N}{lf_{e,k}}, \tag{4.2}$$

where $af_{e,n,k}$ is the language-specific *aspect frequency*, which represents the frequency of the co-occurrences of the topical aspect a_k and the entity e in the context in language l_n; N is the number of languages in the multilingual collection; $lf_{e,k}$ is the *language frequency*, which represents the number of languages in which the contexts of e contain the topical aspect a_k.

4.2.2 Similarity Measure Between Topic Representations

The similarity between entity e's topic vector representations of languages l_1 and l_2 is computed as their cosine similarity:

$$Sim(\mathbf{r_{e,1}}, \mathbf{r_{e,2}}) = \frac{\mathbf{r_{e,1}} \cdot \mathbf{r_{e,2}}}{|\mathbf{r_{e,1}}| \times |\mathbf{r_{e,2}}|}. \tag{4.3}$$

In order to allow for cross-lingual similarity computations, I represent the language-specific contexts in a common language, using machine translation. To simplify the description in this thesis, I always refer to the original language of the entity context, keeping in mind that all the contexts are translated to a common language.

4.2.3 Article-based Context Creation Approach

Wikipedia articles describing an entity in different language editions (i.e., the articles that use the named entity as titles) can be directly employed as contexts to generate the language-specific topic representations. Thus, I first employ a baseline *Article-based* context creation approach, which simply employs the articles describing the entity in different language editions of Wikipedia. Similar with [47,68,109,121,190, 280,289,303], I use all sentences from an article describing the entity in a language edition as the only source of the *Article-based* language-specific context for this entity.

One drawback of this approach is the possible limitation of the topical aspects coverage due to the incompleteness of the Wikipedia articles. Such incompleteness can be more prominent in some language editions, making it difficult to create fair cross-lingual comparisons. For example, when reading the English Wikipedia article about the entity "Angela Merkel", a lot of basic facts about this politician, such as her background and early life, her domestic policy and her foreign affairs, are provided. However, not all topical aspects about Angela Merkel occur in this Wikipedia article. It can be observed that other articles in the same Wikipedia language edition mention other important facts. For example, the Wikipedia article about "Economic Council Germany" mentions Angela Merkel's economic policy: "Although the organisation is both financially and ideologically independent it has traditionally had close ties to the free-market liberal wing of the conservative Christian Democratic Union (CDU) of Chancellor Angela Merkel.". Even the English Wikipedia article about an oil painting, "The Nightmare", which does not seem connected to "Angela Merkel" at the first glance, also mentions "Angela Merkel" as: "On 7 November 2011 Steve Bell produced a cartoon with Angela Merkel as the sleeper and Silvio Berlusconi as the monster." The topical aspects contained in the examples above do not occur in the English Wikipedia article entitled "Angela Merkel". As this example illustrates, just employing the Wikipedia article describing the entity can not entirely satisfy the need to obtain a comprehensive coverage of the language-specific topical aspects.

4.2.4 Graph-based Context Creation Approach

To alleviate the shortcomings of the *Article-based* approach presented above and obtain a more comprehensive coverage of the entity's topical aspects in the entire Wikipedia language edition (rather than in a single article), I propose the *Graph-based* context creation approach. The idea behind this approach is to use the link structure of Wikipedia to obtain a comprehensive set of articles, which may mention the target entity and to use this set to create the context. To this extent, I use the *in-links* to the Wikipedia article describing the entity and the *language-links* of this article to efficiently collect the articles that are likely to mention the target entity in different language editions. I extract the sentences mentioning the target entity

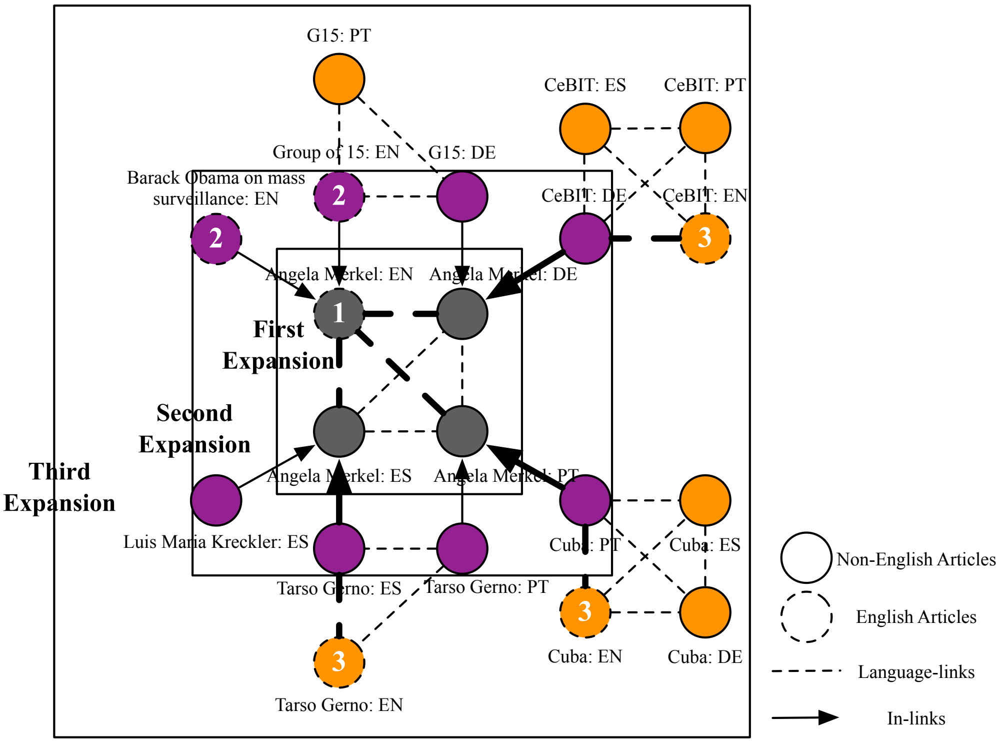

Figure 4.1: An example of the Graph-Based context creation of the English Wikipedia for the entity "Angela Merkel".

in these articles using the state-of-the-art named entity disambiguation method and use these sentences to form language-specific contexts.

To illustrate my approach, I use the creation of the context of the English edition of Wikipedia for the entity "Angela Merkel", as an example. For the Wikipedia article in English entitled "Angela Merkel", there are several *in-links* from other articles in English that mention the entity. Besides that, there are also *language-links* from the articles describing "Angela Merkel" in other Wikipedia language editions to this entity's English Wikipedia article.

In Figure 4.1, I use the arrows to represent the *in-links*, and dashed lines to represent the *language-links*. The nodes with dashed edge lines represent articles in English Wikipedia, the nodes with solid edge lines represent articles in non-English Wikipedia. All the nodes are annotated with the titles and languages of their corresponding Wikipedia articles.

Overall, the creation of the *Graph-based* English context for "Angela Merkel" using these links includes the following steps:

1. *Graph Construction.* I construct a subgraph for "Angela Merkel" from Wikipedia's link structure in the following way: I first expand the node set from the article in English describing the entity (the central node) to all language editions of this Wikipedia article (nodes in black colour in Figure 4.1); I further expand the node set with all the articles having *in-links* to the nodes in the node set (nodes in purple colour in Figure 4.1); I finally expand the node set with all the articles having *language-links* to the existing nodes in the node set, if they have not been included in the node set yet (nodes in orange colour in Figure 4.1). Different types of edges are also added between the nodes based on the *in-link* and the *language-link* relationships.

2. *Article Extraction.* To efficiently extract as many mentions of Angela Merkel from the English Wikipedia as possible, I first extract the article of the central node (e.g., the one annotated with number 1 in Figure 4.1), and then start traversing the Wikipedia link structure from this node.

   Second, all the articles in the graph that have paths of length 1, and the path types are *in-link* to the central node (e.g., the one annotated with number 2 in Figure 4.1), are extracted.

   Third, all the articles in the graph that have paths of length 3, which are in English, and the path types are *language-link* — *in-link* — *language-link* (marked as bold lines in Figure 4.1) to the central node (e.g., the one annotated with number 3 in Figure 4.1), are also extracted. These articles, although they do not have the direct *in-link* paths to the central node, are in English and their other language editions have *in-links* to articles describing "Angela Merkel" in other languages. Therefore, these articles are likely to mention "Angela Merkel". In this way, I tackle the "missing links" problem raised in [30]. The extracted articles contain uninformative metadata, such as HTML tags, references and sub-titles. Therefore, I eliminate these metadata, to obtain plain text for each selected Wikipedia article.

3. *Sentence Extraction.* DBpedia Spotlight [185] is employed, to annotate the extracted articles, to identify the sentences mentioning the target entity "Angela Merkel". DBpedia Spotlight uses the DBpedia Lexicalisation dataset to provide candidate disambiguations for each surface form in the text, and a vector space model to find the most likely disambiguation. All these sentences form

the English *Graph-based* context of Angela Merkel. This step cannot be replaced by string matching, because the target entity can have different *surface forms* in the extracted articles. For example, the surface forms of the entity "Angela Merkel" include: "Angela Merkel", "Angela Dorothea Kasner", "Chancellor of Germany", "Angela Dorothea Merkel", "Angela", "Merkel", "Chancellor Angela Merkel" and "Ms. Merkel". Besides, it is also importance to distinguish the entities that have common surface forms. For example, the surface form "Merkel" can be used to refer either to a person or to a town in the United States, depending on the surrounding text. To reduce the length of texts to be annotated by DBpedia Spotlight, an extra pre-selection step can be performed, by discarding sentences which don't mention any surface form of the target entity. The surface forms of an entity can be extracted from its DBpedia page, as in `http://dbpedia.org/page/Angela_Merkel`. This step only affects the DBpedia Spotlight's performance marginally, as this is also DBpedia Spotlight's first step of Entity Disambiguation [185].

To generate an entity's topic representations based on the *Graph-based* context, besides all the noun phrases extracted from sentences mentioning the entity, I also include the names of all the extracted articles mentioning the entity as topical aspects. I assign these article names with an average language-specific *aspect frequency* computed for the noun phrases, in order to calculate their weights in the topic representations.

The comprehensiveness of the contexts created by the baseline Article-based approach and the proposed Graph-based approach will be examined and compared in Section 4.4 and Section 4.5.

4.3 News Retrieval Using an Entity's Language-specific Topic Representations

In this section, I present the retrieval scenario of searching for relevant articles over news collections in a common language using an entity name as the query. Then I describe my approach that addresses the entity-centric search, using the entity's language-specific topic representations, as presented in Section 4.2.

4.3.1 Entity-centric News Retrieval

When users are interested in the current news about a named entity, they could simply provide the entity name as the query to a retrieval application. This entity-centric retrieval scenario is also referred as *querying by entities* in [316].

On a daily basis, only a limited number of news articles that explicitly mention this named entity are published. However, one named entity is typically related to various other topical aspects, as I observed during the context creation using the Wikipedia link structure. This kind of relationship with topical aspects is demonstrated by the entity-centric topic representations, which I described in Section 4.2. The motivation is that by using these topic representations, I could significantly increase recall of the retrieved documents for the entity-centric queries in a news retrieval application, while keeping high precision. Moreover, some documents could only marginally mention an entity, without providing any comprehensive information for the specific entity. In these cases, the entity's topic representations can help the retrieval application to focus on more relevant documents.

When only using an entity name as the query, traditional information retrieval systems that are based on keyword matching can only return news articles with the named entity's occurrence, which can barely satisfy the users' needs of comprehensive knowledge about the named entity. For example, when using "Angela Merkel" as the query, it would be beneficial to return news articles like `http://www.thelocal.de/20151202/germany-fear-terrorism-if-army-fights-in-syria`, which describe the situation in Germany. Although the content of this article contains neither the term "Angela" nor the term "Merkel", it reports about an event that has a potentially large impact on her political decisions. In order to tackle this problem, my context-based information retrieval model incorporates the entity's contextual topic representations from Wikipedia into the search and ranking process. As a result, the articles discussing similar topical aspects as the entity's context will obtain higher ranks, even if the entity is not mentioned explicitly.

While using the entity's topic representations for retrieval applications, the relevance of a news article to a named entity may be controversial among people with different language backgrounds. For example, a news article containing information about the VW scandal affecting the biggest German car production company `http://www.thelocal.de/20151202/what-the-vw-scandal-means-`

`for-germanys-economy` could be considered as relevant by most German people, as they could think that the German Chancellor should take direct measures to boost the national economy hurt by the scandal. However, the relevance of this article to the query “Angela Merkel” can be considered to be low among the English-speaking communities. These users could think this to be a company problem, and it could be hard for them to understand if this scandal would have a big impact at the national level. I tackle this problem by using the entity’s language-specific topic representations in news retrieval. The users of the retrieval application can select the topic representation of their preferred language when searching for a named entity. The returned news articles and their ranks are then language-specific, based on the background knowledge from the corresponding language edition of Wikipedia.

Besides the retrieval of relevant articles, it is also useful to provide information regarding the topical aspects of the entity influencing their relevance. That is particularly important in case the entity itself is not mentioned in the article. The proposed context-based information retrieval model addresses this problem by creating an overview of each news article discussing language-specific topical aspects related to the entity.

4.3.2 Context-based Entity-centric Information Retrieval Model

For the news article document d where $d \in \{1, \dots, D\}$, I extract all the *noun phrases* in the document as potentially related *topical aspects* to query entities (the named entities whose names are provided as the queries), and then index all the documents by the *topical aspects*. For a query entity e, I generate a query-specific vector representation for the document d, with aspect a_k weighted by:

$$s_{e,d,k} = af_{e,d,k} \times log\frac{N}{lf_{e,k}}, \tag{4.4}$$

where $k \in \{1, \dots, K\}$, $af_{e,d,k}$ is the number of matches of topical aspect a_k with the noun phrases from document d. In this way, document d’s entity-specific vector representation is $\mathbf{s_{e,d}} = (s_{e,d,1}, \dots, s_{e,d,K})$.

I apply the same vector space model and similarity metric as in Section 4.2.2, to compute the similarity between entity e’s topic representation of language l_n, denoted by $\mathbf{r_{e,n}}$, and document d’s entity-specific representation $\mathbf{s_{e,d}}$:

$$Sim(\mathbf{r_{e,n}}, \mathbf{s_{e,d}}) = \frac{\mathbf{r_{e,n}} \cdot \mathbf{s_{e,d}}}{|\mathbf{r_{e,n}}| \times |\mathbf{s_{e,d}}|}. \tag{4.5}$$

The above similarity will be used to measure the levels of relevance between the query entity and documents under this setting. All the documents' entity-specific representations, which have similarities with $\mathbf{r_{e,n}}$ that are higher than a threshold, will be returned. Their ranks will also be decided based on $Sim(\mathbf{r_{e,n}}, \mathbf{s_{e,d}})$.

The top weighted topical aspects in the document d will be returned to provide an overview of what topical aspects this document is discussing the query entity e.

4.4 Analysis of an Entity's Language-specific Topical Representations

The goal of this analysis is to compare the entity's topical representations derived from the *Graph-based* and the *Article-based* contexts. To this extent, I analyse the similarities and the differences of the language-specific topic representations in a case study.

4.4.1 Dataset Description

For my study, I selected five European languages: English, German, Spanish, Portuguese and Dutch, as the target languages. As my approach requires machine translation of the contexts, to enable cross-lingual similarity computation between topic representations, I chose Google Translate[3]— one of the best public translation services for all the involved language pairs. To facilitate my analysis, I selected a total number of 219 famous entities with *worldwide influence* that came from four categories that Internet users were most interested in, or most frequently discussed about, as my target entities. These categories included: multinational corporations, politicians, celebrities and sports stars. For each category, I included entities originating from the countries that use one of these target languages as official languages. For example, the politician entities originating from Dutch speaking countries were selected from the top results returned by Google search, when using "politician + Dutch" as the query.

Based on the approach described in Section 4.2, I created the entity-centric contexts for my target entities from the five Wikipedia language editions listed above using the *Graph-based* and the *Article-based* approach. All the data on multilingual Wikipedia can be accessible through MediaWiki API[4]. The average number of

[3] https://translate.google.co.uk/

[4] http://www.mediawiki.org/wiki/API:Main_page

sentences in the contexts extracted from the Wikipedia article describing the entity using the *Article-based* approach was around 50 in my dataset. With the *Graph-based* context creation approach, which utilised Wikipedia link structure to collect sentences mentioning the entity from multiple articles, the number of sentences referring to an entity was increased by the factor 20 to more than 1,000 sentences per entity in a language edition, on average. This factor reflects the effect of the additional data sources within Wikipedia. The total number of sentences in the entity-centric contexts collected by the *Graph-based* approach is 1,196,403 for the whole dataset under consideration.

4.4.2 Topic Representation Similarity Analysis

I derived the topic representations for entities from the language-specific contexts, according to Section 4.2.1. The topical aspects were extracted by the Stanford POS tagger [269]. The similarity values between language-specific topic representations derived from the *Article-based* and the *Graph-based* contexts are presented in Table 4.1 and Table 4.2, respectively.

To enable cross-lingual similarity computation, I translated all the entity-centric contexts to English. Here I present example similarity values for four randomly selected entities (one per entity type) for seven language pairs. In addition, I present the average similarity and the standard deviation values based on all 219 target entities.

Table 4.1: Topic representation similarity based on the *Article-based* contexts.

Entity	**Language pairs**						
	EN-DE	**EN-ES**	**EN-PT**	**EN-NL**	**DE-ES**	**DE-NL**	**ES-PT**
GlaxoSmithKline	0.43	0.34	0.29	0.29	0.31	0.22	0.26
Angela Merkel	0.68	0.66	0.84	0.54	0.60	0.59	0.66
Shakira	0.71	0.58	0.84	0.75	0.48	0.64	0.58
Lionel Messi	0.71	0.86	0.81	0.89	0.71	0.68	0.82
Average of 219	**0.50**	**0.47**	**0.43**	**0.45**	0.38	0.38	0.37
Stdev of 219	0.16	0.17	0.19	0.19	0.15	0.16	0.17

Table 4.1 and Table 4.2 present the similarity values for four selected entities of different types, as well as the average similarity and the standard deviation for all the 219 target entities, based on the contexts created by the Article-based approach and the Graph-based approach, respectively. The language codes representing the

Table 4.2: Topic representation similarity based on the *Graph-based* contexts.

Entity	**Language pairs**						
	EN-DE	**EN-ES**	**EN-PT**	**EN-NL**	**DE-ES**	**DE-NL**	**ES-PT**
GlaxoSmithKline	0.72	0.73	0.59	0.61	0.63	0.62	0.55
Angela Merkel	0.64	0.62	0.42	0.60	0.75	0.82	0.51
Shakira	0.91	0.94	0.90	0.88	0.94	0.91	0.94
Lionel Messi	0.63	0.76	0.77	0.68	0.70	0.62	0.76
Average of 219	0.52	**0.59**	0.54	0.50	0.56	0.52	**0.64**
Stdev of 219	0.24	0.22	0.21	0.23	0.23	0.23	0.19

original context languages are as follows: “NL” — Dutch, “DE” — German, “EN” — English, “ES” — Spanish, and “PT” — Portuguese.

From Table 4.1, it can be observed that using the *Article-based* context creation approach, the average similarity values of the topic representations between language pairs including English are always higher than those between the other language pairs. Using these computation results, I can make several observations. First, as the topic representations of other languages are very similar to those of English, the English edition builds a reference for the creation of the articles in other language editions. This can be further explained by the fact that the English Wikipedia has the largest number of users, articles, and edits compared with other language editions[5]. Second, as the topic representations of other language pairs are less similar, the overlapping topical aspects between the English edition and the other language editions appear to be language-dependent. Finally, although the cosine similarity values can be any value in the interval [0,1], the absolute similarity values between language-specific topic representations derived from the *Article-based* contexts reach at most 0.5, even for the language pairs which are supposed to have relatively high similarity, such as English and German. Such relatively low absolute similarity values indicate that although the articles contain some overlapping topical aspects, they also include a significant proportion of divergent topical aspects.

In contrast to the *Article-based* contexts, the *Graph-based* contexts include more comprehensive topical aspects spread across different articles in a language edition. From Table 4.2, I can see the topic representations of Spanish and Portuguese are most similar among those of all language pairs. Intuitively, this could be explained by the closeness of the cultures using these two languages, and a more comprehensive coverage of the topical aspects from both languages, compared with the *Article-based*

[5] `http://en.wikipedia.org/wiki/List_of_Wikipedias`

contexts. I can also observe that the average similarity values significantly increase, compared with the similarity values in Table 4.1, and can exceed 0.6 in the dataset.

From a single entity perspective, its topic representations of specific language pairs may achieve higher similarity values than the average similarity, when more common topical aspects are included in the contexts in both languages. For example, this is the case for the EN-NL, DE-ES and DE-NL pairs for the entity "Angela Merkel". On the other hand, its topic representations for specific language pairs may achieve lower similarity values, especially when distinct topical aspects are included into the corresponding contexts, such as the EN-DE, EN-ES, and EN-PT pairs for "Lionel Messi". To illustrate the differences in the language-specific topic representations derived from *Graph-based* contexts, I select the highly weighted topical aspects of the entity "Angela Merkel" extracted from her *Graph-based* contexts, as shown in Table 4.3. In this table, the unique topical aspects that appear with high weights in each topic representation of the entity "Angela Merkel" are underlined. I can observe that the topical aspects that appear with high weights only in the topic representations of non-German languages, e.g. "England", "Kingdom" and "Dilma Rousseff", are more relevant to her international affairs in corresponding language-speaking countries. In contrast, the topical aspects that appear with high weights only in the topic representation of German, such as "German children" and "propaganda", are more relevant to her domestic activities.

Overall, my observations confirm that the *Graph-based* context provides a better knowledge source for the topical aspects of entities than the *Article-based* context. The topic representations derived from the *Graph-based* contexts can determine the similarity values and the differences with respect to the topical aspects related to the entity, independent of the coverage and completeness of any dedicated Wikipedia article. I also have performed the t-test to confirm the statistical significance of the differences in similarity values based on the *Article-based* contexts and the *Graph-based* contexts. The resulted p-values are: 1.93×10^{-1} (EN-DE), 1.71×10^{-11} (EN-ES), 3.55×10^{-10} (EN-PT), 1.25×10^{-3} (EN-NL), 1.79×10^{-23} (DE-ES), 2.38×10^{-26} (DE-PT), 3.81×10^{-17} (DE-NL), 4.65×10^{-45} (ES-PT), 2.22×10^{-40} (ES-NL), 3.85×10^{-39} (PT-NL). It shows that the differences are significant at the 0.01 level, for all language pairs except the EN-DE. This exception can be explained by a relatively high coverage of the German Wikipedia articles with respect to the topical aspects of the target entities.

The analysis results also confirm my intuition that, although the editors of dif-

Table 4.3: Top-30 highly weighted topical aspects of the entity “Angela Merkel” from the *Graph-based* contexts.

Language	**Topical aspects**
English	angela merkel, battle, berlin, cdu, chancellor, chancellor angela merkel, church, edit, election, emperor, empire, england, france, george, german, german chancellor angela merkel, germany, government, jesus, john, kingdom, merkel, minister, party, president, talk, union, university, utc, war
German	academy, angela merkel, article, berlin, cdu, cet, chancellor, chancellor angela merkel, csu, election, example, german, german chancellor angela merkel, german children, germany, government, kasner, merkel, minister, november, october, office, party, president, propaganda, ribbon, september, speech, time, utc
Spanish	administration, angela merkel, berlin, cdu, chancellor, chancellor angela merkel, coalition, council, country, december, decommissioning plan, decreed, election, energy, france, german, german chancellor angela merkel, german federal election, germany, government, government coalition, grand coalition, merkel, minister, october, party, president, spd, union, year
Portuguese	ali, angela merkel, bank, cdu, ceo, chairman, chancellor, chancellor angela merkel, china, co-founder, coalition, csu, dilma rousseff, german chancellor angela merkel, germany, government, government merkel, koch, leader, merkel, minister, november, october, party, petroleum, president, saudi arabia, state, union, york
Dutch	angela merkel, angela dorothea kasner, bundestag, candidate, cdu, chancellor, chancellor angela merkel, coalition, csu, december, fdp, fist, french president, german, german chancellor angela merkel, german christian democrat politician, german federal election, germany, government, majority, merkel, minister, november, october, party, president, right, spd, state, union

ferent Wikipedia language editions describe some common topical aspects for the same entity, they can have different focuses with respect to the topical aspects of interest. These differences are reflected by the complementary information spread across the Wikipedia language editions and can probably be explained by various factors, including the culture and the living environment of the editors, as well as the information available to them. The entity-centric topic representations derived from the *Graph-based* contexts are capable of capturing these differences from different language editions by creating comprehensive language-specific topical aspects overviews.

4.5 Language-specific Retrieval of News Articles for Entity-centric Queries

In this section, I discuss the impact of the entity's language-specific topic representation on a news retrieval application. Since results following the same patterns can be observed across all named entities, I have randomly selected two named entities, one originated from an English speaking country, and the other one originated from a non-English speaking country, as examples to demonstrate the effectiveness of the context-based information retrieval model.

4.5.1 Dataset Description

The two named entities I have chosen were: "Angela Merkel", originating from Germany, and "David Cameron", originating from Great Britain. To enable the comparison of different language-specific topic representations of an entity, I built two datasets, each containing daily news from different sources: the German media news dataset and the British media news dataset. For the German media news dataset, I randomly sampled 300 news articles from three mainstream online English news websites' RSS feeds published on December 2th, 2015 in Germany. These websites were: Deutsche Welle[6], Spiegel Online[7] and The Local[8]. Regarding the British media news dataset, I randomly sampled 300 news articles from two mainstream online English news websites' RSS feeds on December 10th, 2015 in Great Britain. These websites were: The Guardian[9] and Daily Express[10]. Then, I analysed the performance of topic representations of English and German for the entities "Angela Merkel" and "David Cameron" for these two datasets, respectively.

I expected that there were only a few news articles per day mentioning a specific entity, even if this entity was prominent. Nevertheless, daily news can contain many relevant articles that discuss events related to the entity. Therefore, I used the following criteria to annotate the articles as "Relevant":

1. *Is the named entity involved in this event?*

2. *Is the named entity one of the direct causes of this event?*

[6] http://www.dw.com/en/

[7] http://www.spiegel.de/international/

[8] http://www.thelocal.de/

[9] http://www.theguardian.com/

[10] http://www.express.co.uk/

3. *Will the named entity be directly impacted by this event?*

After the annotation, 51 news articles in the German media news dataset were annotated as “Relevant” for the query “Angela Merkel”. In the British media news dataset, 71 news articles were annotated as “Relevant” for the query “David Cameron”[11].

4.5.2 Precision-Recall Analysis

I used a state-of-the-art information retrieval model, BM25 [234] as a baseline. The baseline model retrieved the documents based on the number of matches of terms from the original query.

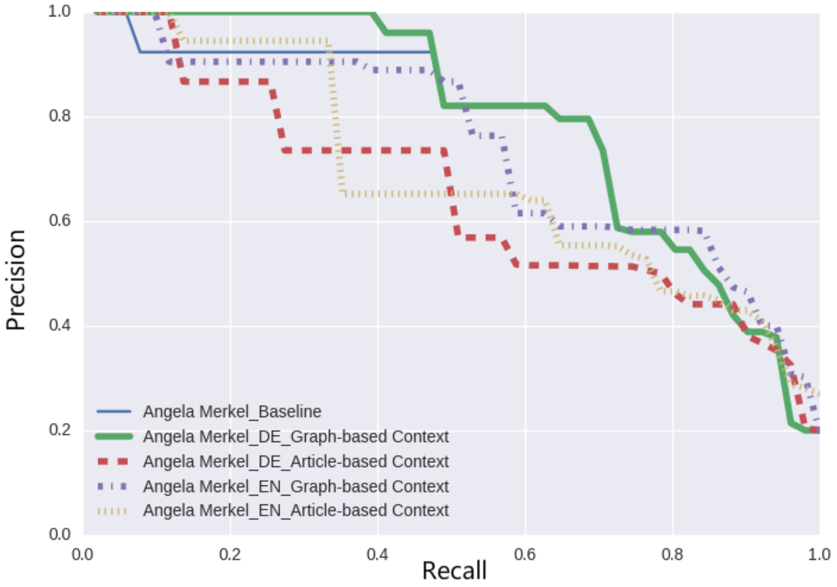

Figure 4.2: Precision-Recall curves of the baseline model and the context-based information retrieval model using different topic representations for “Angela Merkel”.

In Figure 4.2, I present the interpolated precision achieved by the baseline model, and the context-based information retrieval model using topic representations derived from various contexts at different recall levels for the query entity “Angela Merkel”. As it can be observed in Figure 4.2, although the traditional ranking algorithm based on the BM25 scores of the news articles given a query entity can maintain a relatively high precision, the highest

[11] The annotated datasets are accessible at `https://github.com/zhouyiwei/WIKIIRDATA`.

recall it can achieve is about 0.45. That is because a lot of news articles, such as http://www.thelocal.de/20151202/germany-to-send-1200-troops-to-aid-isis-fight, http://www.spiegel.de/international/europe/paris-attacks-pose-challenge-to-european-security-a-1063435.html and http://www.thelocal.de/20151029/germany-maintains-record-low-unemployment, report events either directly driven by "Angela Merkel", or would directly impact her. Although these articles do not mention the query entity by name, they provide indispensable insights into the query entity's current focus or past achievements, which the users issuing this query would consider them to be relevant, especially when the number of articles mentioning the query entity is small. The context-based information retrieval model using the entity's topic representations, no matter whether the topic representation is derived from *Article-based* contexts or *Graph-based* contexts, no matter whether they are extracted from English Wikipedia or German Wikipedia, achieved higher recall for this query.

I can also observe that the context-based information retrieval model using the topic representation derived from German (DE) *Graph-based* context achieves the overall best performance. For most of the time, it achieves higher precision than the ones using topic representations derived from other contexts, while achieving the same recall. This is because this topic representation provides a more comprehensive overview of the topical aspects related to "Angela Merkel".

Moreover, the context-based information retrieval model outperforms the baseline model with respect to precision at all recall levels for this query entity, when utilising the topic representation derived from the German (DE) *Graph-based* context. This is because "Angela" is quite a common term. By incorporating the background information from Wikipedia, the model can perform disambiguation implicitly, by differentiating the Chancellor of Germany from other celebrities, such as Angela Gossow (German singer) and Angela Maurer (German long-distance swimmer), which helps to increase the precision of retrieved results.

The baseline model ranks the news articles mostly based on the occurrences of terms in the query entity. In contrast, my model considers all the topical aspects mentioned in the news articles about the named entity. The ranks are generated based on the similarity values between the articles' entity-specific representations and the named entity's language-specific topic representation, such that news articles that provide a more comprehensive coverage of the entity's language-specific topical aspects are promoted to higher ranks.

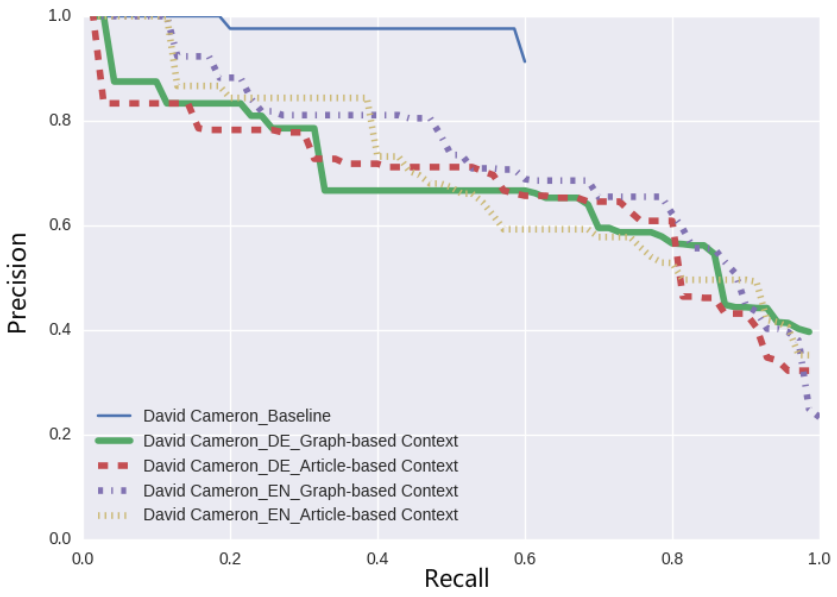

Figure 4.3: Precision-Recall curves of the baseline model and the context-based information retrieval model using different topic representations for “David Cameron”.

The effectiveness of the context-based information retrieval model can also be observed for the query “David Cameron”, presented in the Figure 4.3. As shown in Figure 4.3, the proposed model can achieve a much higher recall than the baseline model for this query as well, while maintaining high precision. As expected, the topic representation derived from the English (EN) *Graph-based* context, which is local for this query, helps the context-based information retrieval model to achieve an overall better performance than the topic representations derived from other contexts.

I did not observe significant differences when using the topic representations derived from the rest of the contexts for the query “David Cameron”. One of the reasons can be the numbers of topical aspects covered in these contexts. The topic representation derived from the English *Graph-based* context of the entity “Angela Merkel” contains 7,317 non-zero weighted topical aspects, the one derived from the German *Graph-based* context contains 6,614. Both of them contain much more non-zero weighted topical aspects than the ones derived from English and German *Article-based* contexts, which contain 562 and 1,069, respectively. Resulting from that, the topic representations derived from the German and English *Graph-based* contexts for "Angela Merkel" are much more "powerful" than the ones derived from German and English *Article-based* contexts. For "David Cameron", the most 'powerful' topic representation is derived from English *Graph-based* context, which contains 10,365 non-zero weighted topical aspects, whereas the numbers for the rest are much smaller and comparable. The topic representation derived from German *Graph-based* context for the entity "David Cameron" only has 1,627 non-zero weighted topical aspects; the numbers for the ones derived from his English and German *Article-based* contexts are 1,143 and 291. Although all of these topic representations can still help to greatly improve the recall while maintaining relatively high precision, their effectiveness is somewhat limited, because of their comprehensiveness.

4.5.3 Analysis of Language-specific Results

Table 4.4 presents the Top-8 results returned by the context-based information retrieval model using the topic representations derived from German and English *Graph-based* contexts of the query "Angela Merkel". As it can be observed, when using the topic representation derived from the German context, German local news such as `http://www.thelocal.de/20141001/german-cabinet-agrees-cap-on-rent-rises-cities` and `http://www.thelocal.de/page/view/german-astronaut-calls-for-peace-and-tolerance` are included in the top-ranked results, which is not the case when using the one derived from the English context. Nevertheless, as many topical aspects related to the entity are shared across both contexts, the results at the top of both rankings are similar.

To better understand the impact of the language-specific topic representations on the retrieved results, I define a measure: *Language Specificity. Language Specificity* (LS) is the percentage of unique documents in top-M results retrieved using two topic representations:

$$LS(R_{e,1}, R_{e,2}) = 1 - \frac{|R_{e,1} \cap R_{e,2}|}{2 \times M}, \tag{4.6}$$

where $R_{e,1}$ is the set of the results retrieved for the entity e using the topic representation $\mathbf{r_{e,1}}$, and $R_{e,2}$ is the set of the results retrieved for the entity e using the topic representation $\mathbf{r_{e,2}}$.

The higher the *Language Specificity*, the less overlapping there will be in the retrieved results, using language-specific topic representations and the more language-

Table 4.4: Top-8 results for the query "Angela Merkel" retrieved using topic representations derived from the German and English *Graph-based* contexts.

Rank	**URL & Topical aspects overview**	
	German	**English**
1	`http://www.spiegel.de/international/germany/angela-merkel-changes-her-stance-on-refugee-limits-a-1063773.html` (minister, idea, germany, merkel, chancellor)	`http://www.spiegel.de/international/germany/angela-merkel-changes-her-stance-on-refugee-limits-a-1063773.html` (minister, idea, germany, merkel, chancellor)
2	`http://www.thelocal.de/20151130/we-owe-future-generations-a-climate-deal-merkel` (prosperity, time, percent, paris, merkel)	`http://www.thelocal.de/20151202/german-forces-will-back-france-in-syria-fight` (bundeswehr, france, germany, thursday, syria)
3	`http://www.thelocal.de/20151202/german-forces-will-back-france-in-syria-fight` (bundeswehr, france, germany, thursday, syria)	`http://www.thelocal.de/20151130/we-owe-future-generations-a-climate-deal-merkel` (prosperity, time, percent, paris, merkel)
4	`http://www.thelocal.de/20151202/no-better-life-for-afghans-in-germany-merkel` (merkel, migration, dec, security, afghanistan)	`http://www.thelocal.de/20151030/the-sailors-who-brought-down-the-german-empire` (revolt, attack, government, battle, wilhelmshaven)
5	`http://www.thelocal.de/page/view/hamburg-bids-farewell-to-its-most-famous-son` (merkel, chancellor, schmidt, flag, terror)	`http://www.spiegel.de/international/germany/editorial-on-anti-refugee-sentiment-in-germany-a-1062442.html` (hitler, culture, germany, time, country)
6	`http://www.spiegel.de/international/germany/editorial-on-anti-refugee-sentiment-in-germany-a-1062442.html` (hitler, culture, germany, time, country)	`http://www.thelocal.de/20151202/no-better-life-for-afghans-in-germany-merkel` (merkel, migration, dec, security, afghanistan)
7	`http://www.thelocal.de/20141001/german-cabinet-agrees-cap-on-rent-rises-cities` (percent, average, law, oct, property)	`http://www.thelocal.de/page/view/hamburg-bids-farewell-to-its-most-famous-son` (merkel, chancellor, schmidt, flag, terror)
8	`http://www.thelocal.de/page/view/german-astronaut-calls-for-peace-and-tolerance` (publicity, vogel, space, space station, photo)	`http://www.thelocal.de/20151202/less-than-half-of-german-jets-ready-for-action` (report, syria, germany, wednesday, dec)

specific the retrieved documents will be.

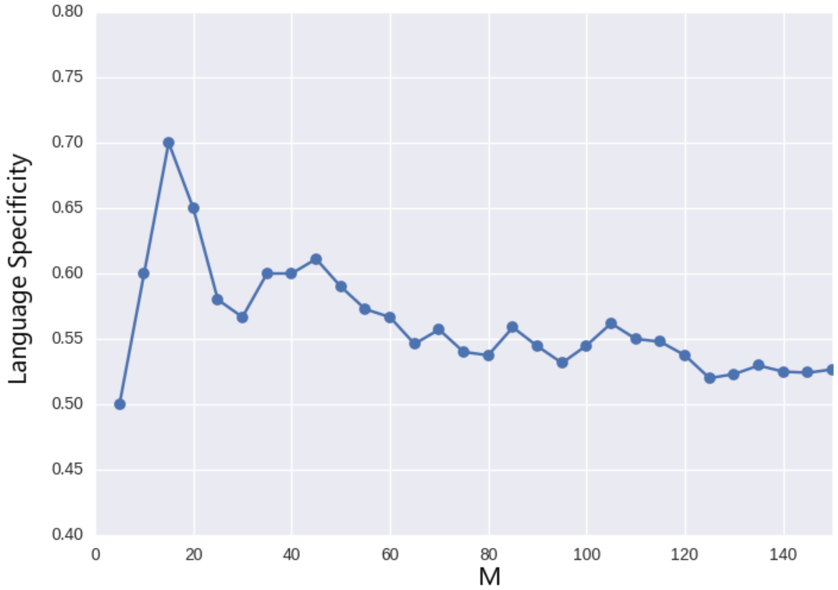

Figure 4.4: *Language Specificity* of the top-M retrieved results using topic representations derived from the German and English *Graph-based* contexts of the query "Angela Merkel".

Figure 4.4 illustrates the trends of the *Language Specificity* with an increasing number of returned results, when using topic representations derived from the German and English *Graph-based* contexts of the query entity "Angela Merkel". Whereas the most relevant results are very similar when using both topic representations, the *Language Specificity* of this pair reaches its maximum of 0.7 when M=15. This means that these language-specific topic representations can help to retrieve distinct and relevant news articles at lower M values. Then, with an increasing M, both relevance and distinctiveness of the retrieved results drop, but the *Language Specificity* still stays above 0.5. On the one hand, this is because many non-zero weighted topical aspects in these two topic representations overlap (as shown in Table 4.2, the similarity value between topic representations derived from English and German *Graph-based* contexts is 0.64). On the other hand, the most relevant news articles have been included in the retrieved results already by lower M values. With an increasing M, divergent articles with lower relevance are further retrieved.

Similar trends can be observed in Figure 4.5 for the query "David Cameron".

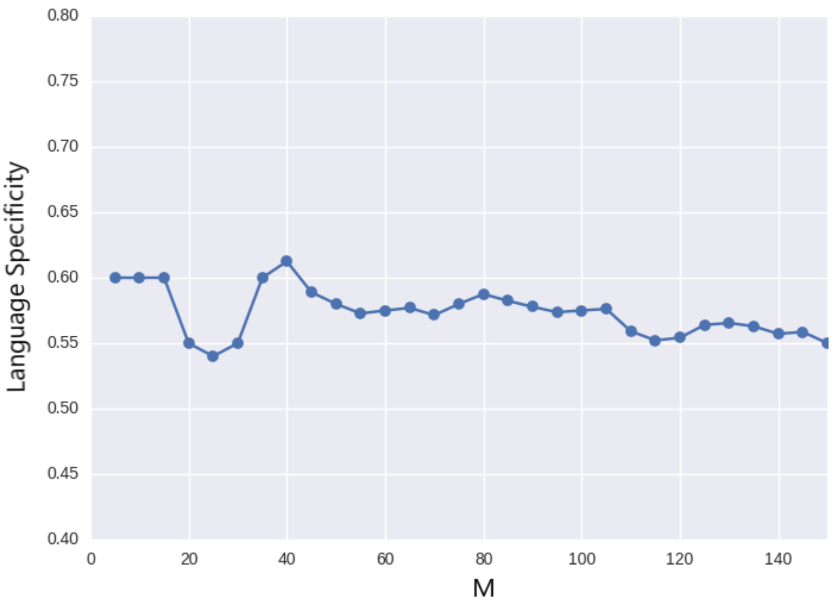

Figure 4.5: *Language Specificity* of the top-M retrieved results using topic representations derived from the German and English *Graph-based* contexts of the query "David Cameron".

4.6 Related Work

As mentioned in Section 2.2, due to its coverage and diversity, Wikipedia has been acting as an outer knowledge source to build semantic representations of entities and documents in various areas. Examples include information retrieval [83, 190, 208], named entity disambiguation [47, 68, 108, 109, 140, 153], text classification [280] and entity ranking [139].

To extract the context of an entity, many studies directly used the Wikipedia article describing the entity [47, 68, 109, 121, 190, 280, 289, 303], similarly with the *Article-based* context creation method; some works extended it with all the other Wikipedia articles linked to the Wikipedia article describing the entity [93, 108, 153]; while some only considered the first paragraph of the Wikipedia article describing the entity [68]. **Different from these approaches**, the *Graph-based* approach not only employs *in-links* and *language-links* to broaden the article set that is likely to mention the entity, but also performs a more fine-grained process: extracting the sentences that mention the entity, such that all the sentences in the context are closely related to the target entity. Thus, the entity-centric contexts extracted via the *Graph-based* approach are more comprehensive and accurate than former approaches.

As to the topic vector representation of the entity, [47, 140] defined it as the binary presence, term frequency (*tf*), or term frequency—inverse term frequency (*tfidf*) values of all the vocabulary words in the context; [68, 221] defined it as the binary presence or *tf* values of other entities in the context; [92, 93, 121, 190, 280] defined it as the similarity values between the target entity's *tfidf* context representation and other entities' *tfidf* context representation; [303] defined it as the visiting probability from the target entity to other entities from Wikipedia; [108,289] used a measurement based on the common entities linked to the target entity and other entities from Wikipedia. **Different from the above works**, I employ *topical aspect weights* that have a different interpretation of the frequency and selectivity than the typical *tfidf* values and take co-occurrence and language specificity of the topical aspects into account. Some studies [47, 68, 121, 153, 190, 280] also employed *category-links* to the Wikipedia article describing the entity. Since the category structure of Wikipedia is language-specific, it is hard to gain insights about the cross-lingual similarity for this case.

With the development of multilingual Wikipedia, there have been many researchers focusing on the differences in the usage and content between the different Wikipedia language editions. In [113], researchers demonstrated the diversities in the concepts discussed in multilingual Wikipedia, as well as the sub-concepts mentioned in the multilingual Wikipedia articles discussing the same concept. They further illustrated that the diversities above had a significant influence on the results of Explicit Semantic Analysis. In [181], researchers explored the question: "Do different language communities develop very diverse versions of equivalent articles in multilingual Wikipedia?". They developed the Manypedia web tool, to present the differences in various features of multilingual Wikipedia articles describing the same concept, which included the images, frequent words, total edits received, the number of different editors, creation date, creator, etc. In [30], researchers developed Omnipedia to visualise the sub-concepts mentioned in multilingual Wikipedia articles discussing the same concept. In [99], researchers proposed a new similarity measure, which combined the textual similarity and the metadata similarity of two multilingual Wikipedia articles discussing the same concept. They also presented the MultiWiki interface to demonstrate the temporal evolvement of the proposed similarity measure. In [16], researchers analysed the differences in patterns of content transclusion in multilingual Wikipedia. Some researchers tried to understand the underlying cultural reasons that drove the differences. For example, researchers in [158] extracted cultural relations, by analysing the content and page views of multilingual Wikipedia articles on cuisines. Besides, the differences in editing behaviour of multilingual Wikipedia were analysed in [142, 201, 218, 302]. **Different from these works**, I propose an automatic approach to systematically analyse the similarities and differences in the entity's related topical aspects extracted from language-specific Wikipedia corpora.

As for incorporating the Wikipedia knowledge in information retrieval applications, [83, 199, 208] applied concept-based approaches that mapped both the documents and queries to the Wikipedia concept space; [190, 242] focused only on query extension; [223, 252] focused only on mapping documents to Wikipedia concept space. To retrieve documents that did not explicitly mention the query entity by name, but were still relevant to the query entity, I choose to map both the query and the documents to the topical aspects space. As for the evaluation metrics of these information retrieval models, all these works used the presence of the query entity as a prerequisite of one document to be relevant. **My research, on the other hand, excludes this restrictive condition**. A document would be annotated as "Relevant" as long as it can satisfy any one of the three much more lenient criteria in Section 4.5.1. When facing a dataset without enough documents mentioning the query entity explicitly, this context-based information retrieval model would still be able to return the most relevant documents, thus achieving higher recall than former works under this setting. Researchers have also been employing multilingual Wikipedia in many multilingual information retrieval applications [223, 242, 252]. However, none of these studies paid attention to the language specificity of multilingual Wikipedia. As different language editions of Wikipedia discussed different topical aspects related to the entity, **in this work, I take a step further to** realise language-specific information retrieval through the entity's language-specific topic representations.

4.7 Conclusion

In this chapter, I have proposed context creation approaches for named entities, and derived language-specific topic representations for entities, to support entity-centric information retrieval. I have compared different ways of context creation, including the *Article-based* and the *Graph-based* approaches. A Wikipedia article describing the entity in a certain language can be seen as the most straightforward source for the language-specific entity context. Nevertheless, such context can be incomplete, lacking important topical aspects. Therefore, in this chapter, I have proposed an alternative approach to create the context, i.e., the *Graph-based* approach. The evaluation results have shown significant differences between the topic representations derived from the contexts that are obtained using different creation approaches. I have suggested that the topic representation derived from the *Graph-based* context provides a comprehensive, language-specific overview of the entity, independent of the coverage of the Wikipedia article describing the entity. To answer RQ1, language-specific topic representations can be constructed for entities from multilingual Wikipedia, and the proposed *Graph-based* approach can improve the comprehensiveness and accuracy of such topic representations.

Furthermore, I have proposed a context-based information retrieval model that applies such language-specific topic representations for entities to improve the recall of entity-centric information retrieval applications, while keeping high precision. The case study presented has illustrated that this model can retrieve documents that contain entity-related information, such as relevant events in the current news articles, even if the entity is not mentioned explicitly. Moreover, by selecting topic representations derived from contexts for different languages, my context-based information retrieval model makes language-specific results possible. This experiment has further proven the comprehensiveness, accuracy and language specificity of the topic representations for entities, which were derived from the Graph-based contexts.

Even though in this chapter I have used a limited number of named entities and languages as examples, neither the proposed approach or the model is dependent on specific languages and entities, thus can be easily extended to all other languages and named entities.

The semantic differences of multilingual Wikipedia, when discussing certain entities, are reflected not only in what related topical aspects are discussed, but also in the contributors' sentiment expressed. In this chapter, I have analysed the semantic differences from the topical aspects perspective to explore the differences between an entity's topic representations for different language editions of Wikipedia. In Chapter 6, I will continue the analysis from the aggregated sentiment perspective, to investigate the differences between the sentiments expressed by Wikipedia contributors in various Wikipedia language editions toward the entity. First, however, Chapter 5 continues the work on Topic Analysis, moving the target from entities to events, as will be further explained.

Chapter 5

Timeline Generation for High-impact Events from the Tweet Stream

In Chapter 4, I have explained how I extracted the topical aspects for *entities*, based on Wikipedia content. In this chapter, I therefore propose algorithms to detect and summarise the fine-grained topics (sub-events) referring to high-impact *events* from the tweet stream, in order to generate chronological event summaries for Internet users. This chapter answers *RQ2. Can timelines of high-impact events be generated automatically from the tweet stream?* This chapter and Chapter 4 represent the part of the study on topic analysis of social media text. Other studies also using Twitter content can be found in Chapter 8. The work in this chapter has been published in [323].

5.1 Introduction

Social media sites, such as Twitter, have become a popular platform for communication in everyday life and in the time of crisis. In the case of critical situations, Twitter demonstrates its usefulness, when users urgently need information, especially if they are directly affected by *major events*, for example, disease outbreaks or natural disasters.

Due to the prevalence of events reporting and collective attention in Twitter, numerous works have leveraged tweets for detecting real-world events. These works can be categorised based on the duration and influence scale of the events they were focusing on. For instance, the events "Gulf of Mexico oil spill", "Ebola outbreak" and "Zika virus outbreak" are regarded as *major events*, because they have long duration and high impact on people worldwide, examples include [66,166,272]; whereas the events "Charlton Road Closure for London Marathon" and "Three people were released from a lift at Pescod Square" refer to *local events*, with short duration and an impact limited to specific group of people, examples include [244, 283, 311].

The consumption of event-related stories in Twitter can be a tedious task that requires cognitive effort, due to the overwhelming amount of tweets, as well as the presence of noisy, redundant and duplicate information. Moreover, a large proportion of tweets contains mundane discussions, irrelevant to real-world event detection. In the case of tweets reporting about an event of interest, they might contain plenty of near-duplicates, in which the main content conveys the same meaning, with slightly different word usages [66, 166, 286].

In this chapter, I focus on a novel problem: detecting *fine-grained topics* of a known *major event*, to automatically generate a real-time timeline for the *major event*, in a format as in `https://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa_timeline`. I choose Wikipedia timeline because it is collaboratively constructed by Internet users, representing their most favourable timeline format that can help them to understand the temporal evolvement of the major event.

The generation of timeline summaries is considered to ease the comprehension of major events from a news stream [271] or social media, such as Twitter. The generated timeline summaries consist of *fine-grained topics*, or *sub-events*, representing key incidents, relevant to a given major event. The sub-events discussed in this chapter show the status of the ongoing major event. The advantage is that they are more fine-grained than the topics/events detected by traditional approaches, such as "Japan tsunami" and "Ebola outbreak". Nevertheless, they can earn compatible attention on a similar scale with their associated major event, but this attention can only last for a few days, or even shorter, as they will be superseded by the following sub-events. For example, "On March 24, two suspected cases in Liberia are announced by the Liberian Ministries of Information, Culture, Tourism, and Health. The government had also stated that Ebola had 'crossed over into Liberia,' but did not confirm the information." is a sub-event of the major event "Ebola outbreak". By using a chronological order, a timeline can represent the temporal development of the major event. Thus, the main task is to detect sub-events and to provide concise and non-redundant summaries. Furthermore, a timeline must be generated in real-time, in order to help users follow recent updates about the high-impact events according to their interest.

Few works have been done in the area of sub-event detection and timeline generation, which include [141, 171, 219]. The approach proposed in this chapter is different from former ones in the following ways: I differentiate between real-world events reporting tweets and other tweets, by applying only *event-independent features*; I employ an online incremental clustering algorithm to handle different levels of duplicated tweets reporting on the same sub-event, which makes real-time timeline generation possible; considering the evolving characteristic of major events, I propose a post-processing step to improve clustering performance and reduce computational cost. As I only use event-independent features overall, the approach can be easily adapted to other major events. I perform a thorough evaluation of the proposed approach on the "Ebola outbreak" tweet stream and verify its advantages based on several evaluation metrics, over the baseline approaches.

The proposed approach can be an efficient supplement or even replacement of the user-generated timeline. Its real-time characteristic can not only eliminate the lag between user-generated timeline on Wikipedia and news reports [89], but also can help to generate early alarms for disasters.

5.2 Timeline Generation from the Tweet Stream

In this section, I present a timeline summarisation approach for real-world major events from the tweet stream, as show in Figure 5.1. As a preprocessing step, POS-tagging is performed on the tweets in the English tweet stream, which mention the pre-known target entity(ies) related to the studied major event. The POS-tags can provide features for the subsequent stages. This is achieved by the CMU Part-of-Speech Tagger [209] for tweets. According to [209], it can achieve more than 90% accuracy on various tweets datasets. After preprocessing, I filter out tweets in the stream that are not *real-world events reporting tweets*. Moreover, I apply an online incremental clustering algorithm to cluster the near-duplicate tweets reporting on the same sub-event, in real-time. Furthermore, I adopt a post-processing step to generate more precise results and remove clusters reporting terminated sub-events. I update the summaries of sub-event clusters, as long as there are new tweets to be included and order these summaries chronologically, which constitute the timeline of the major event.

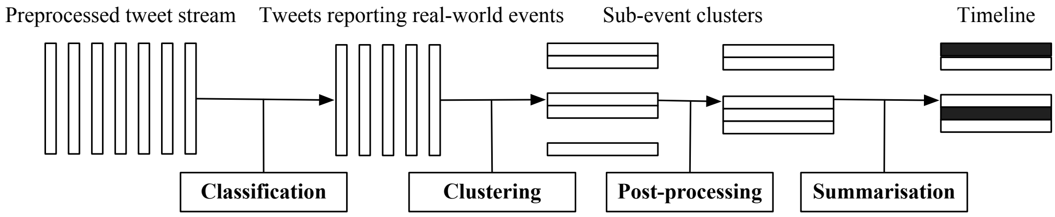

Figure 5.1: Framework for timeline generation from the tweet stream.

5.2.1 Extraction of Tweets Reporting Real-world Events

In [66, 166, 286], researchers have pointed out that about 50% of the tweets on Twitter are not relevant to real-world events. For this reason, in my approach, I first filter all the tweets in the stream by the major event's relevant entity(ies) to reduce the number of irrelevant tweets. I further differentiate the tweets that report real-world events from the tweets that express personal feelings, or pointless "babbles", to avoid the "mundane" and "polluted" information [22].

I train a binary classifier, to determine if one incoming tweet is a real-world event reporting tweet or not. I explore the differences in expression patterns between real-world events reporting tweets and other tweets. I propose the feature set based on the observed differences in expression patterns that are event-independent. The effectiveness of each feature in the feature set is examined by comparing the classifier's performance before and after adding this feature through cross validation. Event-dependent features, such as the n-grams, are excluded. There are obvious Twitter syntax usage differences in tweets reporting real-world events and personal feelings. Thus, one set of features is the Twitter syntax feature set, which is commonly used in tweet-related classification, which include: the number of hashtags in the tweet, the number of at-mentions in the tweet. Another set of features are indicators of other users' reactions to this tweet, which include: the number of retweets of a tweet, whether the tweet has been "favourited", as Twitter users are likely to have different reaction patterns when reading about tweets reporting real-world events from other tweets expressing personal feelings. Compared to real-world events reporting tweets, Twitter users are more likely to include informal language, such as emoticons and abbreviations, in tweets expressing personal feelings. Moreover, Twitter users like to use interjections, exclamation marks, question marks in personal feeling expressing tweets to stress the tone used. In contrast, fact-related information, such as numbers, URLs and locations are frequently mentioned in real-world events reporting tweets. I further include all these above features into the feature set. The number of emoticons, abbreviations, interjections, numbers and URLs in the tweet can be obtained through the CMU POS tagger [209]. The number of exclamation marks and question marks can be obtained by simple character matching. I additionally calculate the number of locations mentioned in the tweet, by checking the inclusion of location names in the noun phrases obtained after POS-tagging. The location names are extracted through gazetteer lookup, the scope and granularity of which can be configured based on the characteristics of the major event to improve efficiency. For example, the country level location names can be extracted from *iso3166*[1].

I do not include the user profile features and the occurrence of a tweet's geo-tag information into the feature set, as experimental results showed that those features cannot help to improve the classifier's performance in the experimental results. This may be due to the fact that the major events usually attract the attention of all kinds of Twitter users, from public accounts of news agencies to regular personal accounts, no matter where their physical locations are. Besides that, the Twitter's retweet function and the "Tweet Button" on webpages make it much easier for Twitter users with different backgrounds to report real-world sub-events related to the major event.

5.2.2 Sub-event Detection in the Tweet Stream

Due to the huge volume of daily posts on Twitter, a large percentage of them can be seen as redundant, as they only report on the sub-events that are already reported by other tweets. In [122], researchers observed that when people were discussing a product, the vocabulary that they used converged for important factual aspects; on the other hand, the vocabulary for personal reviews was often diverse. In [262], researchers distinguished near-duplicate tweets on 5 levels, which were: *exact copy*, *near exact copy*, *strong near-duplicate*, *weak near-duplicate* and *low-overlapping*: for *exact copy*, *near exact copy*, *strong near-duplicate* and *weak near-duplicate* tweets, the main parts are identical or almost the same; for *low-overlapping* tweets, they

[1] `https://pypi.python.org/pypi/iso3166/0.6`

only have a couple of common key terms, but greatly vary in word usages and expression patterns. My sampled tweet dataset illustrates that tweets reporting on the same sub-event are near-duplicate tweets, which is also consistent with the findings in [122].

In [262], researchers treated the near-duplicate detection as a classification problem and the classifier had to make a decision on every pair of tweets that were possible to be near-duplicates. Their near-duplicate detection strategy worked well on a small scale, but it needed human annotation of tweets from various domains to train the classifier and its computation complexity was thus really high. On the other hand, traditional online incremental clustering algorithms, based on the similarities of *tfidf* vector representations (discussed in Section 3.2.1) of tweets' textual content [12, 46, 298], are often employed to process tweet streams online. The traditional incremental online clustering algorithms have low complexity and do not need prior knowledge of the number of clusters. However, they have the following drawbacks: (i) The inverse document frequency (*idf*) information, either iteratively updated or extracted from auxiliary corpus, can be biased, depending on the differences between the term distributions of the processed tweets/auxiliary corpus and the real term distribution; (ii) Tweets are short texts, and therefore the role of some rare and novel terms can be dominating, when calculating the similarities between tweets using their *tfidf* vector representations; (iii) Researchers usually reduce the dimensionality of the vector representations by selecting the terms with high *idf* values only, but it is questionable to equal rareness with importance, especially when the *idf* information is not reliable, as some valuable information can be easily lost; (iv) By setting a reasonable threshold, this kind of online clustering algorithms may have acceptable performance on *exact copy*, *near exact copy*, *strong near-duplicate* and *weak near-duplicate* tweets, but they can hardly deal with *low-overlapping* tweets, which account for 18.8% of all kinds of near-duplicate tweets, according to [262].

Another drawback shared by most current online incremental clustering algorithms for event detection in tweets is that they do not consider the textual variants of terms, which are highly frequent, because of tweets' short and informal characteristics and their rich syntax features. For example, tweet t_1: "#Senegal sends medical teams to border with #Guinea after an outbreak of Ebola there. #Sierra Leone, much closer to the epicenter, hasn't." and tweet t_2: "Senegal has sent a medical team to all its main border crossing points with Guinea after an outbreak of Ebola... `http://fb.me/6WqOZ2b3h`" are talking about the same sub-event. They contain some key terms that vary in representation forms, but have the same meaning, such as *#Senegal* and *Senegal*, *medical teams* and *medical team*, *Guinea* and *#Guinea*, which should be treated as the same terms. Other examples include *#Liberians* and *liberia*, *#EbolaFree* and *ebola-free*, etc. However, traditional online clustering algorithms ignore this phenomenon, they treat these key terms, which share the same meaning, but only vary slightly in representation forms, as different terms. As a result, the similarity of these two tweets decreases and they cannot be included in the same cluster when the *clustering threshold* is high. However, alternatively blindly lowering the *clustering threshold* can cause the decrease of the *clustering precision* (defined in Section 5.3.3).

Algorithm 1: Sub-event detection in tweet stream

```
input  : $T$, tweet stream; $E$, target entity(ies)
output: $ProcessingClusters$, clusters of tweets reporting the same sub-event
$ProcessingClusters = \emptyset$;
foreach Tweet $t \in T$ mentioning $e \in E$ do
    Preprocess $t$;
    if $t$ is reporting a real-world sub-event then
        Initialise a cluster $c_t$ with $t$ using $U_t$ (useful URLs in $t$) and $K_t$ (key terms in $t$);
        foreach cluster $c$ in $ProcessingClusters$ do
            if cluster $c$ has common useful URL with $c_t$ then
                MergeClusters($c$, $c_t$);
            else if GetSimilarity($c$, $c_t$) > clustering threshold then
                MergeClusters($c$, $c_t$);
            else
                add $c_t$ to $ProcessingClusters$;
            end
        end
    end
end
```

To solve the above problem and reduce the dimensionality of tweets' vector representations, as well as increase the *clustering precision* and decrease the *compression ratio* (defined in Section 5.3.3), I propose a variant online incremental clustering algorithm, as shown in Algorithm 1. Algorithm 1 incrementally clusters the tweets based on common URL(s) and key terms sharing the same meaning.

I try to reduce its computational cost as it needs to process the incoming tweets in real-time. When a new tweet comes, after preliminary target entity(ies) filtering, preprocessing and classification, I obtain the tweets reporting real-world sub-events belonging to the pre-known major event. To eliminate the noise, I only use some *key terms* and URLs in the tweet to construct its vector representation, as these parts are crucial to deciding what sub-event the tweet is discussing. The *key terms* are noun phrases, verbs, hashtags, numbers and at-mentions. The choice is made both empirically, based on the observation that for tweets reporting the same sub-event, these key terms would be the same or textually similar, but the other parts of the tweets, such as conjunctions, adjectives and adverbs, often vary; and experimentally, based on the performance of various choices. In this way, the dimensionality of the tweets' vector representations is reduced. Since I already have the POS-tags after the preprocessing step, I only need chunking to extract the key terms, and lemmatisation to transform the verbs from their various inflected forms to their original forms.

It is of high probability that tweets containing URLs are closely related to the content of the linked webpages [3]. Some studies have used this kind of tweets as the summaries or highlights of the sub-events reported by the linked webpages [285]. This has shown that the benefits of the assumption that tweets containing URLs represent highlights of the sub-events reported by the linked webpages, overweigh the risks. Based on the above assumption, new tweets are incorporated into the processing sub-event cluster with which it shares common URL(s). Two sub-event clusters are considered to report on the same sub-event if they contain common URL(s). Similar approach was also employed in [51]. I do not take the full actual content of the webpages into account, to avoid the inclusion of noisy information. Because of the characters limitation of Twitter, the URLs contained in the tweets are mostly shortened in various ways, to save space. After retrieving the original URLs, I consider the URLs that contain nothing but domain and category information, such as `http://NBCNews.com` and `http://www.nbcnews.com/news` to be *useless*, as this kind of URLs provide no information about the sub-events. One original URL would only be consider as *useful*, if it contains the concrete address of a real-world sub-event reporting webpage.

The prioritised *URL-based clustering strategy* can help to enrich the processing sub-event cluster with the key terms that report the sub-event from different angles. On the other hand, if an incoming tweet does not contain any common URL with any processing sub-event cluster, it is still possible to be incorporated into one processing sub-event cluster. This is achieved by the *threshold-based clustering strategy.* As mentioned before, the problem that key terms appear in slightly different forms, but have the same meaning, widely occurs in tweets. For example, the occurrences of "#Liberia", "Liberian" and "Liberia's" have the same effect as the occurrence of the country name "Liberia". To deal with this problem, I treat two different key terms as the same key term, as long as the Jaro-Winkler metric between them is above a threshold. This method can further reduce the dimensionality of the tweets' vector representations. Similar to [250], each dimension in the tweets' condensed vector representations denotes a cluster of key terms that share the same meaning, rather than one individual key term. The Jaro-Winkler metric is specially designed for short strings matching, which is based on the length of the longest common prefix, the number and order of the common characters between two strings [62]. In [62], researchers replaced the exact term matching with approximate term matching, based on the Jaro-Winkler metric; in [61], researchers compared various personal name matching techniques, and the Jaro-Winkler metric was one of the techniques with the best performance. I set the Jaro-Winkler metric threshold to 0.9, following [62]. Based on the above description, I define $GetSimilarity(c_1, c_2)$ in Algorithm 1 as follows:

$$GetSimilarity(c_1, c_2) = J_{JW}(K_1, K_2) = \frac{K_1^{'} \cap K_2^{'}}{K_1^{'} \cup K_2^{'}} \tag{5.1}$$

where: c_1 and c_2 denote two clusters, K_1 and K_2 denote the key term sets in c_1 and c_2. I replace the traditional Jaccard similarity metric based on exact matching (J) with a Jaccard similarity metric based on the Jaro-Winkler matching (J_{JW}). In Equation 5.1, $K^{'}$ represents key term clusters, with all the key terms in the same cluster sharing the same meaning. A new key term can be incorporated into one of the key term clusters, as long as the Jaro-Winkler metric between this new key term and any one of the key terms that are already in the cluster is above 0.9. For two key term clusters, k_i and k_j, they are viewed as the same key term cluster if the Jaro-Winkler metric between one key term from k_i and one key term from k_j is above 0.9.

All information about the processing sub-event cluster, such as the above mentioned key terms and URL(s), will be updated as long as it incorporates new tweets; the information about individual tweets in the sub-event clusters will be discarded to save space.

In [304], researchers pointed out that clustering algorithms utilising the Jaccard similarity metric achieved better performance than the ones utilising the cosine similarity metric, because of the sparsity of tweets. Similar to [315], I choose the Jaccard similarity metric when evaluating the similarity between two tweets' condensed vector representations. Because of the usage of Jaccard similarity, I only need to store the tweet's vector representation as a binary vector, which means the tweet's condensed vector representation is a variant of the binary-based bag-of-words (BOW) vector representation discussed in Section 3.2.1. Since only key terms are considered, I do not have to face the problem that the Jaccard similarity metric cannot deal with terms with different levels of importance.

5.2.3 Post-processing of Detected Sub-events

Algorithm 2: Post-processing of detected sub-events

```
input  : $ProcessingClusters$; $M$, termination threshold; $D$, target period
output: $ProcessingClusters$, $TerminatedClusters$, clusters of tweets
          reporting the same sub-event
$TerminatedClusters = \emptyset$;
foreach Day $d \in D$ do
    $ProcessingClusters$ = HierarchicalClustering($ProcessingClusters$);
    foreach cluster $c$ in $ProcessingClusters$ do
        if $c$ has not incorporated sub-event updates for $M$ days then
            move $c$ from $ProcessingClusters$ to $TerminatedClusters$;
        end
    end
end
```

The proposed online incremental clustering algorithm (Algorithm 1) inevitably has some drawbacks. First, the fact that an incoming tweet is incorporated into one processing sub-event cluster, as long as certain conditions are met, ignores the possibility that there are other processing sub-event clusters, which may also meet these conditions. Second, the information in clusters is dynamic; so one cluster can become more similar to some other clusters after incorporating some tweets. To solve the above problems, I apply a more rigid and computationally-intensive hierarchical clustering algorithm on processing sub-event clusters. The reasons I choose the hierarchical clustering algorithm over other clustering algorithms of similar cost, such as k-means clustering and spectral clustering [133], are that the number of clusters is unknown and I need fine-grained clusters consisting of near-duplicate tweets reporting on the same sub-event.

A hierarchical clustering algorithm needs the distance matrix of all the items to be clustered as the input, and successively merge the sub-event clusters based on their distances. While hierarchical clustering is more robust than online incremental clustering, because of its tendency to compare all pairs of items [298], it is very inefficient, as its computational overhead grows as the square of the number of items to be clustered. Thus, hierarchical clustering is not suitable for online scenarios. However, after the online incremental clustering, the number of items to be clustered (processing sub-event clusters) is much lower than the original number of tweets, which greatly reduces the computational overhead. Moreover, since the hierarchical clustering algorithm aims at fixing the miss-outs and improving the clustering quality of Algorithm 1, it has lower priority thus can be processed offline, at the end of each day, or during any less busy time. I use a similar strategy as in Algorithm 1 to compute the distance matrix of the processing sub-event clusters, as the input of the hierarchical clustering algorithm: for two processing sub-event clusters, their distance is 0 if they mention common *useful* URL(s); otherwise their distance is the Jaccard distance based on the Jaro-Winkler matching of their condensed vector representations. I use the same *clustering threshold* in Algorithm 1 as the *cutting threshold* in the hierarchical clustering algorithm, to guarantee that a similar standard is applied. I choose single-linkage hierarchical clustering, aiming at merging clusters that contain the closest pair of sub-event clusters into a new cluster. In this way, I can deal with the following scenario: if there exist sub-event clusters reporting on the same sub-event from different angles in one intermediate cluster, another intermediate cluster can be further merged with this intermediate cluster, as long as it contains sub-event clusters reporting on the sub-event from any angle. I consider all the tweets in the new generated clusters reporting on the same sub-event.

Since the duration of sub-events is short, I also consider temporal features of processing sub-event clusters to further reduce the computational cost of Algorithm 1 and the hierarchical clustering algorithm. Inspired by the idea of inactive clusters in [8], I set up the following rule: a sub-event can be seen as terminated as long as there is no new tweet reporting on this sub-event for M days since the sub-event cluster's last incorporation. If one sub-event has terminated, its identity will be changed from *processing sub-event cluster* to *terminated sub-event cluster.* I discard the possibility that an incoming tweet reports on a terminated sub-event, thus it is not possible for terminated sub-event clusters to incorporate new tweets. I also discard the possibility that a processing sub-event cluster reports on the same sub-event with any terminated sub-event cluster, thus terminated sub-event clusters are not considered by the hierarchical clustering algorithm either. This rule can improve the efficiency of the whole approach, but it inevitably compromises the overall performance. M can be customised according to the user's interest in obsolete sub-event reporting tweets. I recommend setting M not to a number less than 15, considering the duration of sub-events. The algorithm for the post-processing step is shown in Algorithm 2.

5.2.4 Timeline Summarisation

In this step, I extract the description as well as the timestamp for each sub-event from its corresponding cluster. I perform extractive summarisation of the sub-events described by the clusters and rank them in chronological order to generate the real-time timeline. This summarisation problem is different from traditional summarisation problems from the following perspectives: first, since all the tweets in the same cluster are near-duplicate tweets reporting on the same sub-event, the most representative tweet of each sub-event cluster is selected as its summary, as in [171, 177]; second, it is not feasible to construct separate annotated datasets for different major events, thus the summarisation has to be performed in an unsupervised manner; third, along with the processing of the tweet stream, the sub-event clusters will be updated in real-time, thus it is necessary that the summaries for the sub-event clusters can also be updated in real-time. Considering the above demands, I proposed a heuristic algorithm to generate sub-event summarisations for the timeline, with both temporal and textual features included, as shown in Algorithm 3.

I select the tweets covering the highest number of key term clusters, as the *candidate representative tweets*. This is for the reason that the representative tweet should contain as much information as possible. From the candidate representative tweets, I select the one that has the most recent posted time (P_t) as the summary of this sub-event. This is due to the fact that the summary should contain the newest update of the sub-event.

As for the timestamp of the sub-event, I combine the extracted timestamps from temporal expressions in tweets by the dateparser[2] with the tweets' posted time, similar to [244]. This is because users are likely to post tweets reporting past sub-events.

[2] https://dateparser.readthedocs.org/

Algorithm 3: Generation of items for the timeline

```
input  : $c$, a cluster of tweets reporting the same sub-event
         ($ProcessingClusters + TerminatedClusters$)
output : $s_c$, the summary of this sub-event; $T_c$, the timestamp of this
         sub-event; $P_c$, the posted time of the summary tweet of this
         sub-event
$MaxSimilarity = 0$, $T_c = CurrentTime$, $P_c = CurrentTime$;
foreach new incorporated tweet $t$ in cluster $c$ do
    if extracted timestamp from $t$'s text $< P_t$($t$'s posted time) then
        $T_t$ (the timestamp of the sub-event reported by $t$) $\leftarrow$ extracted
        timestamp from $t$'s text;
    else
        $T_t \leftarrow P_t$;
    end
    Initialise a cluster $c_t$ with $t$'s key terms $K_t$;
    if GetSimilarity($c_t$, $c$) $> MaxSimilarity$ then
        $MaxSimilarity \leftarrow$ GetSimilarity($c_t$, $c$);
        $s_c \leftarrow t$, $P_c \leftarrow P_t$;
    end
    if GetSimilarity($c_t$, $c$) $= MaxSimilarity$ and $P_c < P_t$ then
        $s_c \leftarrow t$, $P_c \leftarrow P_t$;
    end
    if $T_c > T_t$ then
        $T_c \leftarrow T_t$;
    end
end
```

For example, the tweet “Good news! No confirmed cases of Ebola recorded by the Sierra Leone government in their 20 March daily report. `http://reliefweb.int/report/sierra-leone/ebola-outbreak-updates-march-20-2015` ...” is posted on 21st March 2015, one day after the occurrence of the sub-event. The timestamp of the sub-event (T_c) is set to be the earliest posted time of all the tweets in its corresponding cluster, only if no earlier timestamp can be extracted from the tweets; otherwise I use the earliest extracted timestamp instead.

5.3 Experimental Results

5.3.1 Dataset Description

I applied the proposed real-time timeline summarisation approach on the existing Ebola Tweets dataset provided by the TREC Dynamic Domain Track[3]. This dataset contains 165,000 tweet-ids, while only 90,823 of them, which were posted during a period from 31 Jan 2014 to 23 Mar 2015, can be accessed. It should be noted that only a small percentage of tweets in this dataset are related to the "Ebola outbreak". I processed the downloaded tweets in the order of their posted timestamps, to simulate the tweet stream. The known major event for the evaluation was "Ebola outbreak". I filtered out all the non-English tweets, and used "Ebola" as the target entity, in order to filter out tweets that were not related to the considered major event.

5.3.2 Evaluation of Extraction of Real-world Events Reporting Tweets

I utilised CrowdFlower, a crowdsourcing website, to annotate 3,000 tweets, which were randomly sampled from the dataset into two categories: *real-world events reporting tweets* and *other*. Only 2,103 *real-world events reporting tweets* and 333 *other* tweets were left after filtering out all the annotated tweets with confidence lower than 0.9. Since there was a big difference between the numbers of items from these two categories, I balanced the dataset through undersampling [107] to avoid bias. I used the grid search to find the most suitable parameters for a Support Vector Machine (SVM) classifier (discussed in Section 3.3.2) based on the average F1 score. An SVM classifier using the RBF kernel, with the kernel parameter γ set to 0.3125 and penalty parameter C set to 8 achieved the best performance; the detailed definitions of these two parameters can be found in Section 3.3.2. The precision, recall and F1 score of the classifier generated through 10-fold cross validation is shown in Table 5.1.

A recall of 0.850 was achieved on the *real-world events reporting tweets* category using this classifier, which meant about 85.0% of the *real-world events reporting tweets* related to the major event can remain after this stage.

Even though I used the "Ebola outbreak" dataset to train the classifier, only

[3] `http://trec-dd.org/`

Table 5.1: Performance of the extraction of real-world events reporting tweets.

Metric	Other	Real-world	Macro-Avg.
Precision	0.828	0.761	0.795
Recall	0.730	0.850	0.790
F1 score	0.779	0.805	0.792

event-independent features were employed, as illustrated in Section 5.2.1. Thus, the classifier's performance will be less affected than other classifiers that are employing event-dependent features, when categorising tweets related to other major events.

I extracted 7,069 *real-world events reporting tweets* in English relevant to the major event "Ebola outbreak" after processing the whole tweet stream, without performing any clustering.

5.3.3 Evaluation of Sub-event Detection

The cosine similarity between the tweets' *tfidf* vector representations is the most widely employed similarity metric for recent works on online clustering [34, 104, 219, 325], where both the centroids of clusters and IDF weights of the terms were iteratively updated. I implemented the threshold-based online clustering algorithm utilising cosine similarity metric between tweets' *tfidf* vector representations (denoted by **Cosine-tdidf**) as a baseline. Another baseline I implemented was a similar algorithm as Cosine-tfidf but using the Jaccard similarity metric instead (denoted by **Jaccard**). I also compared the performance of the proposed algorithm with and without the post-processing step, denoted by **P** and **No-P**, respectively.

Unlike [183, 244], I define a stricter way to measure the *clustering precision* to reflect the effectiveness of the clustering algorithm at the more intuitive cluster level, rather than the individual tweet level. The *clustering precision* is defined as the percentage of *correct clusters* in all the generated clusters that *contain more than one tweet* after processing the whole tweet stream. A cluster can only be counted as a *correct cluster* if *all* the tweets in the cluster describe the same *sub-event*. The clustering precision evaluation task was divided evenly on three judges. For each sub-event cluster, I provided the judges with all the tweets in the cluster and asked them to annotate it as a *correct cluster* or a *incorrect cluster*. To avoid bias, the judges were kept unaware of any configuration information for each unannotated cluster.

Since I lacked the ground truth about all the sub-events during the "Ebola outbreak", it was infeasible to calculate the recall. In [294], researchers defined *reduction ratio*, as the ratio of the size of the original dataset to the size of the reduced dataset. Similarly, in [200], researchers defined *compression ratio* as the ratio of the size of the summarised text documents to the size of the original text documents. Both of the above evaluation metrics were used to evaluate the compression ability of clustering algorithms. Similarly, I define the *compression ratio* for the application as:

$$CR = \frac{C}{N}. \tag{5.2}$$

where: CR is the *compression ratio*; C is the number of the generated clusters, regardless of the number of tweets in the cluster; N is the total number of the tweets in all clusters. After clustering, all the tweets in the same cluster can be compressed into one summary, as they all described the same sub-event and were near-duplicate tweets. When two clustering algorithms reach the same *clustering precision*, the lower CR one algorithm achieves, the stronger the cluster algorithm's ability is in clustering near-duplicate tweets describing the same sub-event.

I experimentally set the parameter M in Algorithm 2 to 30, based on the observation that for "Ebola outbreak" related tweets, there was hardly any tweet discussing a sub-event, if this sub-event had not been updated for 30 days.

Table 5.2 shows the performance comparison of the proposed algorithm and the baselines, after processing the whole tweet stream, where CT denotes *clustering threshold* used in Algorithm 1 and Algorithm 2, CP denotes *clustering precision*, CR denotes *compression ratio*. I tuned CT in the range of [0.3, 0.9], with 0.1 increments.

Table 5.2: Performance of different sub-event detection algorithms.

CT	**Cosine-tfidf**		**Jaccard**		**No-P**		**P**	
	CP	**CR**	**CP**	**CR**	**CP**	**CR**	**CP**	**CR**
0.3	84.0%	94.0%	67.0%	65.7%	80.0%	60.9%	77.9%	59.2%
0.4	94.4%	97.1%	80.0%	74.6%	85.9%	70.4%	83.9%	68.3%
0.5	97.8%	98.6%	85.0%	78.1%	90.8%	74.4%	90.0%	72.4%
0.6	100.0%	99.2%	88.8%	81.5%	93.0%	76.4%	92.0%	75.6%
0.7	100.0%	99.3%	92.9%	85.1%	94.9%	77.9%	93.5%	77.6%
0.8	100.0%	99.4%	93.9%	90.1%	95.4%	79.9%	95.4%	79.8%
0.9	100.0%	99.5%	100.0%	99.2%	95.9%	80.8%	95.6%	80.8%

Table 5.2 demonstrates that the Cosine-tfidf algorithm has the highest clustering precisions for all the clustering thresholds. However, its high compression ratios show that the Cosine-tfidf algorithm is quite weak in detecting all kinds of near-

duplicate tweets describing the same sub-event. Because of the reasons mentioned in Section 5.2.2, online clustering algorithms based on cosine similarities of the tweets' *tfidf* vector representations are not suitable for clustering near-duplicate tweets.

For the other three algorithms, both of the proposed clustering algorithms with and without the post-processing step perform much better than the online clustering algorithm based on the Jaccard similarity, in both compression ratio and clustering precision, when the clustering threshold is below 0.9. On the one hand, the proposed clustering algorithm only considers the key terms, which can eliminate some noise introduced by tweets mentioning common adjectives and adverbs, but about different objects. On the other hand, the proposed clustering algorithms replace the exact term matching with fuzzy key term matching based on the Jaro-Winkler metric and apply the URL-based clustering strategy, both of which contribute to the large increase in compression ratio. When the clustering threshold is 0.9, the online clustering algorithm based on Jaccard similarity can only detect tweets that are *exact copies* or *near exact copies*, thus it achieves a higher clustering precision but much lower compression ratio than the proposed clustering algorithms.

The choice between the proposed clustering algorithms with and without the post-processing step should be made based on the real-life application, after some consideration on the balance between compression ratio and clustering precision. The proposed clustering algorithm with the post-processing step achieves a slightly better performance in compression ratio than the one without the post-processing step for any clustering threshold, at the price of slightly compromised clustering precision. When setting the clustering threshold to 0.5, the proposed clustering algorithm without the post-processing step's clustering precision is only 0.8% higher than the one with the post-processing step, but its compression ratio is 2.0% higher than the latter one. This is why I decided to choose the proposed clustering algorithm with the post-processing step over the one without the post-processing step.

I set the clustering threshold to 0.5 for the following evaluations, as that was when the proposed clustering algorithm had the lowest compression ratio, when the clustering precision was above 90.0%.

After setting the *clustering threshold* to 0.5, I use Figure 5.2 to further illustrate the proposed clustering algorithm's effectiveness in detecting near-duplicate tweets.

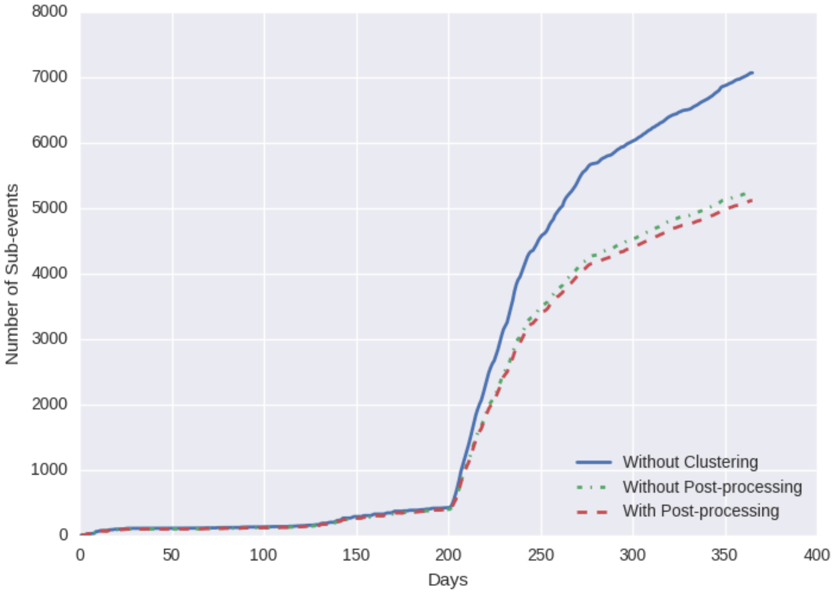

Figure 5.2: Number of detected *sub-events* (C) per day during the target period, with different settings.

5.3.4 Evaluation of Sub-event Summarisation

I further evaluated Algorithm 3's performance on the "Ebola outbreak" tweet stream. I selected one representative tweet for each sub-event cluster, considering both the amount of information and novelty, based on Algorithm 3. The aforementioned three judges were further asked to perform the summarisation task manually. I provided the judges all the generated clusters of tweets and let them choose one tweet for each cluster that can best represent the sub-event this cluster described. For 82.0% of all the clusters, the summarisation algorithm made coherent choices with human judges. Considering the demands for the summarisation algorithm in this real-time timeline generation scenario, mentioned in Section 5.2.4, I compared the performance of Algorithm 3 with a simple but intuitive *Most Recent* algorithm. The *Most Recent* algorithm took the most recently posted tweet as one cluster's summary [129]. The baseline *Most Recent* algorithm achieved 60.5% in accuracy, which is worse than Algorithm 3.

5.3.5 Comparison with the Manually Generated Timeline

A sampled timeline of the "Ebola outbreak" generated with the proposed approach is shown in Table 5.3.

After processing the whole "Ebola outbreak" tweet stream, the automatically generated timeline was compared with the manually generated Wikipedia timeline for the Ebola outbreak in West Africa. I employed the same timeline format as in `https://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa_timeline`, and used all the country names extracted in Section 5.2.1 from tweets in the same cluster as the locations of the sub-events. I did not use the tweets' geo-tags or user profile locations, because as said, unlike tweets reporting *local events*, most of the tweets reporting real-world *major events* were posted by Twitter users from all over the world, rather than from the neighbourhood of the *local events*.

There were 201 sub-events listed in `https://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa_timeline` from 2014 to 2015, and 126 of them can find their corresponding items from the automatically generated timeline. For example, "No serious med infrastructure in the area for response. 'Guinea confirms Ebola as source of epidemic `http://aje.me/1gU7kpU` via @AjEnglish"' and "#Liberia confirms first #Ebola case in weeks, just as the authorities were beginning the countdown to an Ebola-free nation", were included in both the user-generated Wikipedia timeline and the automatically generated Twitter sub-events timeline.

On the other hand, a large number of sub-events detected by the proposed approach, such as "And now a UK Military Health Worker battling #Ebola in #sierraleone `http://edition.cnn.com/2015/03/11/europe/uk-military-ebola/index.html`" and "Good news! No confirmed cases of Ebola recorded by the Sierra Leone government in their 20 March daily report. `http://reliefweb.int/report/sierra-leone/ebola-outbreak-updates-march-20-2015` ...", were only included in the automatically generated timeline. This proves that the automatic timeline summarisation approach for the tweet stream can also work as an efficient supplement of the user-generated timeline.

Moreover, since the proposed approach can detect real-time sub-events of the "Ebola outbreak", it can provide some early alarms for potential outbreaks in some countries. The World Health Organization (WHO), an organisation that always releases convincing worldwide epidemic reports and international travel alarms, usu-

Table 5.3: Example timeline generated for the major event “Ebola outbreak”.

Date	**Location**	**Timeline**
2014.03.23	Guinea	No serious med infrastructure in the area for response. “Guinea confirms Ebola as source of epidemic `http://aje.me/1gU7kpU` via @AjEnglish”
2014.03.24	Senegal, Liberia, Guinea, Sierra Leone	#Senegal & #Liberia mobilise medics to ward off #Ebola spreading in #Guinea. #SierraLeone much closer to epicenter doing/saying nothing.
2014.03.24	Sierra Leone	#EbolaFever has hit eastern Sierra Leone. Fast action needed please madam #MinisterofHealthand-Sanitation. This is very serious. God help us.
2014.03.26	Guinea	#Guinea says it has contained #Ebola outbreak in its southeast, but death toll rises and people are scared Reuters `http://in.reuters.com/article/2014/03/26/guinea-ebola-idINL5N0MN50D20140326` ...
2014.03.26	Guinea	@WHO does not recommend any travel, trade restrictions to #Guinea & neighbouring countries in respect to this #Ebola outbreak #AskEbola
2014.03.26	Guinea, Liberia	#Ebola virus kills 90% of those it strikes - 63 people have died so far in #Guinea in latest outbreak, 5 in #Liberia `http://www.bloomberg.com/news/2014-03-25/ebola-victims-face-90-death-risk-drugs-start-to-emerge.html` ...
2015.03.11	UK, Sierra Leone	And now a UK Military Health Worker battling #Ebola in #sierraleone `http://edition.cnn.com/2015/03/11/europe/uk-military-ebola/index.html`
2015.03.13	Liberia	WHO Confirms No Ebola Case in Liberia in Two Weeks - `http://AllAfrica.com` `http://goo.gl/fb/aDB55B` #LIBERIA
2015.03.20	Liberia	#Liberia confirms first #Ebola case in weeks, just as the authorities were beginning the countdown to an Ebola-free nation.
2015.03.20	Sierra Leone	Good news! No confirmed cases of Ebola recorded by the Sierra Leone government in their 20 March daily report. `http://reliefweb.int/report/sierra-leone/ebola-outbreak-updates-march-20-2015` ...

ally needs more time to gather enough facts than automatic approaches that extract knowledge directly from social media. The real-time timeline generated by the proposed approach, although with much less authority, still can provide some insights for international travellers and local people to avoid some dangerous areas, and also buy some time for them to get prepared for the potential coming outbreak. For example, WHO released its first report about this outbreak's situation in Guinea on 25th March 2014[4], in Liberia on 30th March 2014[5] and in West Africa on 1st April 2014[6]. It also released an international travel alarm for this outbreak on 28th March 2014[7]. However, starting with the 23rd March 2014, the proposed approach has already detected some sub-events describing new Ebola outbreaks in some West African countries, which could provide valuable information for some people who do not want to take any risk, as well as for governmental departments to start corresponding investigation.

5.4 Related Work

As said in Section 2.2, one line of research related to this work is Topic/Event Detection and Tracking on Twitter. According to [22, 286], event detection algorithms can be broadly classified into two categories: document-pivot methods, which detect events by clustering documents based on their semantic distances, and feature-pivot methods, which study the distributions of words and discover events by grouping words. There were a burst of works performing event detection on Twitter utilising *feature-pivot* methods recently. [166, 177] extracted all the topical terms for some given events first, then clustered the topical terms based on their co-occurrences or temporal frequency patterns. [180, 286] detected events by capturing the bursts in the terms' appearances. Some feature-pivot methods applied modified Latent Dirichlet Allocation models, as described in Section 3.2.2, on tweets by incorporating some tweet-specific characteristics. For example, [313] proposed a Twitter-LDA model, which assumed a single topic assignment for an entire tweet; [66] applied the LDA model only on hashtag signals that were identified as events indicators

[4] http://www.afro.who.int/en/clusters-a-programmes/dpc/epidemic-a-pandemic-alert-and-response/outbreak-news/4065-ebola-virus-disease-in-guinea-25-march-2014.html

[5] http://www.afro.who.int/en/clusters-a-programmes/dpc/epidemic-a-pandemic-alert-and-response/outbreak-news/4072-ebola-virus-disease-liberia.html

[6] http://www.afro.who.int/en/clusters-a-programmes/dpc/epidemic-a-pandemic-alert-and-response/outbreak-news/4073-ebola-virus-disease-west-africa-1-april-2014.html

[7] http://www.who.int/ith/updates/20140328/en/

through wavelet signal analysis; [76] proposed a TimeUserLDA model, based on the assumption that tweets reporting global events were likely to follow a global topic distribution that was time-dependent, and tweets reporting personal topics were likely to follow a personal topic distribution that was time-independent; [272] enriched the LDA model with the weights of event terms on timeline and the reliabilities of users to extract social events; [233] applied a LinkLDA model to group tweets from the same event category, based on the assumption that an event term's type distribution was shared across its mentions. This kind of feature-pivot methods can achieve good performance on detecting *major events*. However, they cannot be applied to discover *fine-grained* topics/events, i.e., sub-events, as they did not consider the near-duplicate characteristic of tweets describing the same *sub-event*. Moreover, in some studies, the detected events were groups of terms, with each group representing one event, which made it quite hard to interpret and understand. Besides, most feature-pivot methods can only be applied on offline datasets, thus cannot generate a timeline of one ongoing *major event* from the tweet stream, because no reliable features of all the tweets in the tweet stream can be provided. Different clustering algorithms on tweets have been proposed by *document-pivot* methods. Some research works [34, 104, 219] applied online incremental clustering approaches by measuring the cosine similarities between the tweets' vector representations. [33] proposed an ensemble clustering approach that combined multiple clustering solutions. Their features included terms, time in minutes and geographic locations. They needed labelled training data to tune the cluster thresholds and weights for different clustering solutions, which can be quite labour-intensive to achieve on regular occasions. Olteanu et al. [207] considered two tweets to be near-duplicates, if their longest common subsequence was 75% or more of the length of the shortest tweet. **Unlike these methods**, the proposed approach aims at tackling the problem of clustering near-duplicate tweets describing the same *sub-event* from the tweet stream. Special measures towards the *low-overlapping* level of near-duplicate tweets, such as extracting key terms and considering key terms with high Jaro-Winkler metric as the same key term, are taken. [6, 183, 236] utilised Locality Sensitive Hashing (LSH) techniques to group tweets into buckets; tweets in the same bucket were considered as duplicate tweets. LSH techniques could increase the search efficiency. However, it is not intuitive to incorporate some specific strategies, such as the key term cluster-based representation and the URL-based clustering strategy, into the hash functions.

Recently, a limited number of researchers have been devoting their time and effort to find methods of summarising detected events with timestamps in order to generate timelines for the major events. However, most of them have different focuses than my work. In [126, 182, 295], researchers performed timeline generation for news articles. In [295], researchers generated trajectory timelines by jointly optimising relevance, coverage, coherence and diversity of sentences. Researchers in [126] detected local/global sub-events based on the part-whole relationship with the major event, then they performed extractive summarisation for the sub-events based on the popularity of local/global aspects during a certain period. In [182], researchers generated updated summarisation for news stream by adaptively altering the volume of updates and ranking the candidate summary sentences. Recent timeline generation works on tweets include [50, 170, 171, 246, 282, 301, 329]. Researchers in [50] learnt the underlying hidden state representations of long-running structure-rich events via Hidden Markov Models, each hidden state in their model corresponding to one class of sub-events. However, their model is dependent on features based on all the tweets in the dataset and there can only exist one sub-event at one timestamp; besides, the evolution of many major events on Twitter, such as “Ebola outbreak”, does not have clear underlying structures, which limits the application of their model. In [171], researchers first proposed a language model with dynamic pseudo relevance feedback to retrieve relevant tweets given an event query; then they constructed a multi-view tweet graph with the relevant tweets; after that, they managed to extract the representative tweets by finding a minimum dominant. However, their model also dependent on features based on all the tweets in the dataset. Researchers in [246, 282, 329] developed frameworks for the tweet stream, however, they only generated one item for the timeline if there were quantified variations, which made their timeline not sensitive for major events with a lot of sub-events happening at the same time, with different durations and levels of influence. In [170], researchers introduced a non-parametric multi-level Dirichlet Process model to extract events of interest only to individuals and their followers on Twitter, instead of *major events* of interest to the general public. In [301], researchers employed Determinantal Point Processes to extract a small set of representative tweets by optimising the topical relevance and overall selectional diversity from offline tweet dataset. **My work in this chapter is different from all former works on timeline generation from the following perspectives**: my proposed approach is able to process the tweet stream to generate a real-time timeline;

only tweets reporting facts relevant to the interested major event are considered; the online incremental clustering algorithm considers different levels of near-duplication by employing key terms in the tweets and considering different textual variants of the key terms; a more rigid hierarchical clustering step is applied to improve the clustering quality of online incremental clustering; the scenarios of reporting former sub-events with and without new updates are considered.

Another track of related research is Disaster Surveillance on Twitter. While Event Detection methods are widely used [20, 80, 236] in this research track, there were also some works, such as [306], which tried to correlate the number of Ebola outbreaks with the number of the symptom mentions of Ebola on Twitter. Although [306]'s results showed that the correlation was quite low, my results demonstrate that, with detailed textual analysis, Twitter can still provide some earlier alarms than traditional media about the outbreaks in some countries.

5.5 Conclusion

In this chapter, I have proposed an approach to detect and summarise fine-grained topics (sub-events or sub-topics) for the pre-known high-impact event (major event) in real-time. This approach consists of four stages: real-world events reporting tweets extraction, online incremental clustering, post-processing and sub-events summarisation. Using "Ebola outbreak" as the pre-known major event, I have applied the proposed approach on a tweet stream, and have evaluated the performance of each stage of the approach. The results have shown that the proposed approach was significantly better than several baselines, in terms of *clustering precision* and *compression ratio*, and could generate early alarms for disaster surveillance. As such, this approach is the answer to RQ2, proving that timelines of high-impact events can be generated automatically from the tweet stream.

The proposed approach is generic enough, as only event-independent features are used for all the stages, so it opens up the possibility to be applied to various major events. The automatic timeline generation approach is a promising supplement and replacement of the user-generated timeline, which could provide people with more insights about the real-time status of the major events they care about.

In Chapter 4 and this chapter, I have analysed the topical aspects of *entities* and sub-topics of *major events*, respectively. In the following chapters, I will explore the opinions expressed in social media textual content.

Chapter 6

Analysing Entity-centric Sentiment Bias in Multilingual Wikipedia

From this chapter onwards, I present my work related to opinion mining of textual content on social media, which includes this chapter, Chapter 7 and Chapter 8. In Chapter 4, I have analysed the differences in *topical aspects* related to real-world entities in multilingual Wikipedia. In this chapter, I propose a framework, to systematically extract the variations in *sentiments* associated with real-world entities in different language editions of Wikipedia, in order to answer *RQ3. Is there language-specific sentiment bias in multilingual Wikipedia, when talking about certain entities?* Other studies on Wikipedia can be found in Chapter 4 and Chapter 7. The work in this chapter has been published in [318, 321].

6.1 Introduction

As said in Section 4.1, different language editions of Wikipedia evolve independently and can provide a rich source for cross-cultural analytics. Possible sources for the content on Wikipedia include books, journal articles, newspapers, web pages, sound recordings[1], etc. Due to its openness to multiple forms of contribution, the articles on Wikipedia can be viewed as a summarisation of thoughts in multiple languages, about specific entities and events. Since people with different language backgrounds

[1] `http://en.wikipedia.org/wiki/Wikipedia:Citing_sources`

share different cultures and sources of information, semantic differences between language-specific editions of Wikipedia may occur when discussing certain entities. The semantic differences are reflected in the entity's language-specific topic representations. In Chapter 4, I have analysed the similarities and differences with respect to the entity's related topical aspects in multilingual Wikipedia. In this chapter, I take a step further, by analysing the existence and extent of entity-centric language-specific **sentiment bias** in multilingual Wikipedia.

Although the "Neutral point of view" (NPOV)[2] is Wikipedia's core content policy, implicit sentiment expression is inevitable in this user-generated content. As pointed out in [148], even if an article is written in compliance with the NPOV, the varying cultural, social, national and lingual backgrounds can have an enormous influence; hence, content in Wikipedia can only be as professional and balanced as its authors. Moreover, Wikipedia web pages are actually allowed to contain opinions, as long as they come from reliable authors[3]. In [101, 102], researchers discovered the slant in English Wikipedia articles on US political topics. As for multilingual Wikipedia, most studies discovered the differences in content between articles discussing the same concept in the different language editions of Wikipedia [30, 113, 181]. Few of them analysed the language-specific bias from the sentiment, or tone perspective, except [11, 49, 235]. The language-specific sentiment bias was *manually* examined and verified for events and famous persons in [235] and [49], respectively. In [11], researchers employed statistical classifiers to prove there were differences in views between English and Arabic Wikipedia articles discussing famous persons. These works have proved that although Wikipedia aimed at the NPOV, such NPOV can vary across its language editions, building linguistic points of view (LPOV) [235].

A limitation of the former research on entity-centric *sentiment* bias of *multilingual Wikipedia* is their focus on the comparative analysis of *one entity* at the *article level.* First, considering the size and scale of the multilingual Wikipedia, *automatic and generalisable* approaches for analysing the sentiment bias should be developed. Moreover, as pointed out in Chapter 4, even a dedicated Wikipedia article can typically cover only a part of the descriptions with respect to an entity, and thus cannot fully reflect the collective language-specific sentiment bias associated with this entity in the Wikipedia corpus. To solve the first problem, I exploit

[2] `http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view`

[3] `http://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources`

a *lexicon-based* sentiment analysis approach, which achieves consistent performance on entities from various domains to automatically generate interpretable results that illustrate the degrees of neutrality of multilingual Wikipedia with respect to the entities. Since the development of sentiment analysis techniques, very few studies considered sentiment analysis on multilingual text collections [74]. Moreover, existing sentiment analysis techniques mostly focus on document collections from specific domains with explicit sentiment expressing purpose and often having a clear structure (e.g., movie reviews). Given its NPOV aim, such existing tools are not directly applicable to determine the *collective* language-specific bias in an encyclopaedia corpus like Wikipedia, where I have been expecting much more *moderate and gradable differences*, which I aimed at capturing. As for the second problem, an exhaustive search of all mentions of every entity in Wikipedia does not appear feasible, due to the size and the constant growth of the dataset. In this chapter, I apply the proposed *Graph-based* context creation approach in Section 4.2.4 to collect multiple occurrences of the entity across the articles in a Wikipedia language edition. In summary, my proposed framework is able to *automatically and efficiently quantify the entity-centric language-specific sentiment bias of multilingual Wikipedia at the corpus level.*

I apply the proposed framework to analysis the language-specific sentiment bias of five Wikipedia language editions on 219 entities with worldwide influence. The results show that although the majority of content in multilingual Wikipedia is obeying the NPOV principle, a moderate but stable amount of sentiment-expressing information (around 8% in average, but differs from entity to entity) is to be found in every language edition, representing positive as well as negative sentiments; importantly, these sentiments, and the entities they refer to, are often language-specific.

6.2 Entity-centric Analysis of Sentiment Bias in multilingual Wikipedia

The proposed framework for analysing the sentiment bias in multilingual Wikipedia is presented in Figure 6.1.

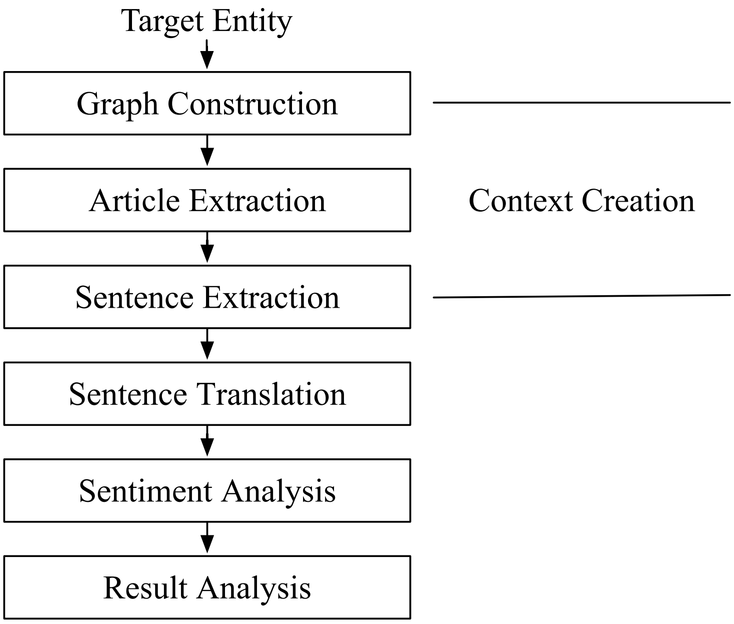

Figure 6.1: Framework for analysing entity-centric sentiment bias in multilingual Wikipedia.

6.2.1 Extracting the Sentences Mentioning the Target Entity in Multilingual Wikipedia

In order to analyse the strength of the sentiment towards the target entity in a given Wikipedia language edition, I need to extract all the texts that mentions the target entity.

For example, as illustrated in Figure 6.2, in English Wikipedia, the entity "GlaxoSmithKline" (a British healthcare company) can be mentioned in articles from various domains, which include the article describing "AT&T" (an American multinational telecommunications conglomerate), the article describing Chlorambucil (a chemotherapy medication), the article describing "DTP-vaccine" (a class of combination vaccines), etc. As currently there are more than four million articles alone in the English Wikipedia, and only a few of them are relevant to the specified entity, it is not efficient to extensively analyse all the articles for each target entity.

In Wikipedia, editors have been using *Interwiki links (in-links)*[4] to link mentions of the entities in one Wikipedia language edition to the Wikipedia articles describing these entities. In English Wikipedia, in-links to the Wikipedia article

[4] https://www.mediawiki.org/wiki/Help:Links#Interwiki_links

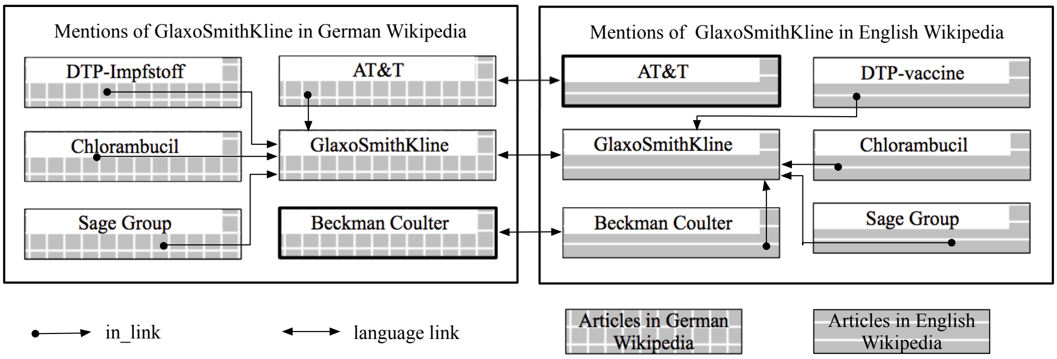

Figure 6.2: Mentions of "GlaxoSmithKline" in the German and English Wikipedia

describing "GlaxoSmithKline" can be detected from Wikipedia articles describing "DTP-vaccine", "Chlorambucil", "Beckman Coulter", "Sage Group", etc., which can provide some clues about where the target entity is mentioned in the Wikipedia corpus. However, since these in-links are manually edited by Wikipedia contributors, a lot of in-links are missing. For example, in the English Wikipedia, "GlaxoSmithKline" is also mentioned in the article describing "AT&T", but there is no in-link from that article to the article describing the target entity. Interestingly, in the German Wikipedia, the article describing "AT&T" contains an in-link to the German Wikipedia article describing "GlaxoSmithKline", and the German Wikipedia article describing "GlaxoSmithKline" is linked to the English Wikipedia article describing "GlaxoSmithKline" by *Interlanguage links (language links)*[5]. Thus, the link structure of Wikipedia can help to locate most mentions of the target entity in a Wikipedia language edition without exhaustively searching in millions of articles. However, the articles mentioning the target entity cannot be directly employed to perform sentiment bias analysis, as not every sentence in these articles are relevant to the target entity. Furthermore, a simple string matching cannot satisfy the aim to be as comprehensive as possible, as the same target entity can occur in various surface forms, and the same surface form can refer to different entities, depending on the surrounding texts.

The *Graph-based* context creation approach proposed in Section 4.2.4 narrows down the search space of articles by employing the linking structure of multilingual Wikipedia, and locates the sentences mentioning the target entity in any of its surface forms exploiting the Entity Disambiguation tool (DBpedia Spotlight), thus the

[5] https://www.mediawiki.org/wiki/Help:Links#Interlanguage_links

resulting sentences can be considered comprehensive and accurate for the following entity-centric sentiment bias analysis.

6.2.2 Sentence Translation

For a fair comparison, the sentiment analysers should have a consistent performance on different languages, which is not feasible to achieve when using multiple sentiment analysers. Besides, there are very limited resources for sentiment analysis for non-English texts. Thus, instead, in order to bring the multilingual texts to a common denominator, all the non-English sentences mentioning the target entity were translated into English, using automatic translation methods, which makes it possible to use the same English sentiment analysis resources to measure the sentiment strength of the multilingual text. If there are some languages that are not supported by the Entity Disambiguation tool employed in the Sentence Extraction step, the Sentence Translation step can be executed before the Sentence Extraction step. Nowadays, machine translation has become a mature technique, which is widely used in research and business, and multiple translation tools have been released by different organisations. Among them, I selected Google Translate[6] to translate all non-English sentences to English, for its good accuracy and rich usage history in the multilingual sentiment analysis area, as well as due to its accessibility and reliability. This practice is relatively common. For example, Wan [276] used Google Translate to close the gap between an English training data set and Chinese test data set; Banea et al. [29] employed Google Translate on Romanian and Spanish texts to use English subjectivity analysis resources on them; in [32], researchers argued that the translated texts were sufficient to accurately capture the sentiment, particularly when aggregating sentiment from multiple documents, which was the same for my case; in [28], researchers conducted several experiments and concluded that the current machine translation systems had reached a reasonable level of maturity to produce reliable training data for languages other than English. Former studies have shown the effectiveness of machine translation services and their influence on the sentiment analysis results is minor. Besides, I apply a lexicon-based sentiment analysis approach, so that grammatical errors that could have been introduced during the translation step will not affect the sentiment analysis step. This is an additional cautionary measure, as I expect the lexicon-based approach to be

[6] `https://translate.google.co.uk/`

affected less by the errors introduced by the translation, unlike the learning-based sentiment-analysing approaches.

6.2.3 Language-specific Sentiment Bias Analysis

I employ a lexicon-based approach, the introduction of which can be found in Section 3.3.1, to measure the language-specific sentiment bias in multilingual Wikipedia. I am interested in the aggregated sentiment strengths of different Wikipedia language editions, rather than the sentiments of separate sentences. As illustrated in [206], for the lexicon-based approaches, when facing a fairly large number of sentences, the errors in polarity detection will cancel out relative to the quality. Moreover, in [119], researchers mentioned that learning-based approaches which were trained to maximise the percent of documents correctly classified, were likely to generate biased estimates of the proportions of documents in given categories. Furthermore, since even for one entity, the sentences mentioning it may come from various domains, the lexicon-based approaches are influenced less by the domain-dependent problem than the learning-based approaches. Besides, the lexicon-based approaches can generate interpretable results, which allow me to perform a more detailed case study.

In order to enable homogeneous processing of entities from different domains and languages, obtain aggregated and graded sentiment strength scores, I select SentiWordNet [25]: a state-of-the-art lexicon containing the sentiment valences of more than 100,000 (word, POS-tag) pairs, being the sentiment lexicon with the widest coverage to date. SentiWordNet has been used in many previous works to analyse sentiment polarity and strength [52, 69, 74, 189, 247]. It annotated all the (word, POS-tag) pairs in WordNet [188] with three numerical scores (adding to 1): positive, negative and objective, each in the range of 0 to 1.

To enable sentiment analysis, each translated sentence mentioning the target entity is POS-tagged with the Stanford POS Tagger [269]. Then the lemmatisation is performed on each (word, POS-tag) pair with NLTK[7] and I use the lemmatised words and their POS tags to obtain the positive, negative and objective sentiment scores from SentiWordNet.

At the sentence level, I aggregate the sentiment scores of the (word, POS-tag) pairs in the sentence. To eliminate the influence of the length differences among

[7] https://www.nltk.org/

sentences, following [52, 69, 74, 189, 247], I normalise the resulting sentiment scores by the number of the (word, POS-tag) pairs that have matches in SentiWordNet from this sentence. I represent the m^{th} sentence that mentions the target entity, in language l's edition of Wikipedia by $s_{l,m}$. The positive, negative and objective sentiment scores of $s_{l,m}$ towards the target entity, which are represented by $POS_{l,m}$, $NEG_{l,m}$ and $OBJ_{l,m}$, respectively, are calculated as follows:

$$POS_{l,m} = \frac{\sum_{n=1}^{N_{l,m}} pos_{l,m,n}}{N_{l,m}}, \tag{6.1}$$

$$NEG_{l,m} = \frac{\sum_{n=1}^{N_{l,m}} neg_{l,m,n}}{N_{l,m}}, \tag{6.2}$$

$$OBJ_{l,m} = \frac{\sum_{n=1}^{N_{l,m}} obj_{l,m,n}}{N_{l,m}}, \tag{6.3}$$

where $pos_{l,m,n}$, $neg_{l,m,n}$ and $obj_{l,m,n}$ represent the positive, negative and objective score of the n^{th} matched (word, POS-tag) pair in sentence $s_{l,m}$; $N_{l,m}$ is the total number of matched (word, POS-tag) pairs in $s_{l,m}$.

The numbers of sentences extracted for a given entity from different Wikipedia language editions vary. Therefore, to make the sentiment scores comparable across different language editions, I need to further normalise the sentiment scores by taking into account the number of sentences extracted from the language edition. To this extent, I build average positive, negative and objective scores per sentence in a language, for each target entity.

The positive, negative and objective sentiment scores for a language l towards the target entity, which are represented by $\mathbf{POS_l}$, $\mathbf{NEG_l}$ and $\mathbf{OBJ_l}$, respectively, are calculated as follows:

$$\mathbf{POS_l} = \frac{\sum_{m=1}^{M_l} POS_{l,m}}{M_l} \tag{6.4}$$

$$\mathbf{NEG_l} = \frac{\sum_{m=1}^{M_l} NEG_{l,m}}{M_l} \tag{6.5}$$

$$\mathbf{OBJ_l} = \frac{\sum_{m=1}^{M_l} OBJ_{l,m}}{M_l} \tag{6.6}$$

where M_l is the total number of sentences that mention the target entity in l's edition of Wikipedia.

6.3 Experimental Results

6.3.1 Experimental Setup

To detect entity-centric language-specific sentiment bias in multilingual Wikipedia, I applied the proposed framework in a case study and provided the insights obtained.

While the framework presented in Section 6.2 is, in principle, language independent, it relies on automatic translation from the target language to English. I only select European languages on which Google Translate achieved desirable performance, as well as supported by DBpedia Spotlight. The Wikipedia language editions included: English (EN) Wikipedia, Dutch (NL) Wikipedia, German (DE) Wikipedia, Spanish (ES) Wikipedia and Portuguese (PT) Wikipedia. These editions differ in size, the largest being English Wikipedia (with more than 4.7 million articles), followed by German Wikipedia and Dutch Wikipedia (with about 1.8 million articles each), Spanish Wikipedia (about 1.1 million) and Portuguese Wikipedia (about 800 thousand articles)[8]. The target entities included a total number of 219 entities with worldwide influence that came from four categories, which were more likely to attract language-specific bias, as my target entities. These four categories were: multinational corporations (55 entities), politicians (53 entities), celebrities (55 entities) and sports stars (56 entities). Each category included entities originating from countries that used one of the five target languages as official languages, in order to verify if the strength of the sentiments towards an entity is different in the countries of their origin. For each entity, about 1,000 sentences that mentioned it were retrieved in one Wikipedia language edition.

6.3.2 Result Analysis

I obtained the objective, positive and negative scores of each Wikipedia language edition towards the target entities according to Section 6.2.3. The sample set of target entities described here, and the summary of their sentiment analysis results, are presented in Table 6.1. Only 10 representative entities from each category are listed. Please note that, due to the open and international nature of Wikipedia,

[8] http://meta.wikimedia.org/wiki/List_of_Wikipedias

with contributions from all around the globe, I do not equate language with the nation.

In Table 6.1, "+" and "−" separately represent the average positive and negative scores of a Wikipedia language edition towards the target entity; "#" represents the number of sentences mentioning the entity extracted from the specific language edition; "L" represents the official language of the entity's origin country.

This table shows that, for some entities, the number of occurrences varies a lot from language to language. The number of occurrences of the entities in the different language editions is influenced by various factors, including the size of the Wikipedia edition, as well as the origin of the entity. Although the English Wikipedia — the largest Wikipedia language edition — contains the majority of entity occurrences, some entities — like Angela Merkel, the Chancellor of Germany, and Mark Rutte, the Prime Minister of the Netherlands — are more frequently mentioned in the local Wikipedia editions.

Because the number of named entities mentioned on Wikipedia is extremely large, it is not possible to apply the proposed framework on all of them. Based on a limited number of 219 entities, the objective information across multilingual Wikipedia constitutes about 92%. The remaining (about 8%) contain positive and negative sentiments, which vary slightly, dependent on the particular entity and language. This has proven the existence of language-specific sentiment bias in multilingual Wikipedia for entities. For all the five target languages, their average proportions of positive sentiment scores and negative scores for each category are at the same level. This means that, for the target entities, there are not some languages which appear to be significantly more positive or negative than other languages. For individual entities, the positive and negative sentiment scores are always in the range of [0.02, 0.09]. This indicates that, although language-specific bias exists in Wikipedia, due to the NPOV policy, this bias can be kept at a relatively low level. Moreover, controversies among different Wikipedia language editions seem to be solved by allowing both positive and negative sentiment expressions to co-exist, instead of removing the bias completely. For example, for the named entity "Thomson Reuters", about 6% of German Wikipedia holds positive sentiment and 3% holds negative sentiment. While in Portuguese Wikipedia, the positive sentiment score and the negative sentiment score change to 4% and 3%, respectively. Maybe it is not unreasonable to say that the German-speaking people like Thomson Reuters more than the Portuguese-speaking people. For other named entities, such as "Unilever",

Table 6.1: Sentiment bias of 219 named entities from four categories in multilingual Wikipedia.

Target entity	**NL**			**DE**			**EN**			**ES**			**PT**			**L**
	#	+	−	#	+	−	#	+	−	#	+	−	#	+	−	
Multinational corporations																
GlaxoSmithKline	51	0.05	0.04	182	0.05	0.03	1076	0.04	0.03	80	0.05	0.03	30	0.05	0.03	EN
News Corporation	229	0.03	0.02	446	0.04	0.02	6879	0.04	0.03	552	0.04	0.03	251	0.04	0.02	EN
Royal Dutch Shell	1185	0.04	0.03	1426	0.04	0.03	6937	0.04	0.03	727	0.04	0.03	312	0.04	0.03	NL
Elsevier	434	0.04	0.02	338	0.04	0.03	1209	0.04	0.03	60	0.04	0.03	4	0.03	0.02	NL
Hugo Boss	70	0.05	0.03	540	0.04	0.03	702	0.04	0.03	144	0.05	0.03	89	0.05	0.05	DE
Unilever	443	**0.04**	**0.03**	557	**0.04**	**0.03**	1826	**0.04**	**0.03**	182	**0.04**	**0.03**	182	**0.04**	**0.03**	EN
Tesla Motors	57	0.05	0.04	321	0.04	0.03	1462	0.04	0.03	622	0.04	0.03	63	0.03	0.02	EN
BMW	1130	0.05	0.03	3760	0.04	0.03	5522	0.04	0.03	868	0.05	0.03	392	0.04	0.03	DE
Thomson Reuters	81	0.04	0.03	428	**0.06**	**0.03**	1802	0.05	0.02	97	0.04	0.02	82	**0.04**	**0.03**	EN
Goldman Sachs	216	0.05	0.03	913	0.04	0.03	4911	0.04	0.03	369	0.05	0.03	189	0.05	0.03	EN
Avg of 55	270	**0.04**	**0.03**	763	**0.04**	**0.03**	3846	**0.04**	**0.03**	508	**0.04**	**0.03**	277	**0.04**	**0.03**	
Politicians																
Bill Clinton	1076	0.05	0.04	3062	0.05	0.04	29351	0.05	0.04	2021	0.05	0.04	1075	0.05	0.04	EN
Stephen Harper	116	0.05	0.03	339	0.04	0.03	5321	0.05	0.04	141	0.04	0.04	69	0.04	0.03	EN
Tony Blair	407	0.05	0.04	1508	0.05	0.04	11739	0.05	0.04	913	0.05	0.04	389	0.05	0.03	EN
David Cameron	181	0.04	0.03	708	0.05	0.03	7710	0.05	0.04	476	0.05	0.05	142	0.04	0.04	EN
Angela Merkel	406	0.05	0.04	**4666**	0.05	0.05	2840	0.05	0.04	583	0.05	0.04	302	0.05	0.04	DE
Mark Rutte	**687**	0.05	0.03	178	0.04	0.03	479	0.05	0.04	74	0.04	0.04	28	0.04	0.04	NL
Dilma Rousseff	169	0.04	0.03	236	0.05	0.04	1106	0.05	0.04	436	0.04	0.03	2315	0.05	0.04	PT
Hillary Clinton	541	0.06	0.03	964	0.05	0.04	13155	0.05	0.04	1051	0.05	0.04	558	0.05	0.04	EN
Michelle Bachelet	48	0.05	0.03	156	0.05	0.03	850	0.04	0.04	2548	0.05	0.04	163	0.05	0.03	ES
Heinz Fischer	33	0.06	0.03	617	0.05	0.03	245	0.05	0.04	37	0.05	0.04	20	0.04	0.04	DE
Avg of 53	282	**0.05**	**0.04**	885	**0.05**	**0.04**	5485	**0.05**	**0.04**	814	**0.05**	**0.04**	286	**0.05**	**0.04**	
Celebrities																
Til Schweiger	12	0.03	0.02	565	0.04	0.03	301	0.05	0.02	37	0.04	0.03	12	0.06	0.02	DE
Eddie Van Halen	166	0.05	0.03	389	0.05	0.03	2669	0.05	0.04	408	0.06	0.04	439	0.05	0.03	NL
Antonio Banderas	116	0.05	0.03	300	0.06	0.03	1412	0.05	0.04	742	0.05	0.03	248	0.05	0.03	ES
Enrique Iglesias	108	0.04	0.02	208	**0.09**	0.04	2985	0.05	0.04	872	0.05	0.04	407	0.04	0.03	ES
Taylor Swift	101	0.04	0.03	633	**0.07**	0.03	6252	0.05	0.03	2222	0.05	0.04	2499	0.05	0.03	EN
Christoph Waltz	36	0.06	0.04	305	0.06	0.02	344	0.06	0.03	103	0.06	0.04	76	0.05	0.02	DE
Rodrigo Santoro	21	0.02	0.02	45	0.05	0.02	254	0.05	0.03	69	0.06	0.03	186	0.05	0.04	PT
Colin Firth	127	0.06	0.03	357	0.06	0.04	1259	0.05	0.03	363	0.06	0.03	212	0.05	0.03	EN
Katy Perry	293	0.04	0.03	781	0.06	0.04	5457	0.05	0.03	1963	0.05	0.04	1756	0.05	0.04	EN
Shakira	223	0.05	0.03	605	**0.07**	0.04	4358	0.05	0.03	2423	0.05	0.04	915	0.04	0.04	ES
Avg of 55	147	**0.05**	0.03	369	**0.05**	0.03	2491	**0.05**	**0.04**	727	**0.05**	**0.04**	521	**0.05**	0.03	
Sports stars																
Andy Murray	315	0.04	0.05	458	0.04	0.04	3701	0.05	0.04	795	0.04	0.06	243	0.04	0.05	EN
Lionel Messi	429	0.05	0.03	382	0.06	0.03	3643	0.05	0.04	1556	0.05	0.03	642	0.05	0.04	ES
David Villa	104	0.04	0.04	151	0.05	0.03	1178	0.05	0.05	443	0.05	0.05	158	0.05	0.04	ES
Arjen Robben	274	0.05	0.04	226	0.04	0.04	1090	0.05	0.04	190	0.05	0.05	160	0.06	0.05	NL
Wesley Sneijder	252	0.04	0.03	136	0.05	0.02	564	0.05	0.04	149	0.05	0.04	108	0.05	0.03	NL
Tiger Woods	539	0.06	0.03	209	**0.07**	0.03	3987	0.05	0.04	182	0.05	0.03	77	0.06	0.05	EN
Lukas Podolski	57	0.05	0.02	306	0.04	0.04	610	0.05	0.05	92	0.05	0.04	63	0.05	0.03	DE
Miroslav Klose	93	0.04	0.04	505	0.05	0.04	682	0.05	0.04	239	0.05	0.03	131	0.05	0.04	DE
Cristiano Ronaldo	314	0.05	0.03	578	0.05	0.03	4099	0.05	0.04	1263	0.05	0.04	1011	0.05	0.04	PT
Rafael Nadal	573	0.04	0.05	766	0.04	0.04	4043	0.04	0.04	1771	0.05	0.06	624	0.04	0.05	ES
Avg of 56	260	**0.05**	0.03	550	**0.05**	**0.04**	2608	**0.05**	**0.04**	580	**0.05**	**0.04**	289	**0.05**	**0.04**	

all the five language editions of Wikipedia contain almost the same level of positive sentiment and negative sentiment, the scores of which are 4% and 3%, respectively. Nevertheless, I did not observe any systematic increase in the positive or negative scores of the language corresponding to the country of the entity origin.

There are some other interesting patterns. For example, all the five languages average proportions of the positive and negative sentiment scores of corporations are slightly lower (about 1%) than their corresponding proportions for the people-related categories. This means that Wikipedia contributors tend, in general, to like people more than corporations — a possibly foreseeable outcome. However, there are some outliers. The exception is formed by the average negative sentiment proportion of celebrities in the Dutch, German and Portuguese Wikipedia, respectively, as well as the average negative sentiment proportion of sports stars in the Dutch Wikipedia. The probability values of the t-test in Table 6.2 confirm the statistical significance of the sentiment differences except the above outliers. In Table 6.2, I use "**M**", "**P**", "**C**" and "**S**" to represent multinational corporations, politicians, celebrities and sports stars, respectively; the other denotations are the same as the ones in Table 6.1. Specifically, the "**M-P(+)**" column is the set of probability values of t-test between the positive sentiment scores of multinational corporations (**M**) and politicians (**P**).

Table 6.2: Probability values of the t-test.

L	**M-P(+)**	**M-P(−)**	**M-C(+)**	**M-C(+)**	**M-S(+)**	**M-S(−)**
NL	4.34×10^{-6}	6.40×10^{-6}	5.59×10^{-3}	4.98×10^{-1}	2.15×10^{-4}	1.18×10^{-3}
DE	1.21×10^{-7}	6.23×10^{-15}	2.46×10^{-8}	1.07×10^{-1}	1.73×10^{-13}	4.00×10^{-7}
EN	6.04×10^{-17}	2.37×10^{-26}	7.64×10^{-17}	4.26×10^{-11}	2.60×10^{-10}	6.33×10^{-22}
ES	1.49×10^{-4}	9.79×10^{-16}	1.09×10^{-11}	5.11×10^{-6}	1.04×10^{-19}	1.90×10^{-15}
PT	2.55×10^{-3}	3.86×10^{-5}	4.73×10^{-9}	1.93×10^{-1}	1.37×10^{-10}	9.31×10^{-9}

On the other hand, some celebrities and sports stars have much higher positive scores than the rest. Examples are Enrique Iglesias, Taylor Swift, Shakira and Tiger Woods in the German Wikipedia. After a preliminary analysis of representative sentences with high positive scores, I attribute this to the following reasons. First, these celebrities and sports stars tend to have larger numbers of fans than average. These fans have very positive feelings towards them, which results in frequent usage of positive sentimental terms when discussing them on Wikipedia. Examples include *"Shakira's Ojos Así performance was chosen as the best Latin Grammy performance* *of all time"* and *"The most successful song of the year was Bailando by Enrique Iglesias"*. Second, these celebrities and sports stars achieve awards or victories more often than average, the inclusion of these awards or victories also greatly contributes to the positive sentiment scores. For example, *"Tiger Woods with his 14 victories since 1997, the most successful active golfer and the second most successful in the eternal ranking"* and *"In addition to that, so Swift received BMI President's Award, Which honours at exceptional individual in entertainment industry deserving of special recognition"*.

In the following, I am going to analyse some of the results in more detail.

6.3.3 Language-specific Affective Facts

To explore the underlying reasons that lead to the language-specific sentiment bias in multilingual Wikipedia, I further analysed the automatically extracted sentences with high positive/negative scores for two entities: GlaxoSmithKline — a British multinational pharmaceutical company, and Angela Merkel — the Chancellor of Germany.

GlaxoSmithKline occurs more frequently in the English and German Wikipedia, while less in the Dutch and Portuguese editions. I find many sentences with high positive scores from the German and English Wikipedia are about the effectiveness of the various vaccines developed by GlaxoSmithKline. However, in the Dutch and Portuguese Wikipedia, the sentences mentioning GlaxoSmithKline with high positive sentiment scores are mostly the description of the economical development of this company. I conjecture that facts relevant to the performance of GlaxoSmithKline's medicine are more likely to provoke the positive sentiment of the English and German speaking community; facts relevant to GlaxoSmithKline's economical growth are more likely to provoke the positive sentiment of the Dutch and Portuguese speaking community. This information could help the company to take language-specific measures to build its reputation in different language-speaking communities. For English, German, Dutch and Spanish language editions of Wikipedia, the sentences with high negative scores are about a mix of facts relevant to medicine safety issues, the company's lawsuits and its corruption. This illustrates these four language-speaking communities are more affective about the above facts than the Portuguese-speaking community. Furthermore, as possibly to be expected, facts showing some level of locality are more likely to be affective. For example, one of the sentences with high positive scores in German Wikipedia mentions: *"Under the umbrella of GlaxoSmithKline, Odol*[9] *has become the largest oral hygiene brand in Germany"* — a fact that is relevant for the German Wikipedia only.

As for Angela Merkel, the majority of the entity occurrences are located, as expected, in the German and English Wikipedia. Nevertheless, for all Wikipedia language editions, the sentences about Angela Merkel's success in the elections and the criticism received during her tenure receive high positive and negative sentiment scores, respectively. However, in the German Wikipedia, some sentences about her life before she went on the political stage get relatively high positive scores. Such an example is the sentence *"She was a well-known student with excellent performance in any event at the University of Leipzig"* and *Angela Merkel (then Kasner) was awarded the "Lessing" medal in silver after the tenth grade (1971) for outstanding social and academic performance.* Moreover, some sentences regarding her haircut and clothes receive very high negative scores in German Wikipedia. Sentences describing similar facts are not in the list of sentences with high positive/negative scores for other Wikipedia editions. In English Wikipedia, some sentences reflecting Angela Merkel's international role receive high positive scores, such as *The Indian government presented the Jawaharlal Nehru Award for International Understanding for the year 2009 to Merkel.* For the Portuguese Wikipedia, sentences about Angela Merkel's performance in the economic crisis and on the financial market receive high positive scores.

As these examples illustrate, the entity-centric affective facts are language-specific, the aggregated effects of which lead to a clear entity-centric language-specific sentiment bias in multilingual Wikipedia.

6.4 Related Work

Related works on analysing the differences in the usage and content between different Wikipedia language editions have been summarised in Section 4.6. However, few studies analysed the semantic differences of multilingual Wikipedia from the sentiment perspective, except [11, 49, 235]. In [49, 235], researchers manually examined the extent to which the content and sentiment varied across multilingual Wikipedia articles about the Srebrenica massacre and selected famous persons, respectively. They discovered that the multilingual Wikipedia expressed diverse points of view,

[9] `https://de.wikipedia.org/wiki/Odol`

attributed to specific sets of editors and the references they employed. Researchers in [11] detected the point of view differences between Arabic and English Wikipedia articles towards selected entities, by employing trained classifiers for corresponding languages. However, their method was language-specific, and would require extra annotation and training in order to be extended to other languages and entities from other domains; they employed the *sentences* in the Wikipedia *article* describing the entity as the sentences directly relevant to the entity, which was not comprehensive and accurate. **Unlike previous research**, I analyse the entity-centric language-specific bias of multilingual Wikipedia from the sentiment perspective, using an unsupervised approach; I search for as many as possible the sentences mentioning the entity at the Wikipedia *corpus level* employing the link structure of Wikipedia. The proposed framework is totally automatic and generalisable, and is able to generate reproducible and interpretable results.

Lexicon-based sentiment analysis approach has been commonly applied in quantifying aggregated sentiment information in corpora. Of all the lexicons annotated with sentiment valences reviewed in Section 3.3.1, many researchers selected SentiWordNet, due to its popularity, coverage and availability. In [189], researchers explored how the sentiment expressed in blog posts, measured by SentiWordNet, 'travelled' through hyperlink networks. In [52], researchers analysed the extent of opinionated queries issued on controversial topics using SentiWordNet. In [296], researchers aimed at enhancing the user location preference model with the sentiment valences of user comments derived by SentiWordNet. In [75], researchers used SentiWordNet to evaluate the sentiment of a virtual interviewer's users. To my best knowledge, the work presented in this chapter is the first attempt to analyse the aggregated sentiment bias towards the entity in multilingual Wikipedia with the lexicon-based approach.

6.5 Conclusion

In this chapter I have proposed a novel, *easily-reproducible, automatic framework to analyse and understand entity-centric language-specific sentiment bias in different Wikipedia language editions.* This framework includes the collection of sentences mentioning the target entity by using *in-links* and *language links*, as well as employing a lexicon-based sentiment analysis approach to numerically quantify the aggregated language-specific differences.

I have applied this framework in a case study over five Wikipedia language editions (more than any predecessor), analysing the language-specific sentiment bias for 219 entities representing multinational corporations, politicians, celebrities and sports stars. The results illustrate that the proportion of objective information for any given entity in my study is similar across language editions and constitutes about 92%. The remaining 8% contains positive and negative sentiments, that vary, dependent on the particular entity and language. This may show that the neutrality in Wikipedia is obtained not by neutralising all statements, but by including both positive and negative statements. Thus, RQ3 has been answered: there is language-specific sentiment bias in multilingual Wikipedia, because different language editions of Wikipedia have different proportions of positive and negative sentiments for a given entity.

Whilst the proportion of 8% seems very low, Internet users can spot the implicit sentiment expression and their perceptions towards the mentioned entities may be affected by it. *It should be noted that the proportion of objective information is not equal to the proportion of objective sentences in Wikipedia, as both subjective and objective sentences can contain objective information, as well as positive/negative sentiment information.* To better explain the results, I have further analysed some of the examples to show that even for well-known, internationally relevant entities, their affective facts vary in different language editions of Wikipedia.

In Chapter 4 and this chapter, I have analysed the semantic difference of multilingual Wikipedia with respect to the entity from the topic and sentiment perspectives, respectively. Both of the works were performed at the *corpus* level, facing all the texts in a Wikipedia language edition. In Chapter 7, I will describe the algorithms I have developed to detect the reputation-influential *sentences*, which lead to the aggregated language-specific sentiment bias discovered in this chapter.

Chapter 7

Detecting Reputation-influential Sentences in Wikipedia

In Section 6.3.3, I have analysed some affective sentences on Wikipedia, with implicit sentiment expression towards the entities, extracted via the lexicon-based sentiment analysis approach according to hand-crafted rules. However, even though the lexicon-based sentiment analysis approach can achieve graded and interpretable results at the *corpus level*, it cannot achieve a high accuracy at the *sentence level*, due to the change in granularity. In this chapter, I train classifiers using various features of Wikipedia sentences, in order to learn the differences between *reputation-influential sentences* and *reputation non-influential sentences* automatically. As stated, this chapter, Chapter 6 and Chapter 8 are the part of the study on Opinion Mining of social media text. This chapter answers *RQ4. Can the positive or negative reputation-influential information in Wikipedia be identified?* Other studies on Wikipedia can be found in Chapter 4 and Chapter 6. The work in this chapter has been published in [317].

7.1 Introduction

Wikipedia has become one of the most frequently used websites in people's daily lives. Even when considering only the English Wikipedia, it contains more than 5 million articles and receives more than 5 million views per hour[1]. Such comprehensive information inclusion and huge visiting traffic make Wikipedia influ-

[1] `http://stats.wikimedia.org/EN/Sitemap.htm`

ential for people worldwide. As said, due to the NPOV[2] policy, most sentences in Wikipedia are factual. However, researchers have proved that besides subjective sentences, which express opinions explicitly, *factual sentences* can also express sentiments implicitly through selection of verbs [168], noun phrases [215, 312], or syntactic patterns [100]. Influenced by the contributors' backgrounds, it is not possible for the content on Wikipedia to be absolutely neutral of view [148]. Wikipedia contributors manage to implicitly express their opinions by including selective facts and varying description patterns [49, 235], either purposely or unconsciously, which lead to language-specific sentiment bias in multilingual Wikipedia, as illustrated in Chapter 6. Thus, some sentences on Wikipedia are *polar facts*, which reveal the contributors' opinions towards the entities mentioned in them and, more importantly, aim at influencing Wikipedia users' perception about these named entities. For example, sentences in Wikipedia like "Chevron did not apologise, nor paid the amount of compensation." and "There are some exceptions, such as striker Wayne Rooney, who became extremely unpopular with fans after changing Everton for Manchester United, and is currently always booed when he returns to the stage of his former club." imply Wikipedia contributors' negative opinion towards the mentioned entities, which are "Chevron Corporation" and "Wayne Rooney". Sentences in Wikipedia like "Lady Gaga won two awards, including the prize for best song for Born This Way at the Europe Music Awards." and "Boeing today is a synonymous name for dynamic, impressive aircraft, global air travel, success and economic strength." imply Wikipedia contributors' positive opinion towards the mentioned entities, which are "Lady Gaga" and "Boeing Company". In analysis from the author's point of view, these sentences are generally referred to sentences with *implicit sentiment expressions*, in studies aiming at identifying consumers' opinions in product reviews [168, 268, 312]; or as *biased sentences* in studies focusing on promoting the neutral point of view policy [230, 300]. From the reader's point of view, these sentences were defined as *sentences with reputation polarities* in studies analysing the sentences' implications for the mentioned entities' reputation [14, 72, 96, 214]. It is hard to speculate on the Wikipedia contributors' implicit sentiments hidden in the factual sentences. Especially for persons and companies, which are the target entities for this chapter, the sentiments towards them are not structured around a fixed set of aspects, as in the case of products [72]. On the other hand, it is unfeasible to define bias without considering a specific domain or topic. For above reasons, as well

[2] `http://wikipedia.org/wiki/Wikipedia:NPOV`

as description and explanation convenience, following [14, 72, 96, 214], I analyse the *polar facts*, along with other subjective sentences with explicit sentiment expressions that may appear in Wikipedia, from the reader's point of view and define them as *reputation-influential* sentences of the mentioned persons, or companies. If a sentence can stimulate positive opinions, or have positive reputation implications for the mentioned named entity, then it is a *positive reputation-influential* sentence; if a sentence can stimulate negative opinions, or have negative reputation implications for the mentioned named entity, then it is a *negative reputation-influential* sentence.

In Section 6.3.3, I found that the lexicon-based approach, even though it was suitable for quantifying the aggregated sentiment information at the corpus level, cannot achieve desirable performance when applied to detect the above affective sentences with implicit sentiment expression. For individual sentences, besides the drawbacks I mentioned in Section 3.3.1, the lexicon approach has the following shortcomings with respect to detecting *reputation-influential* sentences in Wikipedia. First, sentences containing positive/negative words or illustrating favourable/unfavourable facts, are not necessarily positive/negative reputation-influential sentences for the mentioned entities. For example, both the sentence "Repeatedly bullied by white children in her neighbourhood, Parks often fought back physically." and the sentence "Despite initially neglecting to comment, Gomez confirmed in 2015 that she had been diagnosed with the auto-immune disease, lupus, and that she had cancelled the tour and entered rehab to undergo chemotherapy." contain some negative words and illustrate that something unpleasant happened to the mentioned entities, but they have no negative implication for the mentioned entities' reputation at all. Second, the lexicon-based approach cannot differentiate between the nuanced sentiments expressed via varied linguistic patterns. For example, the sentence "Jude Law involved in car accident." has different reputation implications for "Jude Law" than the sentence "Jude Law crashes his vintage Mercedes into a London black cab on Drury Lane.", which the lexicon-based approach is not able to capture. The influence of these drawbacks is diminished when aggregating the results of a large number of sentences [206]. However, a novel approach has to be proposed when analysing the problem at a finer granularity.

This chapter aims at the detection of positive and negative *reputation-influential* sentences from Wikipedia articles. This is not a traditional sentiment analysis problem, as the sentiments towards the entities are only implicitly expressed or even hidden in Wikipedia sentences. However, they have positive or negative implications for the mentioned named entities' reputation and can influence people's opinions towards them implicitly. The sentences which are identified as neutral or negative by traditional sentiment analysers can have positive implications for the mentioned entities' reputation and vice versa. To the best of my knowledge, this is the first work to define such a problem for Wikipedia sentences.

I apply a hierarchical classification method to tackle this multi-classification problem (reputation non-influential, positive reputation-influential and negative reputation influential) on Wikipedia sentences, *which have no natural partitioning into domains*. I use multiple lexicons to generate domain independent features. Because of the lack of large annotated datasets from various domains, I generate unsupervised features from the unlabelled dataset. The experimental results prove that the proposed approach has achieved competitive performance on Wikipedia sentences from various domains.

7.2 Data Annotation

I have employed the same dataset created for Chapter 4 and Chapter 6. The dataset consisted of 1,196,403 Wikipedia sentences explicitly mentioning one of the targeted 219 named entities, which came from four popular categories: multinational corporations, politicians, celebrities and sports stars. I have used a crowdsourcing website: CrowdFlower[3] to annotate 5037 sentences (23 sentences per named entity) selected from the dataset into two categories: *reputation-influential* sentence and *reputation non-influential* sentence.

Due to the NPOV policy and collaborative characteristic of Wikipedia, most sentences in Wikipedia are impartial and narrative. This kind of sentences has a minor influence on the mentioned named entities' reputation, as most words included in these sentences are neutral, non-judgmental and unbiased. To avoid the situation that *reputation non-influential* sentences dominate the dataset to be annotated, I applied the lexicon-based approach employed in Section 6.2.3, to increase the percentage of sentences that carry strong subjective (i.e., weak objective, as these were complementary) words into the dataset to be annotated. First, for each named entity, I calculated the average objective score OBJ of all the words in each sentence that mentioned this named entity, as in Equation 6.3. The objective scores of the words can be obtained from SentiWordNet [25]. Second, half of the sentences in

[3] https://www.crowdflower.com/

the dataset to be annotated were sentences with the lowest OBJs. This was due to the fact that words contained in these sentences were relatively strongly subjective in general, thus they were more likely to be *reputation-influential*, and promoted empathy amongst Wikipedia users. Third, the other half of the sentences in the dataset to be annotated were the sentences randomly sampled from the rest, to alleviate the strong subjective polarisation of the dataset to be annotated. Thus, the dataset to be annotated was a combination of the sentences with low OBJs and the sentences retrieved from random sampling. Results showed that, the above method made the extracted dataset contain more balanced proportions of sentences from different categories than absolute random sampling.

The annotators were provided with the sentences to be annotated and their corresponding mentioned named entities, and were asked to label these sentences, based only on the sentence provided, rather than their pre-known information — if these sentences would influence the mentioned named entities' reputation. For the *reputation-influential* sentences, the annotators were further asked to response what kind of influence these sentences would have, *positive* or *negative*. There were three annotators allocated to pass judgment independently on each sentence and more than 1,000 annotators with different backgrounds participated the task. The annotators were free to annotate any number of sentences. Crowdflower provided the confidence score[4] of each label for each sentence, which was calculated as the agreement among multiple annotators on this label, weighted by their accuracy on several test questions annotated by myself. For each sentence, the label with the highest confidence score was chosen as the annotation of the sentence. Similar to [228, 327], I evaluated the annotation quality based on the confidence score of the annotation. For this application, only annotations with confidence scores higher than 0.75 were applied to train the classifiers, which left me with 1,147 *reputation non-influential* sentences, 461 *positive reputation-influential* sentences and 228 *negative reputation-influential* sentences.

[4] http://success.crowdflower.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score

7.3 Two-step Binary Classification Approach for Ternary Classification

The goal is to detect the *positive reputation-influential* and *negative reputation-influential* sentences from Wikipedia. I cast the *reputation-influential* sentence detection as a ternary classification problem on sentences from various domains. All the sentences are classified into three categories: *positive reputation-influential* sentences, *negative reputation-influential* sentences and *reputation non-influential* sentences. Following [31, 211], I apply a two-step binary classification approach for the ternary classification. Essentially, the two-step binary classification approach for ternary classification is a *Hierarchical Classification* [249] approach, which decomposes the multi-classification problem into a set of smaller problems corresponding to hierarchical splits in the category tree representing the relationships among the categories. The hierarchical classification approach was first employed for text categorisation tasks [81,239,255], on which it achieved better performance than the traditional flat multi-classification approach that ignored the category hierarchy. The hierarchical classification approach results in many accurate and specialised classifiers, because training examples of some related sub-categories can be combined; the features which are not useful for flat multi-classification may show their effectiveness when discriminating examples from combined sub-categories. In the first step, the sentences are classified into two categories: *reputation-influential* sentences and *reputation non-influential* sentences. The *reputation-influential* sentences are further classified into *positive reputation-influential* sentences and *negative reputation-influential* sentences. I select for both steps a Support Vector Machine (SVM) classifier with RBF kernel, a binary classifier that has been proven to be effective in many sentence classification applications [106,195,212]. A more detailed description about the SVM classifier can be found in Section 3.3.2.

As, under the strong influence of the NPOV policy, the numbers of sentences from different categories in the annotated dataset is still quite unbalanced, I perform undersampling [107] on the sentences from the *reputation non-influential* category, to balance the number of *reputation-influential* sentences and *reputation non-influential* sentences.

7.4 Feature Extraction and Selection

It is hard for traditional fully-supervised approaches to achieve good performance on sentences from various domains, because they need a large number of annotated sentences from each domain to start with. In this work, I tackle the problem from the following directions: first, I prioritise domain independent features when performing feature extraction; second, I leverage unlabelled sentences to provide topical and word embedding features in order to boost the performance of traditional classifiers; third, I incorporate many lexicons to provide rich, domain independent, prior knowledge for classification.

Since it is difficult to clarify which features are useful for which step, I ran various tests with different subsets of the full feature set for both steps to select the features that perform best. The results of this process are further presented in Table 7.1. To diminish the risk of introducing too many irrelevant features and reduce the dimensionality of the training matrix, I employ Randomized Logistic Regression [184] as a further feature selection step after fixing the feature set for one classifier. Next, I introduce the full feature set used.

7.4.1 Baseline Features

The first set to choose from are baseline features, represented by **FS1**, which are mostly used in classifiers for sentence classification [2, 23, 31], as follows.

1. *Number of words.* Number of words in the sentence.

2. *N-gram features.* The *tfidf* values of unigrams and bigrams in the sentence, as discussed in Section 3.2.1.

3. *Punctuation features.* Number of question marks and number of exclamation marks in the sentence.

4. *POS-tag features.* I use the Stanford POS tagger [269] to POS-tag all sentences. Numbers of adjectives, adverbs, verbs and nouns are included into the feature set.

5. *Dependency features.* I represent all the dependencies as features to capture grammatical relationships between words in the sentence. This is achieved via the Stanford dependency parser [54]. For example, in the sentence "German

Chancellor Angela Merkel and US Vice President Joe Biden condemned the attack on the US mission.", even trigrams are not able to capture the nominal subject relationship between words "Merkel" and "condemned". I represent this dependency as "nsubj_condemned_Merkel" and include the number of its occurrences in the feature set.

7.4.2 Lexicon Features

I have collected all the commonly used biased lexicons and sentiment lexicons, as detailed in Section 3.3.1, and transfer the prior knowledge contained in these lexicons into features, represented by **FS2**, as follows. Future lexicons can be easily included to enrich the feature set.

1. *Opinion Lexicon features.* The Opinion Lexicon [122] contains a positive opinion words list and a negative opinion words list. I include the numbers of positive and negative opinion words from the lexicon that occur in the sentence into the feature set.

2. *Biased Lexicon features.* The Biased Lexicon [230] contains a list of biased words. I include the number of biased words from the lexicon that occur in the sentence into the feature set.

3. *MPQA Subjectivity Lexicon features.* The MPQA Subjectivity Lexicon [288] contains a list of words, with each word's level of subjectivity (strongly subjective or weakly subjective), POS-tag and prior polarity(positive, neutral or negative) provided. I lemmatise both the words in the lexicon and the words in the sentence, and include the number of strong and weak subjective words from the lexicon that occur in the sentence, as well as the number of positive, neutral and negative words occurring in the sentence into the feature set.

4. *SentiWordNet Lexicon features.* The SentiWordNet Lexicon [25] contains a list of (word, POS-tag) pairs, with each (word, POS-tag) pair's positive score, negative score and objective score provided. I use w_n to denote one (word, POS-tag) pair from the lexicon that occur in the sentence, and pos_n, neg_n and obj_n to denote its positive score, negative score and objective score, respectively, where $obj_n = 1 - pos_n - neg_n$. The following features derived based on SentiWordNet Lexicon are included into the feature set: (i) Number of w_n, denoted by N; (ii) Number of w_n which $obj_n > pos_n + neg_n$; (iii) Number of

w_n which $pos_n > neg_n$; (iv) Number of w_n which $neg_n > pos_n$; (v) The sum of all w_n's obj_n; (vi) The sum of all w_n's pos_n; (vii) The sum of all w_n's neg_n; (viii) The maximum of obj_n; (ix) The maximum of pos_n; (x) The maximum of neg_n; (xi) The average of obj_n (as in Equation 6.3); (xii) The average of pos_n (as in Equation 6.1); (xii) The average of neg_n (as in Equation 6.2).

5. *MSOL Lexicon features.* The MSOL Lexicon [192] provides both single-word entries and multi-word expressions with their sentiment labels. I include the number of positive and negative single-word entries/multi-word expressions from the lexicon that occur in the sentence into the feature set.

7.4.3 Unsupervised Features

As I have a large dataset with only a small part of it annotated, I instead decided to use unsupervised features, aiming at gaining additional knowledge from the whole dataset.

1. *Latent Dirichlet Allocation (LDA) features.* I train LDA models [41], which is discussed in more details in Section 3.2.2, with all the sentences in the original dataset, no matter if they are annotated or unannotated, with different numbers of predefined topics $K \in \{50, 100, 200, 300, 400, 500\}$. Then I represent each sentence with its probabilistic vector representation, denoted by **FS3**, with each dimension in the vector denoting the degree to which the kth topic is referred to in the sentence. I incorporate FS3 into the feature set, and test the classifier's performance with different K.

2. *Word embedding features.* In [187], researchers proposed the continuous Skip-gram model, which is described in Section 3.2.3, to learn word embeddings in a new vector space $\mathbb{R}^{d_0}$, in order to capture syntactic and semantic word relationships. I train the word2vec model [187] on all the sentences in the original dataset, using Gensim [231] with a wide range of $d_0 \in \{50, 100, 200, 300, 400, 500\}$ in order to obtain the most suitable vector representations for all the words occurring in the original dataset.

   Word embedding features have been applied in sentence classification tasks, such as [106]. Inspired by [106], when generating the vector representation for sentences, I use tfidf values to weigh each word in order to decrease the influence of unimportant words. I use $x_n \in \mathbb{R}^{d_0}$ to denote the embedding of word w_n in the sentence and $tfidf_n$ to denote the *tfidf* value of w_n. The vector representation of the sentence can be calculated as: $\frac{\sum_{n=1}^{N} tfidf_n x_n}{N} \in \mathbb{R}^{d_0}$. The weighted average of word embeddings have been proven to be more effective sentence representations in classification-related tasks than the non-weighted average of word embeddings, also in [55]. The word embedding-based vector representation of the sentence is included in the feature set, denoted by **FS4**.

7.5 Experimental Results

I investigated two application scenarios. The first scenario was binary classification, in which I only aimed at detecting *reputation-influential* sentences. The second scenario was ternary classification, in which I aimed at deciding both whether one sentence was *reputation-influential* and the direction in which it influenced the entity's reputation. To jointly consider the precision and recall achieved in different categories, I mainly focused on the average F1 scores achieved in different scenarios.

7.5.1 Reputation-influential Sentence Detection

I performed feature selection manually by analysing the classifier's performance with different feature sets on the basis of Randomized Logistic Regression, using 10-fold cross-validation. I did not totally rely on Randomized Logistic Regression for feature selection in order to discover the most effective features in the feature set and discard redundant features. For different feature sets, I used grid search to choose the most suitable number of topics for the LDA-based topical features K, the dimensionality of the word embeddings d_0, the penalty parameter of the SVM classifier C and the kernel parameter for the RBF kernel γ. More detailed discussion about K can be found in Section 3.2.2; more detailed discussion about C and γ can be found in Section 3.3.2.

In Table 7.1, I use FS1 to denote baseline features described in Section 7.4.1, FS2 to denote lexicon features described in Section 7.4.2, FS3 and FS4 to denote topical features and word embedding features, respectively, which were described in Section 7.4.3. FS1234 represents the combination of FS1, FS2, FS3 and FS4. I use P to represent precision, R to represent recall and F1 to represent the F1 score. Table 7.1 shows that the classifier using lexicon features, topical features and word embedding features (FS234) achieves the best performance, which outperforms the benchmark classifier just using baseline features (FS1). The best performance is

Table 7.1: Performance of reputation-influential sentence detection with different feature sets.

Feature set	**Reputation-influential**			**Non-influential**		
	P	**R**	**F1**	**P**	**R**	**F1**
FS1	0.817	0.386	0.521	0.606	0.916	0.729
FS12	0.745	0.750	0.747	0.755	0.747	0.750
FS123	0.765	0.777	0.771	0.780	0.766	0.773
FS124	0.760	0.781	0.769	0.781	0.758	0.768
FS134	0.768	0.711	0.737	0.738	0.791	0.763
FS34	0.771	0.717	0.743	0.743	0.792	0.766
FS24	0.790	0.770	0.780	0.782	0.799	0.790
FS23	0.711	0.726	0.718	0.728	0.711	0.719
FS234	0.781	0.795	**0.788**	0.807	0.782	**0.795**
FS1234	0.788	0.783	0.786	0.786	0.790	0.783

achieved with FS234 when $K = 100$, $N = 100$, $C = 1$ and $\gamma = 0.005$. I found that the increase in the number of topics and the dimensionality of the word embeddings did not always lead to an improvement of the classifier's performance. This is because larger feature spaces are less able to generalise for sentences from various domains.

Both lexicon features and unsupervised features help to increase the average F1 score. The most helpful features are the word embedding features. This illustrates that word embedding features are the best semantic generalisations of the original Wikipedia sentences from various domains. The average F1 score drops after adding baseline features on the basis of lexicon features, topical features and word embedding features. This is because most baseline features, such as n-grams or dependency features, are domain dependent and the classifier is experiencing the overfitting problem. On the one hand, the lexicon features, topical features and word embedding features already capture the useful patterns that are presented in the baseline features. On the other hand, the baseline features include some irrelevant and redundant information that can hurt the classifier's performance. These factors allow the classifier that excludes the baseline features to outperform other classifiers, including the one with all available features.

7.5.2 Positive Reputation-influential, Reputation Non-influential and Negative Reputation-influential Sentences

I conducted similar experiments as in Section 7.5.1 to select the best feature sets and hyper-parameters for the classifier used to distinguish between *positive reputation-influential* sentences and *negative reputation-influential* sentences, and the classifier for one-vs-one multi-classification approach. Interestingly, the best feature sets for these two classifiers were also FS234. I compared the two-step binary classification approach for ternary classification with the benchmark one-vs-one approach [95]. Table 7.2 shows the performance comparison. A macro-averaged F1 score of 0.717 is achieved with the two-step binary classification approach when classifying all the Wikipedia sentences into three categories, higher than the macro-averaged F1 score of the baseline one-vs-one approach, which is 0.705. This is because the *positive reputation-influential* and *negative reputation-influential* sentences share some common characteristics, thus the combination of the sentences from these two categories provides the classifier with more information than differentiating sentences of these two categories from the sentences of the *reputation non-influential* category separately.

Table 7.2: Performance of the two-step binary classification approach and the one-vs-one approach for ternary classification.

Type	**Positive**			**Non-influential**			**Negative**			**Avg.**
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1**
1-vs-1	0.723	0.695	0.708	0.672	0.684	**0.677**	0.725	0.733	0.729	**0.705**
2-step	0.715	0.713	**0.714**	0.668	0.673	0.670	0.766	0.768	**0.767**	**0.717**

7.6 Related Work

There exists some studies on detecting factual sentences with implicit sentiment expression. In [268], researchers paid special attention to objectively verifiable, but evaluative sentences — *polar facts*, when annotating customer reviews; in [312], researchers identified nouns that imply customer's opinions on products; in [127], researchers mined verb expressions implying negative opinions from sentences describing the products' functionalities in customer reviews. In [300], researchers discovered the existence of biased sentences, which had a "tendency or preference towards a particular perspective, ideology or result", in political blogs They proposed that a biased sentence can purport to communicate factually. In [230], researchers identified linguistic cues for biased language using Wikipedia's *refused historical edits* that violated the NPOV policy. In [11], researchers trained classifiers with n-gram features to classify sentences in Wikipedia articles into three point of view classes, positive, negative and neutral; their targets were limited to person entities. Different from [11], I analysed the implicit sentiment expression from the influence of reputation perspective, which has substantially increased the agreement among annotators; I made sure that all the sentences which are to be annotated mentioned their corresponding target entities, which made the annotation dataset more reliable; I exploited various domain independent features, which were less prone to overfitting and achieved a much better performance on sentences from various domains mentioning either person or company entities than the baseline n-gram features. Works on reputation polarity analysis include [14, 72, 96, 214], which focused on the reputation polarity analysis of tweets towards person and company entities. In this chapter, **I defined and tackled a novel sentence classification problem**: detecting *reputation-influential* sentences from the encyclopaedic content.

Various features have been considered when tackling the sentence classification problem. For example, n-grams [2, 31], POS-tags [2, 31], lexicon-based features [23, 31], dependency features [23], LDA-based topical features [292] and word embedding features [106, 260]. Inspired by [260], I tried to train classifiers with combined hand-crafted features and word embedding features to improve the performance of sentence classification. To the best of my knowledge, the classifiers **jointly considered all the available state-of-the-art features**, and are different from former research in the way of extracting and applying them, such as the SentiWordNet features and the word embedding features. The trained classifiers have achieved a promising performance for the proposed task.

Another relevant track of research is Wikipedia-related sentence classification. In [71], researchers trained classifiers to classify the edits in Wikipedia's revision history; in [112], researchers performed automatic textual vandalism detection; in [291], researchers proposed approaches to label personal attack comments on Wikipedia talk pages. All the above works have different objectives with my work presented in this chapter.

7.7 Conclusion

In this chapter, I have proposed an approach to detect *reputation-influential* sentences in Wikipedia. I have applied several lexicons to generate domain independent lexicon features, and have leveraged an unlabelled dataset to generate topical features and word embedding features. All these features have been proven to be functional in the experiments. The classifier can achieve a macro-averaged F1 score of 0.792 on the reputation-influential Wikipedia sentence detection task, which is a binary classification problem. I have further adopted a two-step binary classification approach when performing the task of classifying all the Wikipedia sentences into three categories: *positive reputation-influential*, *reputation non-influential* and *negative reputation-influential*. This method outperformed a benchmark one-vs-one approach and reached a macro-averaged F1 score of 0.717. Since positive and negative reputation-influential information in Wikipedia can be identified with satisfiable performance, by using the above approaches, RQ4 has been answered.

The detected *positive reputation-influential* sentences and *negative reputation-influential* sentences are the sentences that Wikipedia users are generally very interested in, as they are very likely to be discussed in the Wikipedia talk pages, thus the user experience could be improved by highlighting them; alternatively, they could also help the administrators to better apply the NPOV policy of Wikipedia. Although I have limited the application scenario to *reputation-influential* sentences detection on Wikipedia, the proposed features and two-step binary classification approach for ternary classification could also be helpful for other sentence classification tasks.

This chapter has proposed to employ various features for the SVM classifier on sentence classification, which outperformed the lexicon-based approach applied in Chapter 6 in terms of the F1 score at the *sentence level*. However, it has also proved that the performance of the SVM classifier heavily relies on the feature engineering. In this chapter, all the sentences mentioned their corresponding target entity explicitly by name. In Chapter 8, I will employ attention-based neural network models to perform target-specific sentence classification, which are more complex, but of much stronger expressive powers and inference capabilities, to detect the stances in tweets, even when the targets are not mentioned. The attention-based neural network models proposed in the next chapter outperformed the SVM classifier on the target-specific stance detection task without any feature engineering involved.

Chapter 8

Attention-based Models for Target-specific Stance Detection in Tweets

In Chapter 7, I employed the *SVM classifier* to detect sentences with implicit sentiment expression and opinion implication, i.e., *reputation-influential* sentences, from *Wikipedia articles* for *persons and companies*, as well as the direction in which they influenced the reputation. For each person or company entity, I only focused on the sentences that mentioned the entity by name. In this chapter, I will focus on posts from Twitter (i.e., tweets) that carry more explicit opinion and stance expression, and are generally shorter and noisier than textual content from Wikipedia. I no longer limit the target to persons and companies; instead, it can be any object for stance expression, such as a product, a policy, an event, or a movement. In addition, the target is not necessarily explicitly mentioned in the tweet, the stance can be demonstrated by mentioning the target implicitly, or by talking about other targets. Intuitively, there is no correlation between the overall sentiment expressed in the tweet and the stance of the tweet towards the specific target. To detect the target-specific stance, the proposed model needs to infer the relationship between the given target and the topic discussed int the tweet. For above reasons, models that are of stronger expressive power and inference capability, such as attention-based neural networks, are needed to tackle the target-specific stance detection problem. This chapter answers *RQ5. Can the performance of target-specec stance detection in tweets be improved, and if so, how?* The work in this chapter has been published in [319].

8.1 Introduction

Besides real-world events reporting tweets employed in Chapter 5, there are also a large number of tweets demonstrating Internet users' stances targeting at various objects. Target-specific Stance Detection is a problem that can be formulated as follows: given a tweet X and a target Y, the aim is to classify the stance of X towards Y into three categories, *Favour*, *None* or *Against*. The target may be a person, an organisation, a government policy, a movement, a product, etc. [193]. Target-specific Stance Detection is a different problem from Aspect-level Sentiment Analysis [238, 258, 259] in the following ways: the same stance can be expressed through positive, negative or neutral sentiment [195]; the interested target of the Stance Detection does not necessarily have to occur in the tweet, as the target-specific stance can be expressed by mentioning the target implicitly, or by talking about other relevant targets.

Besides typical tweet characteristics, such as being short and noisy, the main challenge in this task is that the decision made by the classifier has to be target-specific, *whilst having very little contextual information or supervision provided.* Example training data from the benchmark target-specific Stance Detection dataset for SemEval-2016 Task 6 [193] can be found in Table 8.1.

Table 8.1: Examples of target-specific stance detection.

Target	Tweet	Stance
Donald Trump	#DonaldTrump my tell it like it is but his comments speaks to a prejudice and cold heart.	*Against*
Hillary Clinton	I love the smell of Hillary in the morning. It smells like Republican Victory.	*Against*
Hillary Clinton	Just think how many emails Hillary Clinton can delete with today's #leapsecond	*Against*
Climate Change	Coldest and wettest summer in memory.	*Favour*

Deep neural networks enable the continuous vector representations of underlying semantic and syntactic information in natural language texts, and save researchers the efforts of feature engineering [257, 258]. Recently, they have achieved significant improvements in various natural language processing tasks, such as Machine Translation [26,58], Question Answering [53,257], Sentiment Analysis [143,238,258,259,299], etc. However, applying deep neural networks on target-specific Stance Detection has not been successful, as their performance has, up to now, been slightly worse than traditional machine learning algorithms with manual feature engineering, such as Support Vector Machines (SVM) [193].

In this work, the above challenges are tackled based on the intuition that the target information is vital for the Stance Detection and that the vector representations for the tweets should be "aware" of the given targets. Since not all parts in the tweet are equally helpful for the Stance Detection task towards the specified target, I firstly apply the state-of-the-art token-level attention mechanism [26]. This allows neural networks to automatically pay more attention to the tokens that are more relevant to the target and more informative for detecting the target-specific stance. Importantly, a given token can be interpreted differently, according to different targets and the semantic features in the token's vector representation can be of different levels of importance, conditional on the given target. I propose a novel attention mechanism, which extends the current attention mechanism from the *token level* to the *semantic level* through a *gated structure*, whereby the tokens can be encoded adaptively, according to the target. I compare the models I propose based on the token-level attention mechanism and the novel semantic-level attention mechanism with several baselines on the target-specific Stance Detection dataset for the SemEval-2016 Task 6.A [193], which is currently the most widely applied dataset on target-specific Stance Detection in tweets. The experimental results show that substantial improvements can be achieved on this task, compared with all previous neural network-based models, by inferencing conditional tweet vector representations with respect to the given targets; the neural network model with semantic-level attention also outperforms the SVM algorithm, which achieved the previous best performance in this task [193]. Additionally, it should be noted that my results are obtained with a *minimum of supervision*, with *no external domain corpus collected* to pre-train target-specific word embeddings and *no extra sentiment information annotated*. Moreover, there are *no target-specific configurations or hand-engineered features involved*, thus *the proposed models can be easily generalised to other targets* with no additional efforts.

8.2 Neural Network Models for Target-specific Stance Detection in Tweets

In this section, I first describe two baseline models, the bi-directional Gated Recurrent Unit (biGRU) model and the model that stacks a Convolutional Neural Network (CNN) structure on the outputs of the biGRU (biGRU-CNN) model. I then show how I extend these two baseline models by incorporating the target information through *token-level* and *semantic-level attention mechanisms*, obtaining the AT-biGRU model and the AS-biGRU-CNN model, respectively. Finally, I demonstrate methods to generate the target embedding and how to obtain the stance detection result based on the tweet vector representation, as well as other model training details.

8.2.1 BiGRU Model

As discussed in Section 3.2.3, GRU [58] aims at solving the gradient vanishing or exploding problems, by introducing a gating mechanism. It adaptively captures dependencies in sequences, without introducing extra memory cells. GRU maps an input sequence of length N, $[x_1, x_2, \cdots, x_N]$ into a set of hidden states $[h_1, h_2, \cdots, h_N]$ as follows:

$$r_n = \sigma(W_r x_n + U_r h_{n-1} + b_r) \tag{8.1}$$

$$z_n = \sigma(W_z x_n + U_z h_{n-1} + b_z) \tag{8.2}$$

$$\tilde{h_n} = \tanh(W_h x_n + U_h(r_n \odot h_{n-1}) + b_h) \tag{8.3}$$

$$h_n = (1 - z_n) \odot h_{n-1} + z_n \odot \tilde{h_n}. \tag{8.4}$$

where $n \in \{1, \ldots, N\}$; r_n is the reset gate and z_n is the update gate; $\tilde{h_n} \in \mathbb{R}^{d_1}$ represents the "candidate" hidden state generated by the GRU; $h_n \in \mathbb{R}^{d_1}$ represents the real hidden state generated by the GRU; $x_n \in \mathbb{R}^{d_0}$ represents the word embedding vector of a token in the tweet; W_r, W_z, $W_h \in \mathbb{R}^{d_1 \times d_0}$ and U_r, U_z, $U_h \in \mathbb{R}^{d_1 \times d_1}$ represent the weight matrices; b_r, b_z, $b_h \in \mathbb{R}^{d_1}$ represent the bias terms; $\sigma(\cdot)$ represents the sigmoid function; $\odot$ represents the Hadamard product operation (element-wise multiplication).

To capture the information from both the past and the future sequence, the bi-directional GRU (biGRU), which processes the sequence in both the forward and backward directions, has proven to be successful in various applications [26, 53, 299]. In biGRU, the hidden states generated by processing the sequence in opposite directions are concatenated as the new output: $[\overrightarrow{h_1} \parallel \overleftarrow{h_1}, \overrightarrow{h_2} \parallel \overleftarrow{h_2}, \cdots, \overrightarrow{h_N} \parallel \overleftarrow{h_N}]$, where $\overrightarrow{h_n} \parallel \overleftarrow{h_n} \in \mathbb{R}^{2d_1}$ and the arrow represents the direction of the processing.

In the biGRU model, the final hidden states of the input sequence, when processing it in opposite directions, are concatenated to form the vector representation of the tweet s:

$$s = \overrightarrow{h_N} \parallel \overleftarrow{h_1}. \tag{8.5}$$

8.2.2 BiGRU-CNN Model

The biGRU model attempts to propagate all the semantic and syntactic information in a tweet into two fixed hidden state vectors, which could become a bottleneck when there exist some long-distance dependencies in the tweet. In [257], Recurrent Neural Network (RNN) outputs were fed into a CNN structure, as described in Section 3.2.3, to generate a vector representation based on all the hidden states of the RNN, rather than just the final hidden state. Specifically, a filter $w_f \in \mathbb{R}^{2kd_1}$ is applied to k concatenated consecutive hidden states $h_{i:i+k-1} \in \mathbb{R}^{2kd_1}$ to compute c_i, one value in the feature map corresponding to this filter:

$$c_i = f(w_f^T h_{i:i+k-1} + b_f) \tag{8.6}$$

where f is the rectified linear unit function and $b_f \in \mathbb{R}$ is a bias term. A max-pooling operation is further applied over the feature map $\mathbf{c} = (c_1, c_2, \cdots, c_{N-k+1})$ to capture the most important semantic feature $\hat{c}$ in each feature map:

$$\hat{c} = \max\{\mathbf{c}\}. \tag{8.7}$$

Here, $\hat{c}$ is the feature generated by filter w_f. Filters with varying sliding window size k can be applied to obtain multiple features. The features generated by different filters are concatenated to form the vector representation of the tweet s.

8.2.3 AT-biGRU Model

Whilst they solve specific problems as above, neither the biGRU model nor the biGRU-CNN model takes into account the target information. However, when human annotators are asked to label the stance of a tweet towards a given target, they are likely to keep the information about the target in their mind and pay more attention to the parts relevant to the target. The *token-level attention mechanism*, firstly proposed in [26] for Machine Translation, allowed the neural network to automatically search for tokens of a source sentence that were relevant to predicting a target word and mask irrelevant tokens; it released the burden on RNN in compressing the entire source sentence into a static, fixed representation. The attention mechanism has been successfully applied in Question Answering [53, 257], Caption Generation [293], Sentiment Analysis [299], etc.

In this chapter, I propose to apply the attention mechanism to the biGRU model, to enable the model to automatically compute proper alignments in the tweet that reflect the importance levels of different tokens in deciding the tweet's stance towards the given target, as shown in Figure 8.1.

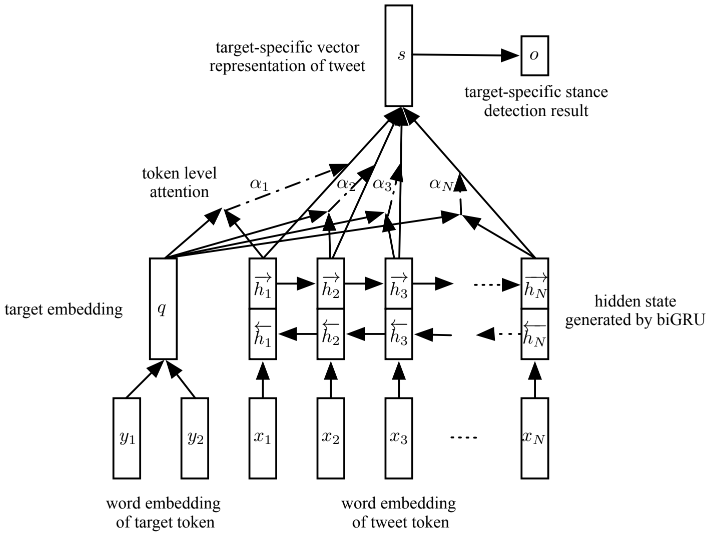

Figure 8.1: The AT-biGRU model for target-specific stance detection.

In the AT-biGRU model, the vector representation s of the tweet is calculated as the weighted sum of the hidden states:

$$s = \sum_{n=1}^{N} \alpha_n h_n. \tag{8.8}$$

In the above equation, the weight α_n of each hidden state h_n is computed by:

$$\alpha_n = \frac{\exp(e_n)}{\sum_{n=1}^{N} \exp(e_n)}, \tag{8.9}$$

where $e_n \in \mathbb{R}$ is calculated through a multi-layer perceptron (discussed in Section 3.3.3) that takes h_n and the target embedding q as input, specifically:

$$e_n = att(h_n, q) = w_m^T(\tanh(W_{ah}h_n + W_{aq}q + b_a)) + b_m. \tag{8.10}$$

where $W_{ah} \in \mathbb{R}^{2d_1 \times 2d_1}$; $W_{aq} \in \mathbb{R}^{2d_1 \times d_2}$; b_a, $w_m \in \mathbb{R}^{2d_1}$; $b_m \in \mathbb{R}$ are token-level attention parameters to optimise. In Section 8.2.5, I explore various ways to generate the target embedding $q \in \mathbb{R}^{d_2}$, based on the embeddings of the tokens in the target Y, denoted by y_1, $y_2 \in \mathbb{R}^{d_0}$. The weight α_n can be interpreted as the degree to which the model attends to token x_n in the tweet, while deciding the stance of the tweet towards the given target.

8.2.4 AS-biGRU-CNN Model

The model I propose above is an improvement on prior research. However, it can be further refined, as follows. The AT-biGRU model applies the attention mechanism at the token level, which enables the model to pay more attention to the tokens that have contributed to the stance decision towards specified targets. However, in the AT-biGRU model, the vector representations of the tokens do not have direct interaction with the vector representation of the target, which is against the intuition that the target can influence the human annotators' interpretation of each token. For example, the token "email" in Table 8.1 implies an *Against* stance towards the target "Hillary Clinton", but has no obvious influence on stances towards other targets; the token "cold" can either reveal the user's *Favour* stance towards the target "Climate Change is a Real Concern", or suggest the user's *Against* stance towards the target "Donald Trump".

Thus, I use a gated structure to extend the current token-level attention mechanism to a *more fine-grained semantic level* by introducing the direct interaction between the hidden states and the vector representation of the target. The gated structure can be embedded into the biGRU-CNN model, which results in the AS-biGRU-CNN model, as shown in Figure 8.2.

In Figure 8.2, I introduce the *target-specific hidden state* $h_n^{'}$, to replace the original hidden state h_n generated by biGRU. The target-specific hidden state is calculated as follows:

$$h_n^{'} = a_n \odot h_n. \tag{8.11}$$

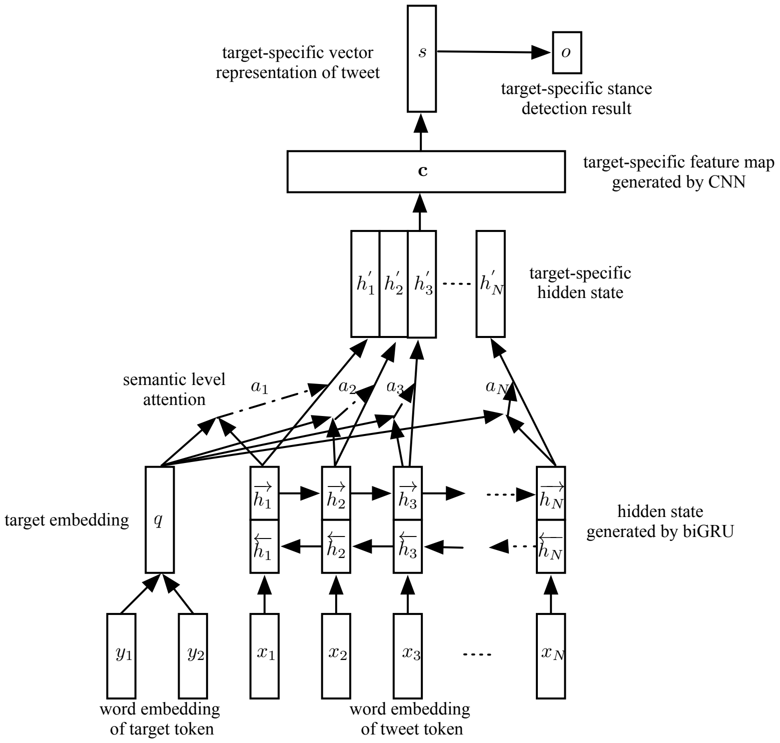
target-specific vector representation of tweet
s
o
target-specific stance detection result
c
target-specific feature map generated by CNN
h'_1 h'_2 h'_3 h'_N
target-specific hidden state
semantic level attention
a_1 a_2 a_3 a_N
target embedding
q
$\overrightarrow{h_1}$ $\overrightarrow{h_2}$ $\overrightarrow{h_3}$ $\overrightarrow{h_N}$
$\overleftarrow{h_1}$ $\overleftarrow{h_2}$ $\overleftarrow{h_3}$ $\overleftarrow{h_N}$
hidden state generated by biGRU
y_1 y_2 x_1 x_2 x_3 x_N
word embedding of target token
word embedding of tweet token

Figure 8.2: The AS-biGRU-CNN model for target-specific stance detection.

The attention vector $a_n \in \mathbb{R}^{2d_1}$ decides which semantic features in each hidden state are meaningful specifically towards the target, which is calculated through a gated structure, as follows:

$$a_n = \sigma(W_m(\tanh(W_{ah}h_n + W_{aq}q + b_a)) + b_m). \tag{8.12}$$

where $W_{ah},\ W_m \in \mathbb{R}^{2d_1 \times 2d_1}$; $W_{aq} \in \mathbb{R}^{2d_1 \times d_2}$; $b_a,\ b_m \in \mathbb{R}^{2d_1}$ are semantic-level attention parameters to optimise in the gated structure. The methods to derive the target embedding $q \in \mathbb{R}^{d_2}$ based on the embeddings of the tokens in the target Y, denoted by $y_1,\ y_2 \in \mathbb{R}^{d_0}$ will be explained in Section 8.2.5. The elements in the attention vector a_n can be understood as the degrees to which the model attends to the semantic features of token x_n in the tweet, while deciding the stance of the tweet towards the given target.

8.2.5 Target Embedding

The models proposed in Section 8.2.3 and Section 8.2.4 employ the embedding of the given target $q \in \mathbb{R}^{d_2}$, which is derived from the embeddings of the tokens in the given target y_1, $y_2 \in \mathbb{R}^{d_0}$. Without loss of generality, here I use a target with two tokens, as an example. However, the methods can be directly applied on targets with any number of tokens. To generate target embeddings of the same dimensionality for the targets with different token numbers, I propose to use a separate biGRU model, described in Section 8.2.1, with the target token embeddings y_1 and y_2 as inputs. For this scenario, the dimensionality of q, denoted by d_2 in Section 8.2.3 and Section 8.2.4, equals the dimensionality of the concatenated final hidden states of the biGRU model denoted by $2d_1$. Results of the AT-biGRU model and the AS-biGRU-CNN model using the biGRU target embedding are reported in Section 8.3.4. In some aspect-level Sentiment Analysis works, researchers have been using the average of the aspect token embeddings to encode the aspect [238, 258, 259]. I also use the averaging method as a baseline target encoding approach to derive the target embedding q by averaging the target token embeddings y_1 and y_2. For this scenario, d_2 equals to the dimensionality of the target token embeddings denoted by d_0. Results of the AT-biGRU model and the AS-biGRU-CNN model using the averaging target embedding are reported in Section 8.3.5.

8.2.6 Model Training

The vector representation of the tweet s is fed as input to a softmax layer after a linear transformation step that transforms it into a vector, whose length is equal to the number of possible stance categories. The outputs of the softmax layer o are the probabilities of the tweet X belonging to the stance category z, given the target Y denoted by $P(z|X,Y)$. The stance category with the maximum probability is selected as the *predicted category*, z^*:

$$z^* = argmax_{z \in \mathbf{z}} P(z|X,Y). \tag{8.13}$$

All the models are smooth and differentiable and they can be trained in an end-to-end manner with standard back-propagation. I use the cross-entropy loss (discussed in Section 3.3.3) as the *objective function* $L(\theta)$, which is defined as follows:

$$L(\theta) = -\sum_{X \in \mathbf{X}} \sum_{z \in \mathbf{z}} P^{'}(z|X,Y) \cdot \log(P(z|X,Y)). \tag{8.14}$$

where $\mathbf{X}$ is the set of training data; $\mathbf{z}$ is the set of stance categories; $P^{'}(z|X,Y)$ denotes the target stance distribution z given X and Y; θ is the set of parameters.

8.3 Experimental Results

8.3.1 Dataset Description

I have evaluated the effectiveness of the proposed models on the benchmark Stance Detection dataset for the SemEval-2016 Task 6.A [193], which is the most widely applied target-specific stance detection dataset for tweets by various studies [24, 82, 195,284,310]. I used the exact same data as provided to the contestants for this task, with no extra labelled data [82] or domain corpus [24,195] employed. The benchmark Stance Detection training dataset contained 2,914 tweets relevant to five targets: "Atheism" (**A**), "Climate Change is a Real Concern" (**CC**), "Feminist Movement" (**FM**), "Hillary Clinton" (**HC**) and "Legalisation of Abortion" (**LA**). Each tweet was annotated as *Favour*, *Neither* or *Against* towards one of the five targets. The benchmark Stance Detection test dataset contained 1,249 tweets, as well as the interested targets. Detailed statistics about the dataset can be found in Table 8.2, where "#" represents the number of tweets, "**%F**", "**%A**" and "**%N**" represent the percentages of tweets with *Favour*, *Against* and *Neither* stances towards the targets, respectively.

Table 8.2: Statistics of the benchmark target-specific stance detection dataset.

Target	**Training**				**Test**			
	#	**%F**	**%A**	**%N**	#	**%F**	**%A**	**%N**
A	513	17.9	59.3	22.8	220	14.5	72.7	12.7
CC	395	53.7	3.8	42.5	169	72.8	6.5	20.7
FM	664	31.6	49.4	19.0	285	20.4	64.2	15.4
HC	689	17.1	57.0	25.8	295	15.3	58.3	26.4
LA	653	18.5	54.4	27.1	280	16.4	67.5	16.1
All	2914	25.8	47.9	26.3	1249	24.3	57.3	18.4

8.3.2 Comparison Models

I compared the proposed models with the two best performing models in the SemEval-2016 Task 6.A: (1) MITRE [310], which trained separate Long Short-Term Memory (LSTM) networks with a voting scheme for different targets — the LSTM networks were pre-trained by an auxiliary hashtag prediction task on 298,973 self-collected tweets; (2) pkudblab [284], which also trained separate CNN classifiers for different targets, with a voting scheme employed both in and out of each epoch to improve the performance. I also compared against the SVM classifiers trained on the corresponding training datasets for the five targets, using word n-grams and character n-grams features, as reported in [193], representing the previous best performer for this task. Additionally, to illustrate the influence of the token-level and semantic-level attention mechanism, I included the performance comparison between the biGRU model (Section 8.2.1) and the AT-biGRU model (Section 8.2.3), the biGRU-CNN model (Section 8.2.2) and the AS-biGRU-CNN model (Section 8.2.4).

8.3.3 Experimental Settings and Model Configuration

In line with former works, I first trained separate classifiers for different targets. To obtain a fair comparison, I employed the *only* evaluation metric in the SemEval-2016 Task 6.A, which was the macro-average of the F1-score for the *Favour* and *Against* stance categories. This evaluation metric will be referred to as "macro-averaged F1 score" in this thesis for simplicity. In the evaluation stage of SemEval-2016 Task 6.A, the target information of each tweet was ignored in order to measure each team's overall performance, rather than performance on each separate target. This was because the training datasets for different targets had different percentages of tweets with Favour, Against and Neither stances, as well as different percentages of tweets expressing stances by mentioning the given target and by mentioning other targets. Thus, this evaluation metric can reflect each team's overall ability in dealing with different scenarios. It should be noted that even though separate classifiers were trained for different targets, I used the same configurations for target-specific classifiers to make sure my proposed models can be easily applied to any other target, as well as effectively demonstrate the advantages of target-specific tweet vector representation by eliminating the effects of target-specific model settings. Various methods were applied to avoid overfitting. To guarantee there were enough samples in the validation dataset, I performed a standard 5-fold cross-validation. For each round of cross-validation, I experimentally set the maximum number of epochs to 50 and located the epoch that achieved the best performance on the validation dataset. The post-softmax probabilities of the 5 trained classifiers were averaged to obtain the probabilities of a tweet in the test dataset belonging to the three stance categories.

I implemented the proposed models using the Theano library [263] and the Keras library [59].

For comparison fairness, *all the neural network-based models in the experiments used the same hyper-parameters (as illustrated below), which were selected using grid search on the baseline biGRU model.* In the experiments, all the word embeddings were initialised by the Glove [216] 100-dimensional pre-trained embeddings on Wikipedia data, i.e., $d_0 = 100$. I applied dropout [253] with probability 0.2 on the embedding layer. The word embeddings were fine-tuned during the training process to capture the stance information. From the preliminary experiments, I have observed that the models that shared the embedding layer between the tweets and the targets performed significantly better than the models that did not. I chose the dimensionality of hidden states (d_1) of both the GRU encoding the tweet and the GRU encoding the target to be 64 and the GRU weights are initialised from a uniform distribution $U(-\epsilon, \epsilon)$. Following [94], I added a dropout level of 0.3 between each recurrent connection in the GRU that encoded the tweets. I further selected the hyper-parameters for the CNN structure on top of the fixed hyper-parameters of the biGRU model. Following [143], I used filters of $k \in \{3, 4, 5\}$ with widths equal to the dimensionality of the outputs of the biGRU, which was 128 in this case. There were 100 filters for each size. To increase the robustness of the models to overfitting, a dropout level of 0.5 was further applied before the softmax layer.

I used the Adam optimiser [145] for back-propagation with the two momentum parameters set to 0.9 and 0.999, respectively. The mini-batch size was set to 16. The code for the experiments is available at `https://github.com/zhouyiwei/tsd`.

8.3.4 Using the biGRU Target Embedding

The experimental results are shown in Table 8.3. Besides the evaluation metric of SemEval-2016 Task6.A, I also provide the macro-averaged F1 scores of different targets as references. From the comparison between the biGRU model and the biGRU-CNN model, it can be seen that the CNN structure on top of the biGRU model can help to generate more compact and abstract vector representations of the tweets for Stance Detection.

Both neural network-based models that incorporate target information when generating vector representations for the tweets, i.e., AT-biGRU and AS-biGRU-CNN, outperform other neural network-based models that did not, i.e., MITRE, pkudblab, biGRU and biGRU-CNN. Specifically, the state-of-the-art token-level attention mechanism helps to increase the performance of the biGRU model by 0.32

in the overall macro-averaged F1 score. The injection of target information through the proposed semantic-level attention mechanism in the biGRU-CNN model, which results in the AS-biGRU-CNN model, leads to a more significant improvement (1.71) on the basis of the biGRU-CNN model, which makes it the best performing model among all the neural network-based models. This demonstrates the effectiveness of attention mechanisms in constructing a composite vector representation between the target and contextual information provided in the tweet. The proposed AS-biGRU-CNN model with semantic-level attention, however, has stronger capability in modelling the complex interaction between the target and each token in the tweet, and generating an expressive conditional vector representation of the tweet, with respect to the target, compared with the AT-biGRU model with the token-level attention mechanism.

Moreover, the AS-biGRU-CNN model outperforms the traditional SVM algorithm (described in Section 3.3.2), with word n-grams and character n-grams features reported in [193] by a substantial margin, in the absence of feature engineering and target-specific tuning, which justifies the motivation to automatically intensify the features that are essential to the target and "dilute" the features that are not.

Table 8.3: Performance of target-specific stance detection based on the macro-averaged F1 score, using separate classifiers.

Model	**Target**					**Overall**
	A	**CC**	**FM**	**HC**	**LA**	
SVM	65.19	42.35	57.46	58.63	66.42	68.98
MITRE	61.47	41.63	62.09	57.67	57.28	67.82
pkudblab	63.34	52.69	51.33	64.41	61.09	67.33
biGRU	65.26	43.08	56.53	55.60	61.39	67.65
biGRU-CNN	63.42	42.91	58.69	55.11	60.55	67.71
AT-biGRU	62.32	43.89	54.15	57.94	64.05	67.97
AS-biGRU-CNN	66.76	43.40	58.83	57.12	65.45	**69.42**

8.3.5 Using the Averaging Target Embedding

In Table 8.3, I used biGRU to generate the vector representations for the targets. Additionally, I further experimented with the AT-biGRU and AS-biGRU-CNN models using the averaging target embeddings. The overall macro-averaged F1 score of the AT-biGRU model increases from 67.65 to 68.30, while the macro-averaged F1 score of the AS-biGRU-CNN model decreases from 69.42 to 68.35. One possible explanation could be that a simple averaging approach is insufficient to capture the semantic meanings of the targets, thus for the biGRU-CNN model, which has stronger expressive power than the biGRU model in target-specific Stance Detection, it is helpful to use more flexible target embeddings to perform complex inference. However, for the AT-biGRU model, the target embeddings generated by biGRU surpass its capability to learn and generalise, which results in overfitting. *This is also the reason why stacking the CNN structure on top of the AT-biGRU model cannot help to improve the performance, as it does in the AS-biGRU-CNN model.*

8.3.6 Using Combined Classifiers

In the Stance Detection dataset for the SemEval-2016 Task 6.A, the training data for all the targets were of similar sizes, except for the target "Climate Change is a Real Concern". There were only 395 items in its training data and they were highly biased, with only 3.8% of them coming from the *Against* category. As a result of this, all the models in Table 8.3 cannot achieve a comparable performance on this target, when compared with other targets. When there was not enough training data for some targets, or the training data for some targets was highly biased, it was not possible to guarantee the performance of independent classifiers for these targets. For this case, I hypothesised that a combined classifier of all the targets can alleviate this problem, through jointly modelling the interaction between the stances and contexts of all the available targets. This way, when performing Stance Detection on the "Climate Change is a Real Concern" target, the classifier can employ — or even transfer — the knowledge about the intricate connection between the stances and contexts learnt from the training data of other targets. Motivated by this idea, I further trained combined classifiers based on the proposed models, using all the training data, rather than trained separate classifiers for different targets. The combined classifiers' performance is shown in Table 8.4.

Table 8.4: Performance of target-specific stance detection based on the macro-averaged F1 score, using combined classifiers.

Model	**CC**	**Overall**
SVM	47.76	62.06
biGRU	54.14	62.82
biGRU-CNN	54.57	62.70
AT-biGRU	55.69	63.36
AS-biGRU-CNN	**58.24**	**67.40**

In Table 8.4, I compare my results with the combined SVM classifier [193], which is the only result achieved through combined classifiers reported on this dataset so far. For combined classifiers, richer semantic and syntactic information was needed in the tweets' vector representations, as it was necessary to additionally encode the relatedness and diversity of different targets in stance expressions. This was a much harder task, as the combined classifier had to employ useful knowledge from other targets and avoid the impairment of useless information. For this reason, I continued to employ the biGRU model to generate the target embeddings, which had stronger expressive power than the averaging method. The difficulty level of this task is illustrated by the significantly diminished overall macro-averaged F1 score of the SVM combined classifier in Table 8.4, compared with the overall macro-averaged F1 score of the SVM separate classifiers in Table 8.3. I experimentally increased the dimensionality of the pre-trained word embedding vectors from 100 to 300, and the dimensionality of the hidden states of GRU from 64 to 256 to satisfy the above requirements. All the other hyper-parameters were kept the same, as illustrated in Section 8.3.3.

From Table 8.4, it can be observed that for the target "Climate Change is a Real Concern", it is helpful for all models to employ the training data from other targets. Comparatively, combined classifiers using models based on neural networks achieve much better macro-averaged F1 scores on this target than the combined classifiers using the traditional SVM algorithm. This is because the neural network-based models employed continuous vector representations of tweets, which allows them to more easily incorporate information from other domains, compared with the traditional SVM algorithm, which employs sparse and discrete vector representations, based on feature engineering. The combined classifier using the proposed AS-biGRU-CNN model yields the best performance so far on the "Climate Change is a Real Concern" target, which further illustrates the model's strong ability to capture the generality in stance expressions of different targets. However, the overall performance of the combined classifiers decreases. This is because the performance for targets with sufficient training data can be negatively influenced by the redundant information from other targets. Nevertheless, the AS-biGRU-CNN model still yields the best overall performance using only combined classifiers, which shows the model's power in modelling the differences in stance expressions of different targets.

8.4 Related Work

Previous research mainly focused on Stance Detection in debates [56], or in rumour spreading conversations [326]. Target-specific Stance Detection on individual tweets, however, is another challenging task, because of the irregularities in language use and the lack of contextual information. The variations in the mentions of the target, the lack of mentions of the target and the mentions of other targets clearly lead to increased difficulty. Thus, existing approaches cannot achieve satisfactory performance on the target-specific Stance Detection task.

Very few recent works have attempted to tackle the target-specific Stance Detection task on tweets [24, 82, 195, 284, 310]. [24] focused on *predicting the stances towards targets with no training data provided*, which was the SemEval-2016 Task 6.B, a different task to the one studied here. For the problem I tackled in this work, there was a training dataset for each specified target to effectively update the states and memories of the encoders. [82] was based on the correlation assumption between sentiment and stance, and it was limited by the need for sentiment labels. Thus, **the settings of both of the above works were different from the settings of the SemEval-2016 Task 6.A**. [284, 310] ignored the target information while performing classification, whereas my experiments have clearly proven that the target-specific vector representation of tweets can substantially boost the performance. [195] relied on feature engineering and a large domain corpus to perform feature selection, which was hard to generalise to other targets; and the collection of domain corpus additionally added difficulty, because of the limitations of the Twitter API. **The attention-based models proposed in this chapter, on the contrary, are fully automatic, with minimum supervision. I did not collect any extra domain corpora or use any linguistic tools and no feature engineering was needed. Since no target-specific configurations are involved, the proposed models can be directly applied to other targets.**

Another track of relevant research is aspect-level Sentiment Analysis on texts [238, 243, 258, 259, 277]. In this task, the text to be analysed, or at least parts of the text, focus on the aspects of interest, which can be easily located in the original text. This eases the problem of modelling the importance and relatedness of tokens with respect to the aspects. **This is not the case for the target-specific Stance Detection task.** Thus, a deeper integration between the target and the tweet, and a more complex inference mechanism, are needed.

8.5 Conclusion

To the best of my knowledge, I am the first one to effectively apply the traditional token-level attention mechanism to the problem of target-specific stance detection in tweets, which achieves better performance than other neural network-based models. Moreover, I have proposed to use a gated structure on the basis of the biGRU-CNN model to embed target information into the tweet's vector representation, aiming at introducing the direct semantic interaction between the target and each token in the tweet to perform *target-specific Stance Detection*. The proposed model employs a *semantic-level attention mechanism*, which is more fine-grained than the token-level attention mechanism. The proposed semantic-level attention mechanism searches for certain semantic features of each token in the tweet, based on the information contribution these semantic features have, in deciding the stance of the tweet, towards the given target. For the resulting AS-biGRU-CNN model, not only the tweet's representation vector, but also the representation vectors of the tokens are target-specific. The experimental results demonstrates that the proposed model outperforms several state-of-the-art baselines, in terms of macro-averaged F1 score, on the benchmark target-specific Stance Detection dataset of tweets, for both the scenario when separate classifiers are allowed for different targets and the scenario when only one combined classifier is allowed. Thus, the AS-biGRU-CNN model has stronger expressive power, and higher generalising capability, to extract target-specific knowledge from annotated datasets to perform target-specific stance detection in tweets. Importantly, unlike previous works on target-specific detection in tweets, the models employed in this work do not rely on any extra annotation, domain corpus or feature engineering and can be easily generalised to other targets of interest. In this way, I have answered RQ5: the performance of target-specific stance detection in tweets can be improved by incorporating the target information into the vector representations of the tweets through the proposed semantic level attention mechanism.

In this chapter, I brought together various strands of my research. I shifted the targeted social media from Wikipedia to Twitter, aiming at increasing the proposed approach's ability in processing short and noisy texts. The proposed approach in this chapter was stronger than former approaches in terms of expressive power and inference capability. It inferred the relationship between the topic discussed in the tweet and the given target, by introducing the direct interaction between the target and the tweet, which had been proven to be effective in detecting target-specific stances.

Chapter 9

Conclusion

Social media is an exciting and growing platform of our time. However, making sense of its content remains a challenge. Facing the development of social media sites, diverse information needs have been generated. For example, the development of multilingual Wikipedia opened the possibility of analysing semantic differences between different language editions when discussing certain entities, as well as the need of detecting reputation-influential sentences in Wikipedia articles; the enthusiasm in expressing personal opinions on Twitter introduced the problem of target-specific stance detection in tweets; the emergence of ambient journalism on Twitter produced the challenge of summarising fact-reporting tweets to provide the Internet users instant insights about the evolution of the events they are interested in.

In response to the above diverse information needs, I have contributed by designing and implementing automatic and effective text mining approaches to *analyse and understand the huge volume of informal texts on social media, from the topic and opinion perspectives.*

9.1 Contributions and Answers to Research Questions

Concretely, this thesis firstly presents contributions in **analysing the semantic differences between language-specific editions of Wikipedia, when discussing certain entities, from the point of view of related topical aspects** in Chapter 4 to answer RQ1:

- I have proposed a novel Graph-based approach to extract more comprehensive and accurate contexts than the baseline Article-based approach for entities from multilingual Wikipedia.

- To the best of my knowledge, I am the first one to derive language-specific topic representations for entities from their language-specific Wikipedia contexts.
- I have analysed the similarities and the differences in language-specific topic representations in a case study including 219 entities and five Wikipedia language editions, and have discovered that: the Spanish Wikipedia and Portuguese Wikipedia are most similar in their interest in topical aspects, when discussing certain entities; each entity's related topical aspects in the multilingual Wikipedia are language-specific.
- I have developed a context-based, entity-centric information retrieval model, which effectively improves the recall of entity-centric information retrieval over the baseline BM25 model, while keeping high precision, and is able to provide language-specific results.

Furthermore, I have developed an automatic approach to **generate a real-time timeline for the major event of interest, which can supplement or replace the cumbersome manually generated timeline** in Chapter 5 to answer RQ2:

- I have extracted real-world events reporting tweets from the tweet stream, employing only event-independent features; I have proposed a new variant of online incremental clustering algorithms to effectively cluster all levels of near-duplicate tweets reporting on the same sub-event; I have introduced a novel post-processing step to improve the clustering quality and efficiency of the online incremental clustering algorithm.
- I have employed an extractive summarisation algorithm to select one summary tweet from each sub-event cluster consisting of tweets reporting on the same sub-event, and have listed the sub-event summaries in chronological order to generate the real-time timeline for the major event.

I have also made the first step towards **analysing the semantic differences between language-specific editions of Wikipedia, when discussing certain entities from the aggregated sentiment perspective** in Chapter 6 to answer RQ3:

- I have proposed a framework combining the Graph-based context creation approach and a lexicon-based sentiment analysis approach to systematically quantify the variations in sentiments associated with real-world entities in different language editions of Wikipedia at the corpus level.

- I have analysed the language-specific sentiment bias for 219 entities in a case study over five Wikipedia language editions and discovered that: the proportion of objective information for any given entity is similar across language editions and constitutes about 92%; the remaining 8% contains positive and negative sentiments, that varied, dependent on the particular entity and language.

Moreover, I have moved the analysis from the corpus level to the sentence level, by proposing and tackling the problem of **detecting reputation-influential sentences with explicit or implicit sentiment expressions towards the mentioned persons or companies from Wikipedia articles** in Chapter 7 to answer RQ4:

- I have created a new dataset, which consists of Wikipedia sentences annotated by whether they have any influence on their mentioned entities' reputation, as well as the direction of the influence (positive or negative).

- I have employed various effective features with minimum domain-dependency, unlike the state-of-the-art approaches, and have applied the hierarchical classification approach to decide if a Wikipedia sentence is reputation-influential for its mentioned entity, and how the reputation of the mentioned entity would be influenced.

Finally, I have brought together various strands of my research, by **detecting target-specific stances in tweets** in Chapter 8 to answer RQ5:

- I have devised a novel AS-biGRU-CNN model, to generate a target-dependent representation for the tweet, by modelling the interaction between the tweet and the given target.

- I have proven that the proposed model with semantic-level attention mechanism is able to achieve the state-of-the-art performance on a benchmark target-specific stance detection dataset of tweets, without applying any extra annotation, domain corpus or feature engineering — whereas the current state-of-the-art approaches use one or more of these additional and, more importantly, time-consuming and expensive methods.

The relationship among different research questions has been elaborated in Section 1.3. From a technical perspective, the key term-based content representation approach employed to answer RQ1 has been adjusted to answer RQ2 by considering textual variants of key terms. The Wikipedia sentence dataset created to answer RQ1 has been further used to answer RQ3 and RQ4. The sentiment scores calculated by the lexicon-based approach to answer RQ3 have been employed to increase the proportion of reputation-influential sentences in the dataset to be annotated, when solving RQ4; these sentiment scores have also been used as features when training the classifiers to answer RQ4. The SVM classifier employed in RQ2 and RQ4 has been leveraged as a baseline approach, when solving RQ5.

9.2 Limitations and Potential Future Research Avenues

There are some issues worth further exploration. Labelled datasets were employed in sentence classification tasks in Chapter 5, Chapter 7 and Chapter 8. The performance of supervised sentence classifiers heavily relied on the labelled datasets. However, the data annotation process can be very labour-intensive and time-consuming. One possibility is to apply transfer learning techniques [64, 210], which leverage the knowledge learnt from other relevant tasks, to boost the classification performance on the new task. Another possibility is to employ semi-supervised learning techniques [191, 324], which additionally exploit the extracted knowledge from the massive unlabelled dataset. While some preliminary attempts on semi-supervised learning and transfer learning have been made in Chapter 7 and Chapter 8, more efforts are needed to improve the effectiveness. Concretely, one open question is whether the knowledge learnt from the target-specific stance detection task would be useful for the reputation-influential sentence detection task, and vice versa.

The topic representations and aggregated sentiment bias for entities in multilingual Wikipedia can change over time. In Chapter 4 and Chapter 6, only the latest version of Wikipedia when performing data collection was considered. The results can be enriched by analysing the edit history of Wikipedia articles, to explore the evolvement of topic representations and aggregated sentiment bias.

Due to the availability of suitable datasets, the timeline generation approach proposed in Chapter 8 was evaluated on the Ebola Tweets dataset only. It would be helpful to extend the experiments to datasets consisting of tweets reporting on other high-impact events.

Another problem intrinsic to neural networks is how to interpret the learnt distributed vector representations. Efforts [4,5,132] have been made recently to analyse the information encoded in the vector representations of texts through some auxiliary prediction tasks. The proposed neural network model with semantic level attention in Chapter 8 can benefit from similar analysis, by providing some insights on the encoded information before and after introducing the semantic level attention mechanism.

The approaches proposed in this thesis can be applied on other related tasks. For example, the context-based information retrieval approach can be employed to retrieve social media posts relevant to some entities or events, even though they are not mentioned by name in these posts; the timeline generation approach for high-impact events in tweets can be employed to generate timelines or biographies for person entities, based on social media data; the proposed target-specific stance detection approach can be employed to discover social media users' collective stance towards some public issues, etc.

Concluding, I can say that, via this thesis, I have made some significant contributions in the cross-disciplinary areas of topic analysis and opinion mining, opening at the same time new avenues for future researchers to further explore.

Bibliography

[1] Multilayer perceptron deeplearning 0.1 documentation. `http://deeplearning.net/tutorial/mlp.html`. (Accessed on 05/01/2017).

[2] A. Abbasi, H. Chen, and A. Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12, 2008.

[3] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *Proceedings of the 8th Extended Semantic Web Conference*, pages 375–389. Springer, 2011.

[4] Y. Adi, E. Kermany, Y. Belinkov, O. Lavi, and Y. Goldberg. Analysis of sentence embedding models using prediction tasks in natural language processing. *IBM Journal of Research and Development*, 61(4):3–1, 2017.

[5] Y. Adi, E. Kermany, Y. Belinkov, O. Lavi, and Y. Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[6] P. Agarwal, R. Vaithiyanathan, S. Sharma, and G. Shroff. Catching the long-tail: Extracting local news events from twitter. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. AAAI, 2012.

[7] C. C. Aggarwal. Mining text streams. In *Mining Text Data*, pages 297–321. Springer, 2012.

[8] C. C. Aggarwal and P. S. Yu. A framework for clustering massive text and categorical data streams. In *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 479–483. SIAM, 2006.

[9] C. C. Aggarwal and C. Zhai. *Mining text data*. Springer, 2012.

[10] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 1st International Conference on Web Search and Data Mining*, pages 183–194. ACM, 2008.

[11] K. Al Khatib, H. Schütze, and C. Kantner. Automatic detection of point of view differences in wikipedia. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 33–50. ACL, 2012.

[12] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45. ACM, 1998.

[13] L. AlSumait, D. Barbará, and C. Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 3–12. IEEE, 2008.

[14] E. Amigó, J. C. De Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Martín, E. Meij, M. De Rijke, and D. Spina. Overview of replab 2013: Evaluating online reputation monitoring systems. In *Proceedings of the 4th International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 333–352. Springer, 2013.

[15] J. An, M. Cha, P. K. Gummadi, and J. Crowcroft. Media landscape in twitter: A world of new conventions and political diversity. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. AAAI, 2011.

[16] M. Anderson, L. Carr, and D. E. Millard. There and here: Patterns of content transclusion in wikipedia. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 115–124. ACM, 2017.

[17] K. R. Apala, M. Jose, S. Motnam, C.-C. Chan, K. J. Liszka, and F. de Gregorio. Prediction of movies box office performance using social media. In *Proceedings of the 2013 International Conference on Advances in Social Networks Analysis and Mining*, pages 1209–1214. ACM, 2013.

[18] S. Aral and D. Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[19] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576. ACL, 2011.

[20] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta. Tweedr: Mining twitter to inform disaster response. In *Proceedings of the 11th International Conference on Information Systems for Crisis Response and Management*, 2014.

[21] S. Asur and B. A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 492–499. IEEE, 2010.

[22] F. Atefeh and W. Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 2013.

[23] A. Athar. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Student Session*, pages 81–87. ACL, 2011.

[24] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885. ACL, 2016.

[25] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, volume 10, pages 2200–2204. ELRA, 2010.

[26] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[27] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, pages 519–528. ACM, 2012.

[28] A. Balahur and M. Turchi. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech and Language*, 28(1):56–75, 2014.

[29] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 127–135. ACL, 2008.

[30] P. Bao, B. Hecht, S. Carton, M. Quaderi, M. Horn, and D. Gergle. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the 2012 SIGCHI Conference on Human Factors in Computing Systems*, pages 1075–1084. ACM, 2012.

[31] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. ACL, 2010.

[32] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs. In *Proceedings of the 2nd International AAAI Conference on Weblogs and Social Media*. AAAI, 2008.

[33] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 291–300. ACM, 2010.

[34] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[35] A. Ben-Hur and J. Weston. A users guide to support vector machines. *Data Mining Techniques for the Life Sciences*, pages 223–239, 2010.

[36] Y. Bengio, Y. LeCun, et al. Scaling learning algorithms towards ai. *Large-scale Kernel Machines*, 34(5):1–41, 2007.

[37] J. R. L. Bernard. *The Macquarie thesaurus : the book of words*. Dee Why, NSW : Macquarie Library Pty. Ltd, 1986.

[38] J. Bian, Y. Liu, E. Agichtein, and H. Zha. Finding the right facts in the crowd: factoid question answering over social media. In *Proceedings of the 17th International Conference on World Wide Web*, pages 467–476. ACM, 2008.

[39] C. M. Bishop. *Neural networks for pattern recognition.* Oxford university press, 1995.

[40] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[41] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

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

[43] F. Bonchi, C. Castillo, A. Gionis, and A. Jaimes. Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):22, 2011.

[44] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992.

[45] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of 19th International Conference on Computational Statistics*, pages 177–186. Springer, 2010.

[46] T. Brants, F. Chen, and A. Farahat. A system for new event detection. In *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 330–337. ACM, 2003.

[47] R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, volume 6, pages 9–16. ACL, 2006.

[48] H. Cai, Y. Yang, X. Li, and Z. Huang. What are popular: exploring twitter features for event detection, tracking and visualization. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 89–98. ACM, 2015.

[49] E. S. Callahan and S. C. Herring. Cultural bias in wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, 62(10):1899–1915, 2011.

[50] D. Chakrabarti and K. Punera. Event summarization using tweets. In *Proceedings of the 5th International Conference on Weblogs and Social Media*. AAAI, 2011.

[51] C. Chang, Y. Zhang, C. Szabo, and Q. Z. Sheng. Extreme user and political rumor detection on twitter. In *Proceedings of the 12th International Conference on Advanced Data Mining and Applications*, pages 751–763. Springer, 2016.

[52] S. Chelaru, I. S. Altingovde, and S. Siersdorfer. Analyzing the polarity of opinionated queries. In *Proceedings of the 34th European Conference on Information Retrieval Research*, pages 463–467. Springer, 2012.

[53] D. Chen, J. Bolton, and C. D. Manning. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL, 2016.

[54] D. Chen and C. D. Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, volume 1, pages 740–750. ACL, 2014.

[55] M. Chen. Efficient vector representation for documents through corruption. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[56] W. Chen and L. Ku. UTCNN: a deep learning model of stance classification on social media text. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 1635–1645. ACL, 2016.

[57] Y. Chen, H. Amiri, Z. Li, and T.-S. Chua. Emerging topic detection for organizations from microblogs. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM, 2013.

[58] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN

encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734. ACL, 2014.

[59] F. Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

[60] M. D. Choudhury, S. Counts, and M. Gamon. Not all moods are created equal! exploring human emotional states in social media. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. AAAI, 2012.

[61] P. Christen. A comparison of personal name matching: Techniques and practical issues. In *Proceedings of the ICDM 2006 Workshop*, pages 290–294. IEEE, 2006.

[62] W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string metrics for matching names and records. In *Proceedings of the KDD 2003 Workshop on Data Cleaning and Object Consolidation*, volume 3, pages 73–78, 2003.

[63] R. Collobert and S. Bengio. Links between perceptrons, mlps and svms. In *Proceedings of the 21st International Conference on Machine Learning*. ACM, 2004.

[64] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 681–691. ACL, 2017.

[65] J. M. Conroy, J. D. Schlesinger, and D. P. O'Leary. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 152–159. ACL, 2006.

[66] M. Cordeiro. Twitter event detection: Combining wavelet analysis and topic inference summarization. In *Proceedings of the 6th Doctoral Symposium on Informatics Engineering*, volume 8, pages 11–16, 2012.

[67] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[68] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716. ACL, 2007.

[69] Y. Dang, Y. Zhang, and H. Chen. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25(4):46–53, 2010.

[70] S. W. Davenport, S. M. Bergman, J. Z. Bergman, and M. E. Fearrington. Twitter versus facebook: Exploring the role of narcissism in the motives and usage of different social media platforms. *Computers in Human Behavior*, 32:212–220, 2014.

[71] J. Daxenberger and I. Gurevych. Automatically classifying edit categories in wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 578–589. ACL, 2013.

[72] J. C. De Albornoz, I. Chugur, and E. Amigó. Using an emotion-based model and sentiment analysis techniques to classify polarity for reputation. In *Proceedings of 3rd Conference and Labs of the Evaluation Forum*, 2012.

[73] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[74] K. Denecke. Using sentiwordnet for multilingual sentiment analysis. In *Proceedings of the ICDE 2008 Workshop on Data Engineering for Blogs, Social Media, and Web 2.0*, pages 507–512. IEEE, 2008.

[75] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, pages 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[76] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 536–544. ACL, 2012.

[77] Y. Ding, J. Yu, and J. Jiang. Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3436–3442. AAAI, 2017.

[78] R. Dong, M. P. O'Mahony, M. Schaal, K. McCarthy, and B. Smyth. Sentimental product recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 411–414. ACM, 2013.

[79] R. Dong, M. P. OMahony, and B. Smyth. Further experiments in opinionated product recommendation. In *Proceedings of the 22nd International Conference on Case-Based Reasoning*, pages 110–124. Springer, 2014.

[80] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Proceedings of 2012 IEEE Conference on Visual Analytics Science and Technology*, pages 93–102. IEEE, 2012.

[81] S. Dumais and H. Chen. Hierarchical classification of web content. In *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 256–263. ACM, 2000.

[82] J. Ebrahimi, D. Dou, and D. Lowd. A joint sentiment-target-stance model for stance classification in tweets. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 2656–2665. ACL, 2016.

[83] O. Egozi, S. Markovitch, and E. Gabrilovich. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, 29(2):8, 2011.

[84] J. Eisenstein. Unsupervised learning for lexicon-based classification. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3188–3194. AAAI, 2017.

[85] J. L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[86] W. Fan and M. D. Gordon. The power of social media analytics. *Communications of the ACM*, 57(6):74–81, 2014.

[87] R. Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[88] R. Feldman and J. Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[89] B. Fetahu, A. Anand, and A. Anand. How much is wikipedia lagging behind news. In *Proceedings of the 7th International ACM Web Science Conference*. ACM, 2015.

[90] E. Gabrilovich and S. Markovitch. Feature generation for text categorization using world knowledge. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, volume 5, pages 1048–1053. AAAI, 2005.

[91] E. Gabrilovich and S. Markovitch. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence and the 8th Conference on Innovative Applications of Artificial Intelligence*, volume 6, pages 1301–1306. AAAI, 2006.

[92] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, volume 7, pages 1606–1611. AAAI, 2007.

[93] E. Gabrilovich and S. Markovitch. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34:443–498, 2009.

[94] Y. Gal and Z. Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Proceedings of the Advances in Neural Information Processing Systems 29*, pages 1019–1027, 2016.

[95] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, 44(8):1761–1776, 2011.

[96] C. Gârbacea, M. Tsagkias, and M. de Rijke. Detecting the reputation polarity of microblog posts. In *Proceedings of the 21st European Conference on Artificial Intelligence*, pages 339–344. IOS Press, 2014.

[97] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In *Proceedings of the 1st International AAAI Conference on Weblogs and Social Media*. AAAI, 2007.

[98] P. Golik, P. Doetsch, and H. Ney. Cross-entropy vs. squared error training: a theoretical and experimental comparison. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, volume 13, pages 1756–1760, 2013.

[99] S. Gottschalk and E. Demidova. Analysing temporal evolution of interlingual wikipedia article pairs. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1089–1092. ACM, 2016.

[100] S. Greene and P. Resnik. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of the 2009 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 503–511. ACL, 2009.

[101] S. Greenstein and F. Zhu. Collective intelligence and neutral point of view: the case of wikipedia. Technical report, National Bureau of Economic Research, 2012.

[102] S. Greenstein and F. Zhu. Do experts or collective intelligence write with more bias? evidence from encyclopædia britannica and wikipedia. Technical report, Harvard Business School, 2014.

[103] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[104] H. Gu, X. Xie, Q. Lv, Y. Ruan, and L. Shang. Etree: Effective and efficient event modeling for real-time online social media networks. In *Proceedings of the 2011 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 300–307. IEEE, 2011.

[105] X. Guo, Y. Xiang, Q. Chen, Z. Huang, and Y. Hao. Lda-based online topic detection using tensor factorization. *Journal of Information Science*, 39(4):459–469, 2013.

[106] I. Habernal and I. Gurevych. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137. ACL, 2015.

[107] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[108] X. Han, L. Sun, and J. Zhao. Collective entity linking in web text: a graph-based method. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 765–774. ACM, 2011.

[109] X. Han and J. Zhao. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 215–224. ACM, 2009.

[110] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the 4th ACM Conference on Recommender Systems*, pages 199–206. ACM, 2010.

[111] S. Harabagiu and F. Lacatusu. Topic themes for multi-document summarization. In *Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 202–209. ACM, 2005.

[112] M. Harpalani, M. Hart, S. Singh, R. Johnson, and Y. Choi. Language of vandalism: Improving wikipedia vandalism detection via stylometric analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers*, pages 83–88. ACL, 2011.

[113] B. Hecht and D. Gergle. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the*

2010 SIGCHI Conference on Human Factors in Computing Systems, pages 291–300. ACM, 2010.

[114] R. Herbrich. *Learning kernel classifiers: theory and algorithms.* MIT Press, 2001.

[115] A. Hermida. Twittering the news: The emergence of ambient journalism. *Journalism Practice*, 4(3):297–308, 2010.

[116] G. E. Hinton. Learning distributed representations of concepts. In *Proceedings of the 8th Annual Conference of the Cognitive Science Society*, volume 1, page 12, 1986.

[117] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[118] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57. ACM, 1999.

[119] D. J. Hopkins and G. King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 2010.

[120] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[121] J. Hu, L. Fang, Y. Cao, H.-J. Zeng, H. Li, Q. Yang, and Z. Chen. Enhancing text clustering by leveraging wikipedia semantics. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 179–186. ACM, 2008.

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

[123] X. Hu and H. Liu. Text analytics in social media. In *Mining Text Data*, pages 385–414. Springer, 2012.

[124] X. Hu, N. Sun, C. Zhang, and T.-S. Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings*

of the 18th ACM Conference on Information and Knowledge Management, pages 919–928. ACM, 2009.

[125] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou. Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 389–396. ACM, 2009.

[126] L. Huang and L. Huang. Optimized event storyline generation based on mixture-event-aspect model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 726–735. ACL, 2013.

[127] L. Huang, D. Milne, E. Frank, and I. H. Witten. Learning a concept-based document similarity measure. *Journal of the American Society for Information Science and Technology*, 63(8):1593–1608, 2012.

[128] C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*. AAAI, 2014.

[129] D. Inouye and J. K. Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *Proceedings of the 3rd IEEE International Conference on Privacy, Security, Risk and Trust, and the 3rd IEEE International Conference on Social Computing*, pages 298–306. IEEE, 2011.

[130] R. Irfan, C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes, et al. A survey on text mining in social networks. *The Knowledge Engineering Review*, 30(02):157–170, 2015.

[131] M. Iyyer, V. Manjunatha, J. L. Boyd-Graber, and H. Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing*, pages 1681–1691. ACL, 2015.

[132] G. J, M. Gupta, and V. Varma. Interpretation of semantic tweet representations. In *Proceedings of the 2017 International Conference on Advances in Social Networks Analysis and Mining*, pages 95–102. ACM, 2017.

[133] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[134] A. K. Jain, J. Mao, and K. M. Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.

[135] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160. ACL, 2011.

[136] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 655–665. ACL, 2014.

[137] H. Kanayama and T. Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363. ACL, 2006.

[138] A. M. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68, 2010.

[139] R. Kaptein and J. Kamps. Exploiting the category structure of wikipedia for entity ranking. *Artificial Intelligence*, 194:111–129, 2013.

[140] S. S. Kataria, K. S. Kumar, R. R. Rastogi, P. Sen, and S. H. Sengamedu. Entity disambiguation with hierarchical topic models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1037–1045. ACM, 2011.

[141] A. Khurdiya, L. Dey, D. Mahajan, and I. Verma. Extraction and compilation of events and sub-events from twitter. In *Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, pages 504–508. IEEE, 2012.

[142] S. Kim, S. Park, S. A. Hale, S. Kim, J. Byun, and A. H. Oh. Understanding editing behaviors in multilingual wikipedia. *PloS one*, 11(5):e0155305, 2016.

[143] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751. ACL, 2014.

[144] Y. Kim, C. Denton, L. Hoang, and A. M. Rush. Structured attention networks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[145] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations: Poster Session*, 2015.

[146] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *Proceedings of the Advances in Neural Information Processing Systems 28*, pages 3294–3302, 2015.

[147] D. M. Kline and V. L. Berardi. Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Computing and Applications*, 14(4):310–318, 2005.

[148] J. Kolbitsch and H. A. Maurer. The transformation of the web: How emerging communities shape the information we consume. *J. UCS*, 12(2):187–213, 2006.

[149] R. Kosala and H. Blockeel. Web mining research: A survey. *ACM SIGKDD Explorations Newsletter*, 2(1):1–15, 2000.

[150] E. Kouloumpis, T. Wilson, and J. D. Moore. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the 5th International Conference on Weblogs and Social Media*. AAAI, 2011.

[151] P. Kranen, I. Assent, C. Baldauf, and T. Seidl. The clustree: indexing micro-clusters for anytime stream mining. *Knowledge and Information Systems*, 29(2):249–272, 2011.

[152] A. Krogh and J. A. Hertz. A simple weight decay can improve generalization. In *Proceedings of the Advances in Neural Information Processing Systems 5*, pages 950–957, 1992.

[153] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 457–466. ACM, 2009.

[154] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 297–304. ACM, 2004.

[155] S. Lacoste-Julien, F. Sha, and M. I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *Proceedings of the Advances in Neural Information Processing Systems 21*, pages 897–904, 2009.

[156] H. Lakkaraju, R. Socher, and C. Manning. Aspect specific sentiment analysis using hierarchical deep learning. In *Proceedings of the NIPS 2014 Workshop on Deep Learning and Representation Learning*, 2014.

[157] A. Lamb, M. J. Paul, and M. Dredze. Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795. ACL, 2013.

[158] P. Laufer, C. Wagner, F. Flöck, and M. Strohmaier. Mining cross-cultural relations from wikipedia: a study of 31 european food cultures. In *Proceedings of the 7th International ACM Web Science Conference*, pages 3:1–3:10. ACM, 2015.

[159] Q. V. Le et al. A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks. Technical report, Stanford University, 2015.

[160] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 14, pages 1188–1196. ACM, 2014.

[161] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[162] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[163] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural Networks: Tricks of the Trade*, pages 9–48. Springer, 2012.

[164] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary. Twitter trending topic classification. In *Proceedings of the 11th International Conference on Data Mining Workshops*, pages 251–258. IEEE, 2011.

[165] K. Lerman, S. Blair-Goldensohn, and R. McDonald. Sentiment summarization: evaluating and learning user preferences. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 514–522. ACL, 2009.

[166] C. Li, A. Sun, and A. Datta. Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 155–164. ACM, 2012.

[167] H. Li and W. Lu. Learning latent sentiment scopes for entity-level sentiment analysis. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3482–3489. AAAI, 2017.

[168] H. Li, A. Mukherjee, J. Si, and B. Liu. Extracting verb expressions implying negative opinions. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2411–2417. AAAI, 2015.

[169] H. Li and K. Yamanishi. Topic analysis using a finite mixture model. *Information Processing and Management*, 39(4):521–541, 2003.

[170] J. Li and C. Cardie. Timeline generation: Tracking individuals on twitter. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 643–652. ACM, 2014.

[171] C. Lin, C. Lin, J. Li, D. Wang, Y. Chen, and T. Li. Generating event storylines from microblogs. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 175–184. ACM, 2012.

[172] C. X. Lin, B. Zhao, Q. Mei, and J. Han. Pet: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 929–938. ACM, 2010.

[173] C.-Y. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference on Computational Linguistics*, pages 495–501. ACL, 2000.

[174] B. Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions.* Cambridge University Press, 2015.

[175] B. Liu and L. Zhang. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463. Springer, 2012.

[176] Y.-B. Liu, J.-R. Cai, J. Yin, and A. W.-C. Fu. Clustering text data streams. *Journal of Computer Science and Technology*, 23(1):112–128, 2008.

[177] R. Long, H. Wang, Y. Chen, O. Jin, and Y. Yu. Towards effective event detection, tracking and summarization on microblog data. In *Proceedings of the 12th International Conference on Web-Age Information Management*, pages 652–663. Springer, 2011.

[178] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th International Conference on World Wide Web*, pages 347–356. ACM, 2011.

[179] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th International Conference on World Wide Web*, pages 131–140. ACM, 2009.

[180] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 SIGCHI Conference on Human Factors in Computing Systems*, pages 227–236. ACM, 2011.

[181] P. Massa and F. Scrinzi. Manypedia: Comparing language points of view of wikipedia communities. In *Proceedings of the 8th Annual International Symposium on Wikis and Open Collaboration*, page 21. ACM, 2012.

[182] R. McCreadie, C. Macdonald, and I. Ounis. Incremental update summarization: Adaptive sentence selection based on prevalence and novelty. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, pages 301–310. ACM, 2014.

[183] A. J. McMinn, Y. Moshfeghi, and J. M. Jose. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pages 409–418. ACM, 2013.

[184] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[185] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

[186] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang. Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 379–387. ACM, 2012.

[187] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations: Workshop Track*, 2013.

[188] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[189] M. Miller, C. Sathi, D. Wiesenthal, J. Leskovec, and C. Potts. Sentiment flow through hyperlink networks. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. AAAI, 2011.

[190] D. N. Milne, I. H. Witten, and D. M. Nichols. A knowledge-based search engine powered by wikipedia. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pages 445–454. ACM, 2007.

[191] T. Miyato, A. M. Dai, and I. Goodfellow. Adversarial training methods for semi-supervised text classification. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[192] S. Mohammad, C. Dunne, and B. Dorr. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 599–608. ACL, 2009.

[193] S. M. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, volume 16. ACL, 2016.

[194] S. M. 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*, 2013.

[195] S. M. Mohammad, P. Sobhani, and S. Kiritchenko. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):26, 2017.

[196] T. S. Motwani and R. J. Mooney. Improving video activity recognition using object recognition and text mining. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 600–605. IOS Press, 2012.

[197] A. Mukherjee and B. Liu. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 339–348. ACL, 2012.

[198] T. Mullen and N. Collier. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, volume 4, pages 412–418. ACL, 2004.

[199] C. Müller and I. Gurevych. Using wikipedia and wiktionary in domain-specific information retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 219–226. Springer, 2009.

[200] N. Nagwani. Summarizing large text collection using topic modeling and clustering based on mapreduce framework. *Journal of Big Data*, 2(1):1–18, 2015.

[201] K. Nemoto and P. A. Gloor. Analyzing cultural differences in collaborative innovation networks by analyzing editing behavior in different-language wikipedias. *Procedia-Social and Behavioral Sciences*, 26:180–190, 2011.

[202] A. Ng. Cs229 lecture notes, part v: Support vector machines. Technical report, Stanford University, 2012.

[203] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, Q. V. Le, and A. Y. Ng. On optimization methods for deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 265–272. ACM, 2011.

[204] B. Ni, Z. Song, and S. Yan. Web image mining towards universal age estimator. In *Proceedings of the 17th ACM International Conference on Multimedia*, pages 85–94. ACM, 2009.

[205] F. Å. Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC 2011 Workshop on Making Sense of Microposts: Big Things Come in Small Packages*, CEUR Workshop Proceedings, pages 93–98, 2011.

[206] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media.* AAAI, 2010.

[207] A. Olteanu, S. Vieweg, and C. Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 994–1009. ACM, 2015.

[208] A. Otegi, X. Arregi, O. Ansa, and E. Agirre. Using knowledge-based relatedness for information retrieval. *Knowledge and Information Systems*, pages 1–30, 2014.

[209] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* ACL, 2013.

[210] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[211] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd*

Annual Meeting on Association for Computational Linguistics, pages 271–278. ACL, 2004.

[212] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86. ACL, 2002.

[213] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1310–1318, 2013.

[214] M.-H. Peetz, M. de Rijke, and R. Kaptein. Estimating reputation polarity on microblog posts. *Information Processing and Management*, 52(2):193–216, 2016.

[215] F. Peleja, J. Santos, and J. Magalhães. Reputation analysis with a ranked sentiment-lexicon. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1207–1210. ACM, 2014.

[216] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, volume 14, pages 1532–1543. ACL, 2014.

[217] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 181–189. ACL, 2010.

[218] U. Pfeil, P. Zaphiris, and C. S. Ang. Cultural differences in collaborative authoring of wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113, 2006.

[219] S. Phuvipadawat and T. Murata. Breaking news detection and tracking in twitter. In *Proceedings of the 2010 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 120–123. IEEE, 2010.

[220] J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Microsoft Research, 1998.

[221] D. Ploch. Exploring entity relations for named entity disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Student Session*, pages 18–23. ACL, 2011.

[222] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016.

[223] M. Potthast, B. Stein, and M. Anderka. A wikipedia-based multilingual retrieval model. In *Proceedings of the 30th European Conference on Information Retrieval*, pages 522–530. Springer, 2008.

[224] L. Prechelt. Early stoppingbut when? In *Neural Networks: Tricks of the Trade*, pages 53–67. Springer, 2012.

[225] D. Quercia, H. Askham, and J. Crowcroft. Tweetlda: supervised topic classification and link prediction in twitter. In *Proceedings of the 4th International ACM Web Science Conference*, pages 247–250. ACM, 2012.

[226] K. Rajaraman and A.-H. Tan. Topic detection, tracking, and trend analysis using self-organizing neural networks. In *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 102–107. Springer, 2001.

[227] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256. ACL, 2009.

[228] A. Ramesh, S. H. Kumar, J. Foulds, and L. Getoor. Weakly supervised models of aspect-sentiment for online course discussion forums. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing*, pages 74–83. ACL, 2015.

[229] T. Rao and S. Srivastava. Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining*, pages 119–123. ACM, 2012.

[230] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1650–1659. ACL, 2013.

[231] R. Rehurek and P. Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, 2010.

[232] Z. Ren, M.-H. Peetz, S. Liang, W. Van Dolen, and M. De Rijke. Hierarchical multi-label classification of social text streams. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 213–222. ACM, 2014.

[233] A. Ritter, O. Etzioni, S. Clark, et al. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1104–1112. ACM, 2012.

[234] S. Robertson, H. Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[235] R. Rogers. Wikipedia as cultural reference. *Digital Methods*, 2013.

[236] J. Rogstadius, M. Vukovic, C. Teixeira, V. Kostakos, E. Karapanos, and J. Laredo. Crisistracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development*, 57(5):1–4, 2013.

[237] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

[238] S. Ruder, P. Ghaffari, and J. G. Breslin. A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 999–1005. ACL, 2016.

[239] M. E. Ruiz and P. Srinivasan. Hierarchical neural networks for text categorization (poster abstract). In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 281–282. ACM, 1999.

[240] G. Salton and M. J. McGill. Introduction to modern information retrieval. 1986.

[241] G. E. Sandra, M. P. OMahony, and B. Smyth. Towards tagging and categorization for micro-blogs. In *Proceedings of the 21st National Conference on Artificial Intelligence and Cognitive Science*, 2010.

[242] P. Schönhofen, A. Benczúr, I. Bíró, and K. Csalogány. Cross-language retrieval with wikipedia. In *Proceedings of the 8th Workshop of the Cross-Language Evaluation Forum*, pages 72–79. Springer, 2008.

[243] K. Schouten, F. Baas, O. Bus, A. Osinga, N. van de Ven, S. van Loenhout, L. Vrolijk, and F. Frasincar. Aspect-based sentiment analysis using lexico-semantic patterns. In *Proceedings of the 17th International Conference on Web Information Systems Engineering*, pages 35–42. Springer, 2016.

[244] A. Schulz, B. Schmidt, and T. Strufe. Small-scale incident detection based on microposts. In *Proceedings of the 26th ACM Conference on Hypertext and Social Media*, pages 3–12. ACM, 2015.

[245] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th International Conference on Machine learning*, pages 807–814. ACM, 2007.

[246] L. Shou, Z. Wang, K. Chen, and G. Chen. Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 533–542. ACM, 2013.

[247] S. Siersdorfer, E. Minack, F. Deng, and J. Hare. Analyzing and predicting sentiment of images on the social web. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 715–718. ACM, 2010.

[248] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467, 2011.

[249] C. N. Silla Jr and A. A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, 2011.

[250] N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 208–215. ACM, 2000.

[251] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, volume 1631, pages 1631–1642. ACL, 2013.

[252] P. Sorg and P. Cimiano. Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data and Knowledge Engineering*, 74:26–45, 2012.

[253] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[254] N. Stokes and J. Carthy. First story detection using a composite document representation. In *Proceedings of the 1st International Conference on Human Language Technology Research*, pages 1–8. ACL, 2001.

[255] A. Sun and E.-P. Lim. Hierarchical text classification and evaluation. In *Proceedings of the 1st International Conference on Data Mining*, pages 521–528. IEEE, 2001.

[256] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011.

[257] M. Tan, B. Xiang, and B. Zhou. Lstm-based deep learning models for non-factoid answer selection. In *Proceedings of the 4th International Conference on Learning Representations: Workshop Track*, 2016.

[258] D. Tang, B. Qin, X. Feng, and T. Liu. Effective lstms for target-dependent sentiment classification. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 3298–3307. ACL, 2016.

[259] D. Tang, B. Qin, and T. Liu. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224. ACL, 2016.

[260] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1555–1565. ACL, 2014.

[261] Y. Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

[262] K. Tao, F. Abel, C. Hauff, G.-J. Houben, and U. Gadiraju. Groundhog day: near-duplicate detection on twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1273–1284. ACM, 2013.

[263] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, 2016.

[264] R. Tinati, L. Carr, W. Hall, and J. Bentwood. Identifying communicator roles in twitter. In *Proceedings of the 21st International Conference on World Wide Web*, pages 1161–1168. ACM, 2012.

[265] R. Tinati, S. Halford, L. Carr, and C. Pope. Big data: methodological challenges and approaches for sociological analysis. *Sociology*, 48(4):663–681, 2014.

[266] R. Tinati, T. Tiropanis, and L. Carr. An approach for using wikipedia to measure the flow of trends across countries. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1373–1378. ACM, 2013.

[267] I. Titov and R. T. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, volume 8, pages 308–316. ACL, 2008.

[268] C. Toprak, N. Jakob, and I. Gurevych. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584. ACL, 2010.

[269] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology*, pages 173–180. ACL, 2003.

[270] R. Townsend, A. Tsakalidis, Y. Zhou, B. Wang, M. Liakata, A. Zubiaga, A. I. Cristea, and R. Procter. Warwickdcs: From phrase-based to target-specific sentiment recognition. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 657–663. ACL, 2015.

[271] T. A. Tran, C. Niederée, N. Kanhabua, U. Gadiraju, and A. Anand. Balancing novelty and salience: Adaptive learning to rank entities for timeline summarization of high-impact events. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 1201–1210. ACM, 2015.

[272] B. Tsolmon and K.-S. Lee. An event extraction model based on timeline and user analysis in latent dirichlet allocation. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1187–1190. ACM, 2014.

[273] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*. AAAI, 2010.

[274] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424. ACL, 2002.

[275] R. Van Zwol. Flickr: Who is looking? In *Proceedings of the 2007 IEEE/WIC/ACM international Conference on Web Intelligence*, pages 184–190. IEEE, 2007.

[276] X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference of the Asian Federation of Natural Language Processing*, pages 235–243. ACL, 2009.

[277] B. Wang, M. Liakata, A. Zubiaga, and R. Procter. Tdparse-multi-target-specific sentiment recognition on twitter. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.* ACL, 2017.

[278] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 115–120. ACL, 2012.

[279] P. Wang and C. Domeniconi. Building semantic kernels for text classification using wikipedia. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 713–721. ACM, 2008.

[280] P. Wang, J. Hu, H.-J. Zeng, and Z. Chen. Using wikipedia knowledge to improve text classification. *Knowledge and Information Systems*, 19(3):265–281, 2009.

[281] S. Wang and C. D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 90–94. ACL, 2012.

[282] Z. Wang, L. Shou, K. Chen, G. Chen, and S. Mehrotra. On summarization and timeline generation for evolutionary tweet streams. *IEEE Transactions on Knowledge and Data Engineering*, 27(5):1301–1315, 2015.

[283] K. Watanabe, M. Ochi, M. Okabe, and R. Onai. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 2541–2544. ACM, 2011.

[284] W. Wei, X. Zhang, X. Liu, W. Chen, and T. Wang. pkudblab at semeval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 384–388. ACL, 2016.

[285] Z. Wei and W. Gao. Gibberish, assistant, or master?: Using tweets linking to news for extractive single-document summarization. In *Proceedings of the*

38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1003–1006. ACM, 2015.

[286] J. Weng and B.-S. Lee. Event detection in twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. AAAI, 2011.

[287] D. Williams and G. Hinton. Learning representations by back-propagating errors. *Nature*, 323(6088):533–538, 1986.

[288] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the 2005 Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. ACL, 2005.

[289] I. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25–30. AAAI, 2008.

[290] P. Wu, S. C.-H. Hoi, P. Zhao, and Y. He. Mining social images with distance metric learning for automated image tagging. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 197–206. ACM, 2011.

[291] E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. ACM, 2017.

[292] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1980–1984. ACM, 2012.

[293] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2048–2057. ACM, 2015.

[294] D. Yan, L. Huang, and M. I. Jordan. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 907–916. ACM, 2009.

[295] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 745–754. ACM, 2011.

[296] D. Yang, D. Zhang, Z. Yu, and Z. Wang. A sentiment-enhanced personalized location recommendation system. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 119–128. ACM, 2013.

[297] J. Yang, J. McAuley, and J. Leskovec. Community detection in networks with node attributes. In *Proceedings of the 13th International Conference on Data Mining*, pages 1151–1156. IEEE, 2013.

[298] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 28–36. ACM, 1998.

[299] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. ACL, 2016.

[300] T. Yano, P. Resnik, and N. A. Smith. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 152–158. ACL, 2010.

[301] J.-g. Yao, F. Fan, W. X. Zhao, X. Wan, E. Chang, and J. Xiao. Tweet timeline generation with determinantal point processes. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 3080–3086. AAAI, 2016.

[302] T. Yasseri, R. Sumi, and J. Kertész. Circadian patterns of wikipedia editorial activity: A demographic analysis. *PloS one*, 7(1):e30091, 2012.

[303] M. Yazdani and A. Popescu-Belis. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 3185–3189. AAAI, 2013.

[304] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, (6):52–59, 2012.

[305] W. Yin, K. Kann, M. Yu, and H. Schütze. Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923*, 2017.

[306] E. Yom-Tov. Ebola data from the internet: An opportunity for syndromic surveillance or a news event? In *Proceedings of the 5th International Conference on Digital Health*, pages 115–119. ACM, 2015.

[307] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua. Aspect ranking: identifying important product aspects from online consumer reviews. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1496–1505. ACL, 2011.

[308] J. Yun, L. Jing, J. Yu, and H. Huang. A multi-layer text classification framework based on two-level representation model. *Expert Systems with Applications*, 39(2):2035–2046, 2012.

[309] R. Zafarani, M. A. Abbasi, and H. Liu. *Social media mining: an introduction.* Cambridge University Press, 2014.

[310] G. Zarrella and A. Marsh. MITRE at semeval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 458–463. ACL, 2016.

[311] C. Zhang, G. Zhou, Q. Yuan, H. Zhuang, Y. Zheng, L. Kaplan, S. Wang, and J. Han. Geoburst: Real-time local event detection in geo-tagged tweet streams. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 513–522. ACM, 2016.

[312] L. Zhang and B. Liu. Identifying noun product features that imply opinions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 575–580. ACL, 2011.

[313] W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, and X. Li. Topical keyphrase extraction from twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 379–388. ACL, 2011.

[314] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Information Retrieval*, pages 338–349. Springer, 2011.

[315] Z. Zhao, P. Resnick, and Q. Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1395–1405. ACM, 2015.

[316] M. Zhou. *Entity-centric search: querying by entities and for entities.* University of Illinois at Urbana-Champaign, 2014.

[317] Y. Zhou and A. I. Cristea. Towards detection of influential sentences affecting reputation in wikipedia. In *Proceedings of the 8th International ACM Web Science Conference*, pages 244–248. ACM, 2016.

[318] Y. Zhou, A. I. Cristea, and Z. Roberts. Is wikipedia really neutral? A sentiment perspective study of war-related wikipedia articles since 1945. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation.* ACL, 2015.

[319] Y. Zhou, A. I. Cristea, and L. Shi. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *Proceedings of the 18th International Conference on Web Information Systems Engineering*, pages 18–32. Springer, 2017.

[320] Y. Zhou, E. Demidova, and A. I. Cristea. Analysing entity context in multilingual wikipedia to support entity-centric retrieval applications. In *Proceedings of the 1st International KEYSTONE (semantic keyword-based search on structured data sources) Conference*, pages 197–208. Springer, 2015.

[321] Y. Zhou, E. Demidova, and A. I. Cristea. Who likes me more?: analysing entity-centric language-specific bias in multilingual wikipedia. In *Proceedings of the 31st ACM Symposium on Applied Computing*, pages 750–757. ACM, 2016.

[322] Y. Zhou, E. Demidova, and A. I. Cristea. What's new? analysing language-specific wikipedia entity contexts to support entity-centric news retrieval. *Transactions on Computational Collective Intelligence*, 26:210–231, 2017.

[323] Y. Zhou, N. Kanhabua, and A. I. Cristea. Real-time timeline summarisation for high-impact events in twitter. In *Proceedings of the 22nd European Conference on Artificial Intelligence*, volume 285, pages 1158–1166. IOS Press, 2016.

[324] X. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison, 2006.

[325] M. Zimmermann, I. Ntoutsi, Z. F. Siddiqui, M. Spiliopoulou, and H.-P. Kriegel. Discovering global and local bursts in a stream of news. In *Proceedings of the 27th ACM Symposium on Applied Computing*, pages 807–812. ACM, 2012.

[326] A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, and M. Lukasik. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 2438–2448. ACL, 2016.

[327] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, and P. Tolmie. Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 347–353. ACM, 2015.

[328] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, and P. Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989, 2016.

[329] A. Zubiaga, D. Spina, E. Amigó, and J. Gonzalo. Towards real-time summarization of scheduled events from twitter streams. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, pages 319–320. ACM, 2012.