UNIVERSITÄT KOBLENZ · LANDAU
Faculty 4: Computer Science

WeST People and Knowledge Networks
Institute for Web Science and Technologies

Analysing Sentiments on Wikipedia Concepts with varying Time and Geolocation Attributes

Master's Thesis

in partial fulfillment of the requirements for
the degree of Master of Science (M.Sc.)
in Informatik

submitted by
Qianhong Ye

First supervisor:	JProf. Dr. Claudia Wagner Institute for Web Science and Technologies
Second supervisor:	Dr. Fabian Flöck Institute for Web Science and Technologies

Koblenz, December 2018

Statement

I hereby certify that this thesis has been composed by me and is based on my own work, that I did not use any further resources than specified – in particular no references unmentioned in the reference section – and that I did not submit this thesis to another examination before. The paper submission is identical to the submitted electronic version.

	Yes	No
I agree to have this thesis published in the library.	☐	☐
I agree to have this thesis published on the Web.	☐	☐
The thesis text is available under a Creative Commons License (CC BY-SA 4.0).	☐	☐
The source code is available under a GNU General Public License (GPLv3).	☐	☐
The collected data is available under a Creative Commons License (CC BY-SA 4.0).	☐	☐

.....

(Place, Date) (Signature)

Note

- If you would like us to contact you for the graduation ceremony, please provide your personal E-mail address: .....
- If you would like us to send you an invite to join the WeST Alumni and Members group on LinkedIn, please provide your LinkedIn ID : .....

Abstract

The purpose of this thesis is to explore the sentiment distributions of Wikipedia concepts. We analyse the sentiment of the entire English Wikipedia corpus, which includes 5,669,867 articles and 1,906,375 talks, by using a lexicon-based method with four different lexicons. Also, we explore the sentiment distributions from a time perspective using the sentiment scores obtained from our selected corpus. The results obtained have been compared not only between articles and talks but also among four lexicons: OL, MPQA, LIWC, and ANEW. Our findings show that among the four lexicons, MPQA has the highest sensitivity and ANEW has the lowest sensitivity to emotional expressions. Wikipedia articles show more sentiments than talks according to OL, MPQA, and LIWC, whereas Wikipedia talks show more sentiments than articles according to ANEW. Besides, the sentiment has a trend regarding time series, and each lexicon has its own bias regarding text describing different things. Moreover, our research provides three interactive widgets for visualising sentiment distributions for Wikipedia concepts regarding the time and geolocation attributes of concepts.

Acknowledgement

I would like to thank, first and foremost, my supervisors Dr. Fabian Flöck and JProf. Dr. Claudia Wagner for their professional guidance during the whole process of my research. Without their feedback and support, the thesis would never be accomplished. I am particularly grateful to Dr. Fabian Flöck for providing the core idea of visualisation widgets and multiple analysing approaches.

Next, I would like to thank Dr. Christoph Carl Kling for providing this interesting research topic and support during the preliminary research. I would also like to thank Daniel Janke for his assistance regarding Virtual Machine server during the process of data collection and data pre-processing.

Furthermore, I would like to thank Min Ke for her valuable suggestions on academic writing and continued encouragement during my research.

Special thanks to Ding-Ngim Ju Sing, Mujtaba Rafiq, and Shreya Chatterjee for proofreading this thesis.

Finally, I wish to thank my family and friends for their endless support and encouragement throughout my study.

Contents

1. Introduction

1.1. Background and Problem Statement

Wikipedia is a multilingual, free-content encyclopedia project built in 2001[1]. It is the largest and most widely used encyclopedia in the world [MMLW09]. According to the statistics of Wikipedia, there are more than 48 million articles in 302 languages and more than 5 million of them are English articles[2]. More and more researchers are regarding Wikipedia as a goldmine of information and want to apply the concepts and relationships inside to a host of tasks [MMLW09].

Wikipedia is supposed to be neutral according to the *Neutral Point of View* (NPOV) policy[3], but some studies and findings are indicating that it is not true. It is important to note that, NPOV policy does not mean the exclusion of opinions, instead, it encourages editors to include complete verifiable points of view. Zhou et al. [ZCR15] made a sentiment analysis on multilingual Wikipedia articles toward war-related topics and their empirical results proved that articles from different language background hold different emotions or different extent of emotions. Greenstein and Zhu [GZ12] studied all articles relating to American politics and collected their political bias. They summarised a conclusion that "Wikipedia contains a bias, and the level or direction of bias is not fixed over time".

However, these researches focus more on the bias of the opinion, besides, they only use specific topics to perform the analysis. There has been little research conducted on questions such as "To what extent the English Wikipedia contains emotional language or opinions?".

Moreover, there are a handful of studies providing intuitive and effective visualisation tools for visualising the distribution of sentimental expressions among the Wikipedia entities. Wikipedia contains article pages and talk pages[4] for each concept. An article page is used to describe a concept and the corresponding talk page is used for users to discuss the content of the article. In this thesis, the term *articles* indicates Wikipedia article pages, and the term *talks* indicates Wikipedia talk pages. The terms *concepts* and *entities* will be used interchangeably to mean the entries of Wikipedia articles or Wikipedia talks. According to the definition provided by Pang and Lee [PL+08], the terms *sentiment analysis* and *opinion mining* are used interchangeably to mean the computational treatment of sentiment.

[1] `https://en.wikipedia.org/wiki/Wikipedia`, as seen on Nov. 8, 2018

[2] `https://en.wikipedia.org/wiki/Wikipedia:About`, as seen on Nov. 8, 2018

[3] `https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view`, as seen on Nov. 8, 2018

[4] `https://en.wikipedia.org/wiki/Help:Talk_pages`, as seen on Nov. 8, 2018

So far, there are millions of concepts on Wikipedia. However, most existing sentiment visualisation systems are designed to visualise people's opinion about products from online reviews or opinions about events in social media [Bou15]. These tools are only used to analyse a very narrow spectrum of topics. Tree Map, River Flow, Bar Chart, et al. are the most commonly used visual technologies. Therefore, the existing sentiment visualisation systems will not be able to apply the numerous Wikipedia concepts directly.

Furthermore, attributes for Wikipedia concepts, such as category, geolocation, and related dates, could be used as assistance to better visualise the sentiment distribution. Since such structured data has been collected by a community known as DBpedia [BLK+09, LIJ+15], it will be easy to obtain the required attributes via DBpedia.

1.2. Research Objectives and Contributions

The major objective of this study is to investigate the sentiment distributions of Wikipedia corpus, which includes 5,669,867 articles and 1,906,375 talks. By comparing the sentiments calculated with four different lexicons, which are OL, MPQA, LIWC, and ANEW, we focus on the respective characteristics of each of the lexicons. Besides, we aim to design visualisation tools to display the sentiment distributions of Wikipedia concepts from the perspective of time and geography.

In order to make the codes of visualisation systems editable for users, we use Jupyter Notebook as our programming tool. Most existing visualisation systems are presented in packaged application software, which means the functions and usage are predefined and unchangeable. In this case, the internal theories and execution processes are invisible to users, and users are unable to do any modification to the system to meet their own requirements. To break such limitations and facilitate scientific research, the interactive visualisation systems we designed in this thesis allow users not only to set parameters through the graphical user interface, but also to edit source code directly so that the system can be modified according to the specific needs. Jupyter Notebook has the ability to achieve this purpose as it is an open-source software with interactive computing capabilities[5]. The widgets we designed could aid the analysis of data scientist, either to find out untrivial patterns from Wikipedia or help Wikis to locate articles with extreme sentiments.

Our contributions can be summarised as follows:

1. We calculate sentiment scores for entire English Wikipedia articles and talks.
2. We compare sentiment distributions for Wikipedia concepts based on various lexicons, including OL, MPQA, LIWC, and ANEW, to see the sensitivity to emotions of each of these lexicons.

[5] `http://jupyter.org`, as seen on Aug. 14, 2018

3. We present the changes of sentiments over time regarding Wikipedia concepts about people and events.

4. We provide three interactive widgets for visualising the sentiment distributions of Wikipedia concepts with varying time and geolocation attributes.

1.3. Thesis Structure

The rest of the thesis is organised as follows. Chapter 2 introduces the related work in the field of sentiment analysis on Wikipedia as well as sentiment visualisation techniques. Chapter 3 provides an overview of the methodology of sentiment analysis in this thesis. Chapter 4 explains the process of data collection and pre-processing we carried out. Chapter 5 describes the approach to document-level sentiment analysis in terms of articles and talks separately. Chapter 6 illustrates the analysis we did on the sentiment distributions of Wikipedia entities. Chapter 7 presents three interactive visualisation widgets we designed for Wikipedia entities. The last chapter concludes the research work in this thesis, and summarises the limitations of our work, as well as the future directions.

2. Related Work

In this chapter, we will discuss some of the related work in the area of sentiment analysis on Wikipedia, as well as the recent sentiment visualisation techniques.

2.1. Sentiment Analysis on Wikipedia

In order to study whether Wikipedia is neutral, Zhou et al. [ZCR15] propose an approach based on article-level and concept-level sentiment analysis on multilingual Wikipedia articles. They take war-related topics as examples, and employ lexicon-based sentiment analysis with subjectivity analysis on the extracted simple plain descriptive text. The analysis of multilingual text has been implemented by translating the other languages into English firstly. Their empirical results reveal that articles from different language backgrounds focus on different concepts regarding the same topic, and express different sentiments toward these concepts. Regarding the same question, Greenstein and Zhu [GZ12] investigate the bias between Democrat and Republican by analysing the Wikipedia articles which are related to American politics. They however have not conducted sentiment analysis, but measured the bias by constructing slant indexes[1]. Their findings show that many articles contain bias, and this bias evolves over time.

Nielsen et al. [NEH13] present an online server to monitor the sentiment of Wikipedia articles which are describing companies. The sentiment analysis has been carried out with a lexicon-based method with AFINN word list [Nie11]. Moreover, it focuses on real-time edits, which means the sentiment of a new revision of article will be compared with the previous revision, and the information of edit, such as timestamp and editor, is taken into consideration. To visualise the results, the relative sentiment of each edit for the specific company has been plotted with a weekly time axis. Also, a sequential collaboration network (SCN) [INPG10] is generated to show the sentiment of change made by each user, where nodes indicate users, and links represent the sequence of edits. The edit of inserting positive text or removing negative text connects to positive sentiment, while an opposite of it connects to negative sentiment. It states that the accuracy of this system can be increased by refining the pre-process of text, such as handling upper or lower case, or dealing with templates of Wikipedia text, which has been considered in this thesis. Similarly, Chandy [Cha08] design a system, named Wikiganda, with a lexicon-based approach, which

[1] Slant index measures the bias between Democrat and Republican based on the slant of phrases. It has been constructed with a method developed by Gentzkow and Shapiro [GS06].

aims to detect potential malicious propaganda on Wikipedia by analysing the sentiment of the edits. These systems are designed to detect articles against the *Conflict of Interest Policy* (COI) policy[2].

Some researches place great emphasis on Wikipedia talk pages. Grigore and Rosenkranz [GR11] analyse the level of sentiments on Wikipedia talk pages by using the SentiStrength tool [MKGD10, TBP11]. They aim to examine the relation between the sentiment expressed on the talk page and the degree of trust for editors according to the collaborative article page. The trust between editors is measured by the number of reused words between editors, and it can be examined through the edit logs of the article page. Their finding shows that the sentiments on talk pages do affect the collaborative work on article pages. Laniado et al. [LKCM12] measure the emotional content on Wikipedia talk pages with *Affective Norms for English Words* (ANEW) [BL99] from three dimensions: valence, arousal, and dominance. Valence measures emotions such as happiness or sadness, arousal represents the emotions by excitement, and dominance focuses on feelings of being in control or not. Their analysis reveals how the emotions on talk pages related to the profiles of editors such as gender and experience. In the same vein, Iosub et al. [ILC+14] explore the relation between the emotion and the profile, as well as the relation between the emotion and the response of editor, by using lexicons LIWC and SentiStrength. These researches aim to reveal novel insights of online collaboration by using Wikipedia talks as the study case.

In addition to Wikipedia, sentiment analysis is widely used to social media and review. For example, Bautin et al. [BVS08] implement the sentiment analysis on news and blogs with multi-language. They visualise the result by using an international sentiment map, which is similar to the output of one of our widgets (WikiSentiViewer). Singh et al. [SPUW13] employ an aspect-level sentiment analysis by using the machine learning technique to classify movie review. Differ from the analysis described above, aspect-level sentiment analysis examine the sentiment for each feature (e.g. dialogue, script, and music) separately. Mudinas [MZL12] propose a concept-level approach to sentiment analysis by combining lexicon-based and learning-based method, and apply it in software review and movie review. Reagan et al. [RTW+15] focus on different dictionary-based approaches and compare the attributes and words of each dictionary. Further, Liu [Liu12] and Pang and Lee [PL+08] present a detailed description of various sentiment analysis as well as the applications on multiple fields.

2.2. Sentiment Visualisation Techniques

With respect to sentiment visualisation techniques, Boumaiza [Bou15] presents a concise review of recent approaches, including topic-based method and feature-based method, to visualise people's opinion toward products or topics. Topic-based

[2] `https://en.wikipedia.org/wiki/Wikipedia:Conflict_of_interest`, as seen on Nov. 8, 2018

methods aim to extract the topics or events from text corpora, while feature-based methods aim to describe sentiment for extracted features. This survey reveals that all techniques have their own objectives and drawbacks, and there is no technique outperforming the others. Moreover, it suggests that a combination of different visualisation techniques will overcome their individual drawbacks and therefore enhance the final performance.

Wu et al. [WLY+14], and Havre et al. [HHN00] use a river (or flow) metaphor to represent the changes of opinions or topics over time. Wu et al. [WLY+14] propose a visual analysis system, named OpinionFlow, to trace and analyse opinion diffusion among a large number of people based on the dataset of Twitter. The interface contains three parts. On the left side, it consists of a stacked tree which is the hierarchical structure of the topics. In the center, an opinion flow combining a Sankey graph[3] [RHF05] and a density map has been presented. On the right side, it explains the detailed information of tweets. This system is able to visualise the opinion flow with a multi-scale timeline. It uses the color to indicate the polarity of opinion and the density to indicate the strength of the opinion. It is designed for a small number of topics by labeling each of the topics on the flow. Havre et al. [HHN00] provide a prototype system named ThemeRiver, which is used to visualise the variations of themes over time within a large collection of documents by a river plot. The width of the river at a specific point of time indicates the number of documents related to this theme at that time. ThemeRiver is not designed to visualise the sentiment, but it provides the inspiration of using river to visual the increase and decrease of value on temporal dimension. Similarly, TextFlow [CLT+11] and EventFlow [LYK+12] apply the representation of river in their systems as well.

Instead of plotting river, Mishne and Rijke [MDR+06] propose a system, named MoodViews, to track and analyse states-of-mind of massive bloggers. They plot emotions with a line graph on the time axis. Different from the other sentiment visualisation tools, it is able to visualise multiple types of moods.

There are also visualisation techniques regardless of time. Graells-Garrido et al. [GGLBY16] propose visualisation widgets to show both positive and negative sentiments for exploratory search by using scatter plot and parallel coordinates. The scatter plot represents the positive sentiment with the value on the x-axis and the negative sentiment with the value on the y-axis. The parallel coordinates represent the positive sentiment on the left axis and the negative sentiment on the right axis. Their evaluation suggests that scatter plot is more suitable for exploratory search.

To mining customer opinions on products, Gamon et al. [GACOR05] and Carenini et al. [CNP06] employ a tree-map [Shn92] to display and summarise the opinion from a large number of evaluative text. Tree-map depicts each node in a tree as a rectangle, and uses the size and color of the node to represent the value of two dimensions. Gamon et al. [GACOR05] provide a system named Pulse, which is designed to visualise the clusters obtained from customer opinions and their associated sentiments by using tree-maps. This system represents each cluster as a box, in

[3]Sankey graph is a flow graph used to illustrate complex information flow.

which the size of the box indicates the number of sentences included in this cluster, and the color of the box ranging from red to green indicates the sentiment of the cluster from very positive to very negative. Differ from Pulse, each box in the interactive multimedia interface proposed by Carenini et al. [CNP06] represents a component of the product, and these components are allowed to retain their hierarchies by positioning the box nested. Another visualisation technique, named OpinionBlocks, has been applied by Alper et al. [AYHK11]. They propose an interactive visualisation tool able to progressively disclose increasingly details of the review. The summary view of this tool is divided on the horizontal axis among commonly discussed features (extracted from the reviews). The vertical axis of the summary view points out the extracted sentiment for the certain feature by locating the positive and negative sentiment separately at the upper and bottom side.

Besides, a rose plot has been employed by Gregory et al. [GCW+06] to visualise the affective content of documents. They design an interactive system based on the rose plot, which can be used for a further interactive exploration of a large corpus toward multiple sentiments rather than polarity sentiments (i.e. positive and negative). This kind of plot is able to show the variation of median or quartile clearly.

3. Research Methodology

In this chapter, we will describe the approaches we used to analyse the sentiment on Wikipedia concepts.

Considering that each Wikipedia article is a description text for one concept, we choose document-level sentiment analysis to give sentiment score for each document (article or talk). More precisely, we use lexicon-based method of sentiment analysis combining bag-of-words model [SM86, ZJZ10] to measure the sentiment.

Figure 3.1 shows a straightforward processing pipeline.

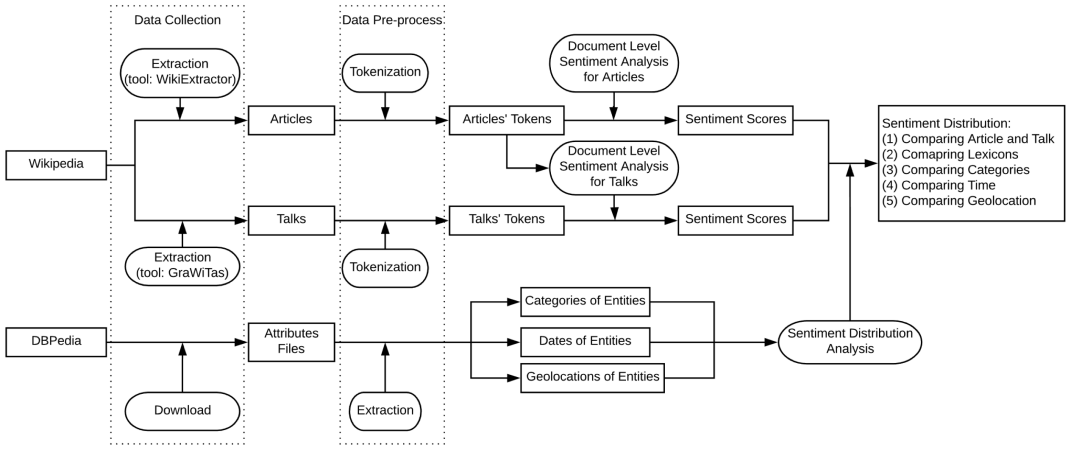


Figure 3.1.: Processing pipeline

First, we prepare the necessary data, which includes two parts: data collection and data pre-processing. Wikipedia concepts and their corresponding attributes are the principal source data. We collect Wikipedia concepts by extracting the whole English Wikipedia articles and talks with the help of existing tools WikiExtractor and GraWiTas. In order to implement lexicon-based sentiment analysis, we carry out tokenization, token cleaning, and lemmatisation during data pre-processing. With regard to DBpedia data, we extract the categories, dates, and geolocations of entities as attributes.

After that, we execute document-level sentiment analysis for articles and talks. The approach to articles is different from the approach to talks. Basically, the sentiment score of an article has been determined by the frequency of sentiment words appearing in this article, and the sentiment words are distinguished by the used lexicon. As opposed to article that the whole page is used to describe the concept, talk is used for editors to present opinions about the text on the article page, which

means the terms occurred in the article text tend to be referred in the talk text. These terms act as references in the talk text however carry no opinions. Therefore, we lower down the weight for these terms according to the number of times they have occurred in the corresponding article text. In other words, the sentiment score of each talk has been determined by the frequency of sentiment words in both article and talk text. The detailed approaches are described in Chapter 5.

After calculating the sentiment score for the whole Wikipedia entities, we will take a closer look at the sentiment distributions in terms of articles and talks separately, and compare the distributions against multiple lexicons to get untrivial properties for each lexicon. As a further exploration, we will examine the changes of general sentiments over time by taking entities of people and events as instances, and probe into novel insights between score and various features (e.g. score of person entities and the date of birth, score of event entities and the type of events, or score and lexicons).

4. Data Collection and Pre-processing

In order to get sentiment score for Wikipedia concepts, we need to collect Wikipedia text for concepts. For a further exploration of sentiment distributions with varying time and geolocation attributes, we need to collect the attributes for each concepts.

In this chapter, the explanation of approaches to data collection and data pre-processing will be presented, followed by a description and statistics of data. In total, we collected and pre-processed 5,669,867 articles and 1,906,375 talks.

4.1. Data Description

In this section, we explain the main source data we used, including data from Wikipedia and DBpedia.

4.1.1. Wikipedia Data Description

Overview

Wikipedia has article page, talk page, and user page. Article page contains description about one concept, talk page is an area for editors to discuss opinions regarding the corresponding article, and user page describes user profile.

Wikipedia provides several different ways to access its database, either by crawling from a server in real time or by downloading a copy of it.

The copy is called Wikipedia Dump, which is provided by Wikipedia[1]. There are a variety of versions of dumps distinguished by date and content.

Structure of Wikipedia Articles

Wikipedia lists its elements in article into four sections: (1) Before the lead section; (2) Body; (3) Appendices; (4) Bottom matter[2]. As described in its official website, *Before the lead section* contains Hatnotes, Infoboxes, etc. *Body* contains Introduction, Table of contents, Content. *Appendices* contains Works or publications, See also, Notes and references, Further reading, and External links. *Bottom matter* contains Other navigation templates, Categories, etc.

[1] `https://en.wikipedia.org/wiki/Wikipedia:Database_download`, as seen on Oct. 10, 2018

[2] `https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Layout`, as seen on Oct. 10, 2018

There are two kinds of article pages different from common pages, namely disambiguation page and redirect page, and these might affect the result of the analysis.

Disambiguation page[3] is used when different entities hold the same name. Since Wikipedia distinguishes entities by name, when the same name being used by more than one entities, the conflict appears. For example, "Mercury" is the name, which different entities (planets, chemical, etc.) are using. To resolve it, Wikipedia creates a *disambiguation page* with the list of the entities sharing the same name and their corresponding links. This *disambiguation page* is entitled the shared name, and these entities will have a parenthesis to label the difference in the subject. The content of this kind of pages usually makes no sense for sentiment analysis since there are only a few words on the page served to announce the disambiguation. For example, the *disambiguation page* for "Mercury" only contains "Mercury usually refers to:" and "Mercury may also refer to:" two sentences, which do not make sense for sentiment analysis.

Redirect page[4] is used when the same entity holds more than one names. Wikipedia creates a main page for one name, and creates a *redirect page* for each of the rest of the names. While a *redirect page* is being visited, Wikipedia will send visitors to the main page. Wikipedia Dump leaves the content empty but remarks a redirect label for each *redirect page*.

Structure of Wikipedia talks

Wikipedia talks is a space offered by Wikipedia for editors to discuss the changes or improvements of articles. Each article has its own talk page. Talk page is sectioned by topics. Comments for the same topic are separated by indentations. Specifically, each comment is supposed to have one more indentation than the comment it replies to[5].

A Special feature for talk page is Archives. Wikipedia periodically archives old discussions on a talk page as Archives on that page when that page becomes too large[6]. Therefore the content of Archives is also part of the talk page.

Parsing Tool for Wikipedia Dump

There exists several tools designed to parse article pages and talk pages. As we researched, WikiExtractor and GraWiTas meet our requirements better than others, and therefore have been used in this thesis.

WikiExtractor WikiExtractor is a python script to get the plain text of articles (not for talks) from Wikipedia dump, by discarding any other information or annotation,

[3] `https://en.wikipedia.org/wiki/Wikipedia:Disambiguation`, as seen on Oct. 10, 2018
[4] `https://en.wikipedia.org/wiki/Wikipedia:Redirect`, as seen on Oct. 10, 2018
[5] `https://en.wikipedia.org/wiki/Help:Talk_pages`, as seen on Oct. 10, 2018
[6] The significance of Archives and the method to archive a talk page could be found in `https://en.wikipedia.org/wiki/Help:Archiving_a_talk_page`

```
<page>
  <title>Koblenz</title>
  <ns>0</ns>
  <id>167926</id>
  <revision>
    <id>841530375</id>
    ......
    <text xml:space="preserve">{{About|the city in Germany}}
    {{Infobox ......}}
    '''Koblenz''' ({{IPA-de|'ko:blɛnts|lang|De-Koblenz.ogg}};
    {{lang-fr|Coblence}}), spelled '''Coblenz'''&lt;ref&gt;Other historical
    spellings include ''Covelenz'' and ''Cobelenz''. In the local dialect the name
    is ''Kowelenz''.&lt;/ref&gt; before 1926, is a [[Germany|German]] city
    situated on both banks of the [[Rhine]] where it is joined by the [[Moselle]].
    ......
    ==Economy==
    [[File:KoblenzFromTheISS.jpg|thumb|Koblenz, as seen from the [[International
    Space Station]]]]
    [[File:Königsbacher Brauerei Koblenz.jpg|thumb|200px|Königsbacher brewery]]
    Koblenz is a principal seat of the Mosel and Rhenish wine trade, and also does
    a large business in the export of mineral waters.
    ......
    ==Education==
    The campus Koblenz of [[University of Koblenz and Landau]] is located in the
    city.
    The University of Applied Sciences Koblenz ([[German language|German]]:
    ''Hochschule Koblenz'') is also located in the city.
    ......
    The children's toy [[yo-yo]] was nicknamed ''de Coblenz (Koblenz)'' in
    18th-century France, referring to the large number of noble French émigrées
    then living in the city.
    ......
    ==External links==
    {{commons|Koblenz}}
    {{wikivoyage|Koblenz}}
    * [https://www.koblenz.de/stadtleben_kultur/koblenz_allgemeine_infos_e.html
    Official website]
    ......
  </revision>
</page>
```

```
<doc id="167926" url="https://
en.wikipedia.org/wiki?curid=167926"
title="Koblenz">
Koblenz
Koblenz (; ), spelled Coblenz before 1926,
is a German city situated on both banks of
the Rhine where it is joined by the Moselle.
......
Koblenz is a principal seat of the Mosel
and Rhenish wine trade, and also does a
large business in the export of mineral
waters.
......
The campus Koblenz of University of Koblenz
and Landau is located in the city.
The University of Applied Sciences Koblenz (
German: "Hochschule Koblenz") is also
located in the city.
......
The children's toy yo-yo was nicknamed "de
Coblenz (Koblenz)" in 18th century France,
referring to the large number of noble
French émigrées then living in the city.
</doc>
```

Figure 4.1.: Format of Wikipedia Dump before and after it has been parsed by WikiExtractor. The left one refers to the text of article page "Koblenz" in Wikipedia Dump, and the right one refers to the corresponding text after being parsed by WikiExtractor.

such as images, tables, references, and lists[7]. In other words, WikiExtractor only retains *Introduction* and *Content* from *Body* section among the entire four sections as described in Section 4.1.1. Moreover, it removes all the *Section name* in the *Content*. With regard to *Internal and External links*[8] and *Templates*[9], it leaves front text aside and discards the rests. An example of how the data in Wikipedia Dump looks like before and after it has been parsed by WikiExtractor is shown in Figure 4.1. With regard to *redirect page*, WikiExtractor leaves the page out by default, in order to avoid any repetition (The page which has been redirected to is parsed, so the *redirect page* is a repetition).

GraWiTas WikiExtractor is unable to parse talk page directly. Instead, GraWiTas [CSR17] as a lightweight and fast tool for Wikipedia talk page, is able to parse each comment by distinguishing the timestamp, IP address, etc. Also, it extracts *Archives*[10].

[7] `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`, as seen on Oct. 10, 2018
[8] `https://www.mediawiki.org/wiki/Help:Links`, as seen on Oct. 10, 2018
[9] `https://en.wikipedia.org/wiki/Help:Template`, as seen on Oct. 10, 2018
[10] Archives is a section for Wikipedia talks, which has been described earlier of this section.

There are three components in GraWiTas that can perform three different functions. The first one is *Crawler Component*, which is used to extract talk page content for given Wiki URLs. The second one is *Dump Component*, which is used to process the full Wikipedia XML dump and export the comments regarding certain articles or editors. The third one is *Core Parser Component*, which can be fed with a raw talk page in Wiki markup[11] and export the parsed talk page into the designated format.

The output of *Dump Component* is formatted as three tables, as shown in Table 4.1, stored in a sqlite3 SQL database. Each entry in table *comment* is an individual comment.

Name of Table	Contained Columns
comment	id, parent_id, user_id, article_id, date, section, text
user	id, name
article	id, title

Table 4.1.: Information of SQL tables obtained from GraWiTas[12]

4.1.2. DBpedia Data Description

DBpedia is a community, which extracts structured information from Wikipedia and re-organises them in a semantic format[13]. It is a typical case of Linked Data [BL06, BHBL11]. Wikipedia contains a lot of structured data, such as categories, geographical coordinates, external links, and information inside the Infobox. DBpedia extracts them to build a huge knowledge database and publishes the database online. It helps users to easily get attributes of Wikipedia entities.

DBpedia uses Resource Description Framework (RDF) [KCM04] as data model for extracted information. Entities in the database are all represented by the Internationalised Resource Identifier (IRI) or Uniform Resource Identifier (URI). Wikipedia gives each article a URI, which is the same as the URL of the article. The format of the URI is <http://en.wikipedia.org/wiki/Name>. Correspondingly, the URI of an entity on DBpedia is formatted as <http://dbpedia.org/resource/Name>, where the *Name* is the same as the *Name* in the URI of Wikipedia, and it is actually the title of the Wikipedia article.

DBpedia provides turtle and quad-turtle two different formats for the dataset. We use turtle file as the source dataset. The format of turtle file is in N-Triples, and the format of quad-turtle is adding extra information to every triple. Each triple of N-Triples consists of a subject, a predicate, an object, and a dot at the end. An example of N-Triples has been shown in Table 4.2.

[11] Wiki markup as a markup for Wikipedia, consists of the syntax and keywords. `https://en.wikipedia.org/wiki/Help:Wikitext`, as seen on Aug. 14, 2018

[12] `https://github.com/bencabrera/GraWiTas`, as seen on Aug. 14, 2018

[13] `http://wiki.dbpedia.org/about`, as seen on Aug. 14, 2018

Format of N-Triples	<subject> <predicate> <object>.
Example	<http://dbpedia.org/resource/Fuzhou_University> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://schema.org/CollegeOrUniversity>.

Table 4.2.: Format of N-Triples in DBpedia dataset

Table 4.3 shows the data files from DBpedia which are involved in this thesis as well as their descriptions. Section 4.3.2 will explain the detailed usage of these files.

File	**Description**
instance_types_transitive_en.ttl	Category file. Describe the category of entities, such as people, events, etc.
mappingbased_literals_en.ttl	Literal file. Describe properties of entities which are literal value, such as name and date.
geo_coordinates_en.ttl	Geolocation file 1. Describe geographic coordinates of entities extracted from the attributes of Wikipedia articles.
geo_coordinates_mappingbased_en.ttl	Geolocation file 2. Describe geographic coordinates of entities extracted from Infoboxes.
infobox_properties_en.ttl	Infobox file. Present properties listed in Infoboxes, such as occupations.

Table 4.3.: Description of dataset on DBpedia

4.2. Data Collection

In this section, we will introduce the procedure of data collection, including data from Wikipedia and DBpedia. DBpedia data is directly collected by downloading from DBpedia online database. The version has been used is 04.2016[14]. The following text will focus on the collection of Wikipedia data.

In order to obtain Wikipedia text more efficiently, we use Wikipedia Dump as source dataset. The Wikipedia dump we used contains only current revisions with both article pages and talk pages[15]. In total, we collected 5,669,867 articles and 1,906,375 talks.

Considering the performance of several existing parsing tools, WikiExtractor has been chosen to collect and parse the Wikipedia articles. The reason to apply it is that it discards sections such as *Appendices* and *Bottom matter* which has few relations to

[14] `https://wiki.dbpedia.org/downloads-2016-04`, as seen on Jul. 01, 2017

[15] `https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-meta-current.xml.bz2`, as seen on July. 03, 2018

sentiment, and only extracts plain text (see Section 4.1.1). Moreover, it is able to deal with *Templates* and HTML markup.

The collecting and parsing of Wikipedia talks are performed by GraWiTas, since it is able to deal with individual comments, and take Archives into consideration. As we described in Section 4.1.1, *Dump Component* of GraWiTas is able to parse talks into comments, and it stores the results into sqlite3 SQL database with a table named *comment*. Each entry in this table is an individual comment annotated with *article_id*. Depend on the table, comments toward the same entity, including comments in *Archives*, have been grouped by identifying their *article_id*.

One of the limitations of GraWiTas is that it is unable to handle the title of entities which contain special symbols, such as double quotation mark (") or ampersand mark (&), in their titles. More specifically, while parsing the title, GraWiTas replaces string after the special symbol (including the special symbol itself) by a space. For instance, talk page named "Chicago & North Western Railway Stone Arch Bridge"[16] has been entitled "Chicago " (consists of *Chicago* and a space) by GraWiTas, and talk page named "The "E" Ticket" has been entitled "The ". This would not only cause the problem of a mismatch between title and text, but also lead to the grouping of comments of several concepts. For example, the concepts "Chicago & North Western Railway Stone Arch Bridge" and "Chicago "L" rolling stock" would share the same title, as a result, comments for both of them will be clustered under name "Chicago ". In order to avoid either the mismatch between title and text, or the mix-up between truncated titles and real title (e.g. mix-up between "Chicago " and "Chicago"), entities which contain special symbols[17] in their titles are excluded. The number of excluded entities is 1328, and the number of left entities is 1,906,375.

It is important to note that WikiExtractor carries out different process from GraWiTas during parsing. For example, WikiExtractor deals with *Internal* and *External links* by leaving their front text aside, and deals with different *Templates* in different ways to get different output text, which GraWiTas is not designed. For the sake of consistency between treatments of articles and talks, we apply the function about text processing from WikiExtractor (with slight changes to the code) to the talks that have been parsed by GraWiTas as a second parsing.

4.3. Data Pre-processing

4.3.1. Wikipedia Data Pre-processing

Before doing lexicon-based sentiment analysis, we need to pre-process the Wikipedia text, such as cleaning, tokenizing, and lemmatising the text. Figure 4.2 shows the

[16] To clarify the quotation used in this paragraph: Curly double quotations are used to mark out title for the talk page, and straight double quotation mark is used to represent the mark itself (i.e. double quotation mark) inside title.

[17] Double quotation mark and ampersand mark are the only discovered mark with such exceptional situation, therefore we are intending to say these two symbols when we mention special symbols in the context.

overview of Wikipedia data pre-processing.

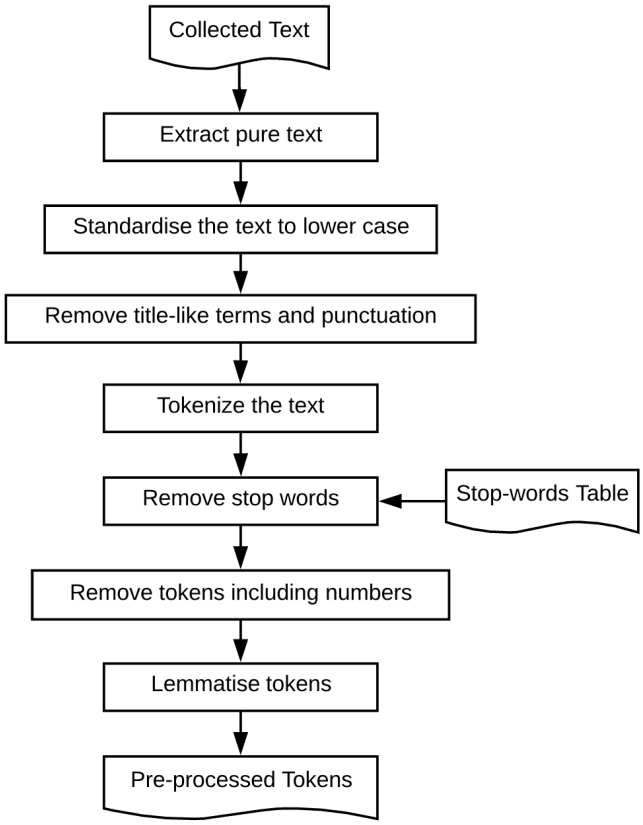


Figure 4.2.: Pre-processing of Wikipedia text

The detailed process has been described as follows:

1. Extraction of pure text from articles and talks. Talks collected are already in pure text. Articles collected are in the form of a XML file (see Figure 4.1). More specifically, text for each entity is delimited with tag <doc> and </doc>, and the name (i.e. title) of this entity is declared by attribute *title* inside tag <doc>.

2. Standardisation of the text to lower case for the sake of the matching in the following steps.

3. Removal of terms which are same as the title. Since Wikipedia article is the descriptive text about an entity, the name of the entity must be appearing in the article with high frequency (Wikipedia talk as the text about communication would be the same). However, the sentiment words in the title carry no opinion. It is just a part of the title. For example, the sentimental word *great* included in the title of entity "Great St. Wilfrid Stakes", but in the context, it is

just a part of the name. Therefore, we remove all terms that same as the title to avoid the influence described above[18].

4. Removal of punctuation except for hyphen connecting words. Punctuation, such as comma and full stops, are meaningless and decrease the sentiment score by increasing the length of the text. Therefore we remove all the punctuation with the exception of the hyphen. A hyphenated word should be regarded as one word such as *well-educated*. Removing hyphen might change the meaning of the original word. So hyphen within the word is retained.

5. Tokenization of the text. This process is implemented by NLTK package[19]. More specifically, the function *RegexpTokenizer* from library *tokenize* of NLTK has been used. After tokenizing, we get bags of tokens for each text.

6. Removal of stop words. To make the result more precise, we remove all stop words listed in NLTK package. A total of 153 stop words are taken into consideration, which are listed in Appendix A.

7. Removal of words which include numbers. Since digital number has little relationship with sentimental expression, all words containing digits, such as "800", "29th", have been removed.

8. Lemmatisation of tokens. The function *WordNetLemmatizer* from NLTK package has been used to lemmatise tokens. Specifically, it is carried out by, for example, removing "s" from the end of plural nouns or from the end of the verb in the form of the third person singular.

9. Storage of results. For each text, the tokens and the number of tokens (i.e. the length of the text) are recorded accordingly.

By this process, the tokens and lengths of the entities for articles and talks are recorded respectively.

4.3.2. DBpedia Data Pre-processing

In this part, we will introduce the procedure of extracting attributes of Wikipedia entities, including categories, dates, and geolocations, from the collected DBpedia data.

[18] We exclude terms which are exactly same as the title instead of excluding individual keywords in the title because those individual keywords could be used to describe other stuff in the context. For example, we remove the term "great st. wilfrid stakes" instead of "great", as "great" could be used to describe other things.

[19] `https://www.nltk.org`, as seen on Oct. 10, 2018

Category Extraction

In this part, we focus on entities of people and events by extracting entities with these two categories.

The categories of Wikipedia entities can be identified by the definition in category file (see Table 4.3). Specifically, different URIs are used to identify different categories. Table 4.4 shows the URIs being used in this thesis to identify the category of *Person* and *Event*. Each category might be identified with various URIs. As we examined, the listed three URIs for *Person* are able to cover most of entities of people, and the URI for *Event* covers all of entities of events[20].

No.	URIs to Identify Person	URIs to Identify Event
1	http://dbpedia.org/ontology/Person	http://schema.org/Event
2	http://schema.org/Person	
3	http://xmlns.com/foaf/0.1/Person	

Table 4.4.: URIs to identify entities of people and events

Date Extraction

In this part, the approaches to extract date information will be explained. Specifically, the birth date will be focused for entities with category of *Person*, and occurrence date will be focused for entities with category of *Event*. Date information is stored in the literal file (see Table 4.3).

To locate birth date information, we regard all predicates (URIs) with suffix "/birthDate" as our target predicates. Combining category information, we obtained 804,631 people with birth date information in total.

With regard to date of events, there are three principal properties relating to date: *date*[21], *startDate*[22] and *endDate*[23]. As the name suggests, *date* indicates the date of occurrence, *startDate* indicates the start date of the event, and *endDate* indicates the end date. Obviously, *startDate* and *endDate* are being used when the event lasts for a period. Among these three properties, *date* has been set with the highest priority. Considering that event usually catches few attention on the day of the start, whereas it collects more opinion on the day of the end, we consequently set the priority of the end date over the start date. In other words, we adopt end date if both of them are present. Combing the category information, we obtained a total of 23,865 events with date information.

[20] As we researched, there are 7 URIs referring to *Event*, and all entities covered by these 7 URIs are covered by URI <http://schema.org/Event>.

[21] *date* is a property in DBpedia ontology. Its URI is <http://dbpedia.org/ontology/date>.

[22] *startDate* is a property in DBpedia ontology. Its URI is <http://dbpedia.org/ontology/startDate>.

[23] *endDate* is a property in DBpedia ontology. Its URI is <http://dbpedia.org/ontology/endDate>.

Geolocation Extraction

In this part, the approaches to get geolocation for entities will be introduced.

Wikipedia provides geolocations in two places, one is the attribute of the article page, the other one is the Infobox. DBpedia stores them separately (see Table 4.3). This two geolocations could be different even regard the same entity. Take entity *Algeria*[24] as an example, the geographical coordinates extracted from the attribute of article are (2° E, 28° N), which is around the middle of the country, whereas the coordinates extracted from Infobox are (36°42′ N, 3°13′ E), which is located in the Capital of Algeria, named Algiers. The URIs of properties we used to locate latitude and longitude are shown in table 4.5.

Property	URI to Identify the Property
longitude	<http://www.w3.org/2003/01/geo/wgs84_pos#long>
latitude	<http://www.w3.org/2003/01/geo/wgs84_pos#lat>

Table 4.5.: URIs to locate latitude and longitude

Besides, the number of geolocation extracted from the attribute of article page is at most one, but geolocations from Infobox could be multiple. For example, geolocations from Infobox of *Amazon River* contain source location and mouth location[25]. Moreover, a region might have the highest position, capital position, or government position. A person could have positions such as birthplace or resting place. As compared, geolocation from the attribute is more representative. Therefore, geolocation from attribute is set as a higher priority. If this one is absent then the geolocations from the Infobox are being considered. With regard to the priority of multiple geolocations from Infobox, the first processed one will be used. An exception to the rule is entities of rivers: the mouth position of a river will be collected if there exists one. The reason for the exception is that there is a large number of entities belonging to rivers and 4,590 of them are possessing information of mouth position.

In order to manipulate geographic data spatially, we convert geographical coordinates into data structure named *GeoDataFrame*[26].

GeoDataFrame, provided by GeoPandas library, is a data structure specifically for geographic data. It is extended from DataFrame, a data structure in Pandas

[24] Algeria is a country in North Africa. `https://en.wikipedia.org/wiki/Algeria`, as seen on Oct. 10, 2018

[25] The URI of entity extracted from Infobox might be a derivative one. For example, one URI extracted for Amazon River is <http://dbpedia.org/resource/Amazon_River__sourcePosition__1>, which is a derivative URI indicating the position of Amazon River's source. Another extracted URI is <http://dbpedia.org/resource/Amazon_River__mouthPosition__1>, which indicating the mouth position of the Amazon river. All those derivative URI are taken into consideration in this thesis.

[26] `http://geopandas.org/data_structures.html`, as seen on Oct. 10, 2018

library[27]. DataFrame is a two-dimensional data structure, represented as a table. Comparing to DataFrame, GeoDataFrame has a special column named *geometry*, which stores the geometry object. GeoPandas has three basic type for a geometric object: Points/ Multi-Points, Line/ Multi-Lines, and Polygons/ Multi-Polygons. These types come from Shapely library[28]. While carrying out spatiality related operations on data with GeoDataFrame structure, the operations will be automatically applied to the column *geometry*.

The procedure of converting geographic coordinates into GeoDataFrame structure has been shown in Figure 4.3. The obtained GeoDataFrame contains coordinates, country, and continent for each entity.

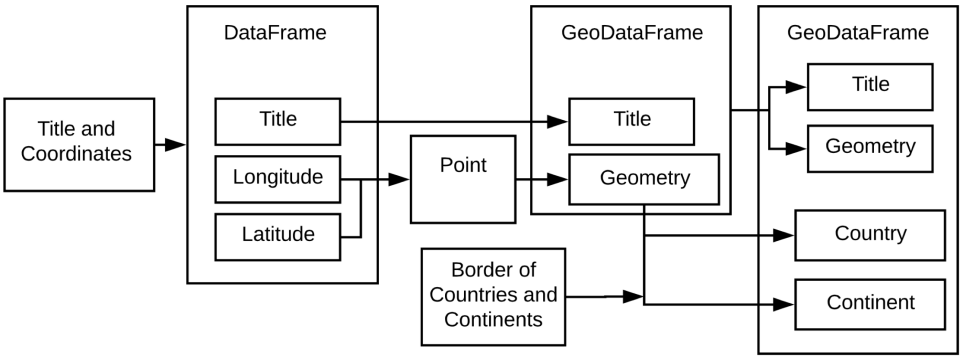


Figure 4.3.: Procedure of converting from coordinates into GeoDataFrame structure

The detailed procedure has been described as follows:

1. Convert entities into the DataFrame structure that containing their geographic coordinates, and create *Point*[29] structure by combining longitude and latitude of entities.

2. Create a preliminary GeoDataFrame structure to store title and geometry of entities. Geometry is created with *Point* from the previous step.

3. Generate GeoDataFrame containing countries and continents of entities. The recognition of country and continent for entity has been done by using GeoPandas library, which provides border between countries all over the world (i.e. world map dataset), and a method *sjoin*[30] to determine the country and continent a geolocation belongs to. Table 4.6 shows the key information of world map offered by GeoPandas taking Germany as an instance.

[27] `https://pandas.pydata.org/pandas-docs/stable/dsintro.html`, as seen on Oct. 10, 2018

[28] `https://pypi.org/project/Shapely/`, as seen on Oct. 10, 2018

[29] Point is a data type consist of longitude and latitude. `http://toblerity.org/shapely/shapely.geometry.html#module-shapely.geometry.point`, as seen on Oct. 10, 2018

[30] `http://geopandas.org/reference/geopandas.sjoin.html`, as seen on Oct. 10, 2018

continent	name	iso_a3	geometry
Europe	Germany	DEU	POLYGON ((9.9219 54.9831,...

Table 4.6.: Example of information describing world map stored in GeoPandas

Geopandas world map dataset contains eight continents: North America, Africa, Asia, Europe, South America, Antarctica, Oceania, and Seven seas (open ocean). However, by looking into the test result, we found that the continent of *Seven seas* is not containing the well-known seven oceanic bodies of water, but only *French Southern* and *Antarctic Lands*. Therefore entities located out of the mentioned area would not have country and continent assigned. There is a special case of assigning countries. The coordinates (24° N, 25° E) belong to both Egypt and Libya, and GeoPandas assign both countries to these entities. In order to keep each entity tagged to one country, the country this entity belongs to will be picked randomly[31].

4.4. Data Statistics

The number of collected and pre-processed data is presented in Table 4.7.

Data	**Statistics**
Entities with article page	5,669,867
Entities with talk page	1,906,375
Entities of People	1,190,565
Entities of Events	76,022
Entities of people with birth date / exclude BC[32]	804,631 / 804,604
Entities of events with occurrence date / exclude BC	23,865 / 23,830
Entities with geographical coordinates	956,916
Entities with continent and country	898,148

Table 4.7.: Statistics of collected and pre-processed data

The number of entities with article page is counted after excluding redirect page and empty page. The number of talk pages collected by GraWiTas is 1,909,117, but 1414 of them are empty, 1328 of them are assigned with the wrong title (see Section 4.2. Consequently, the number of pages being pre-processed is 1,906,375.

With respect to attributes of entities, we extract birth date for people, occurrence date for events, and geolocation for all categories. In terms of geolocation, it can be

[31] In this thesis, there are only three entities: *Senussi Campaign*, *Libyan Desert*, and *Western Desert Campaign*, have such situation.

[32] "exclude BC" here means exclude entities whose date before Christ. Date before Christ will not be processed in our analysis because of library's limitations. Therefore we clarify the number in total and the number after excluding BC.

seen from the number in Table 4.7 that the number of entities with country information is less than the number of entities with geographic coordinates, since entities located in ocean might have no country allocated, furthermore, most of entities located in ocean are erroneous data according to our manual examination. The correction of this error will be considered as one of the future works (see Chapter 8).

5. Document-Level Sentiment Analysis

In this Chapter, we describe the approaches to sentiment analysis of Wikipedia articles and talks separately with a lexicon-based and document-level method.

5.1. Sentiment Lexicon

A sentiment lexicon is a dictionary containing a list of sentiment words, acting as an indicator of sentiments. Sentiment words usually refer to positive sentiment words and negative sentiment words, used to express positive emotions and negative emotions respectively. It is important to note that, word listed in sentiment lexicon might express no sentiments in the context, and word absent in sentiment lexicon might bear sentiments [Liu10]. Nevertheless, sentiment lexicon is instrumental in sentiment analysis. The following paragraphs describe the four sentiment lexicons: Opinion Lexicon (OL), Multi-Perspective Question Answering Subjectivity Lexicon (MPQA), Linguistic Inquiry and Word Count (LIWC), and Affective Norms for English Words (ANEW), which are involved in this thesis.

OL Opinion Lexicon is a list of positive sentiment words and negative sentiment words developed by Hu and Liu [HL04], and it was compiled over many years. There are 6789 words consisting of 2006 positive words and 4783 negative words in the version we are using[1]. Besides, as the source announced[2], there are many misspelled words in OL, which are not mistakes, but vocabularies commonly used in social media content.

MPQA Multi-Perspective Question Answering Subjectivity Lexicon[3] as a sentiment lexicon collected from both manually developed resources and automatically identified resources with annotated and unannotated data [WWH05, RW03]. A total of 6457 sentiment words have been extracted consisting of 2304 positive words and 4153 negative words. For each word it offers six attributes described as follows:

- type - It is used to describe the strength of the subjectivity of word. The value could be either *strongsubj* or *weaksubj*, in which *strongsubj* indicates that the

[1] Since it is keeping on being compiled, the number of sentiment words is increasing over time

[2] `https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html`, as seen on Oct.10, 2018

[3] `http://mpqa.cs.pitt.edu/lexicons/subj_lexicon`, as seen on Oct.10, 2018

word is considered strongly subjective in most context, and *weaksubj* indicates that the word is considered subjective in certain context.

- len - It represents the number of words in this term. Since all terms in MPQA are single word, value for this field is always 1.
- word1 - It shows the token (if stemmed1=*y*) or stem (if stemmed1=*n*) of this word.
- pos1 - It presents the part-of-speech of this word.
- stemmed1 - It used to indicate whether the word is stemmed or not. Value of it could be *y* (yes) or *n* (no).
- priorpolarity - It indicates the prior polarity out of context, which could be *positive*, *negative*, *neutral*, or *both*.

However, there are two implicit issues. First, we applied lemmatisation for all tokens while MPQA contains stemmed tokens and unstemmed tokens. In other words, the lemmatised token from text might be regarded as a non-sentiment word because it fails to match the unstemmed sentiment word from MPQA. Second, the ignorance of part-of-speech in our sentiment analysis will occur issues while different part-of-speech for the same word have conflict prior polarities. To evaluate the effect of these two issues, we make a statistics by examining MPQA. In total, there are 7629 items in MPQA with either positive or negative as the prior polarity, in which 1172 items are repeated with different parts of speech but the same polarity[4], 6 items are repeated with different parts of speech and conflict polarities[5]. Inside the 6457 extracted words, 6415 of them have no problem to match the pre-processed text (in which 6250 words are already in the form of lemmatisation and 165 words have the lemmatised form in MPQA[6]). In the end, there are 42 words that have no lemmatised form in MPQA, which might miss out the matches between text and lexicon. Considering that the problematic portion is small, these issues have been ignored in our analysis.

LIWC Linguistic Inquiry and Word Count 2007 Dictionary [PCI+] is a lexicon developed for the study of sentiments in verbal and written text of individuals[7]. LIWC2007 English lexicon consists of two sections, one is definition for categories, the other one is information for words. Category definition section defines each category with a unique label and a unique id. The label for positive sentiment word is *posemo* with id 126 and the label for negative sentiment word is *negemo* with id 127. Word information section is a list of words with their matched categories. There is a total

[4]These 1172 items will be excluded in our analysis.

[5]These 6 items will be retained in both positive and negative set in our analysis, but positive will be chosen as polarity while determining sentiment for words.

[6]This result is produced by using NLTK package.

[7]`http://www.liwc.net/LIWC2007LanguageManual.pdf`, as seen on Oct. 10, 2018

of 4482 words in this lexicon. In our research, we extract words either matching category id 126 or id 127. Eventually, there are 407 positive words and 499 negative words have been obtained. Items in LIWC could be either the complete form or the form with an asterisk as the wild card. For example, "deligh*" is an item in LIWC, which matches delight, delighted, delightedly, delightful, etc.

ANEW Affective Norms for English Words [BL99] is different from the other three lexicons. It provides normative emotional ratings of valence (pleasure), arousal, and dominance for words. There are three rating results respectively from male, female and all subjects. Here we only focus on valence with the subject of all (both male and female). The rating scale ranges from 1 (low pleasure) to 9 (high pleasure). We regard score 5 as neutral, scores more than 5 as positive words, and scores less than 5 as negative words. There are 1030 words in ANEW list of version 1999, consisting of 580 positive words, 449 negative words, and 1 neutral word. In 2011, the number of words grew to 1034, and this is the version has been used in our thesis, with 584 positive words, 449 negative words, and 1 neutral word.

5.2. Sentiment Score Calculation

In this part, we describe the detailed approaches to sentiment analysis regarding Wikipedia concepts. As we mentioned in Chapter 3, the sentiment analysis in this thesis is carried out in document-level with a lexicon-based method. The principal idea is to determine sentiment by the frequencies of sentiment words presented in the text. Instead of classifying a text to be positive or negative by examining the weight of each polarity (usually based on sentence-level classification), we collect both positive and negative sentiments, to see the extent of sentiment expressed. Consequently, the term *total score* has been proposed to measure the total sentiment for individual entities, and it is determined by the frequency of both positive and negative tokens. Besides, *total score* will play an instrumental role in the analysis of sentiment distribution in the next Chapter.

5.2.1. Sentiment Analysis for Wikipedia Articles

Since each Wikipedia article is edited to describe an entity, the sentiment for each entity is determined by processing the plain text on its article page and measuring the frequency of sentiment words. The basic process of sentiment score calculation for each article has been presented in Figure 5.1.

First, the set of positive words and negative words has been extracted from the described four lexicons respectively. Then the sentiment tokens in the article have been identified by matching with the extracted sentiment words set. After that, the sentiment score has been calculated with both the length of the article (i.e. the number of tokens in the article) and the sentiment tokens in the article.

Since lexicons OL, MPQA, and LIWC classify words into polarity, whereas lexicon ANEW rates words into continuous valence, we treat the calculation with those lexicons in different ways.

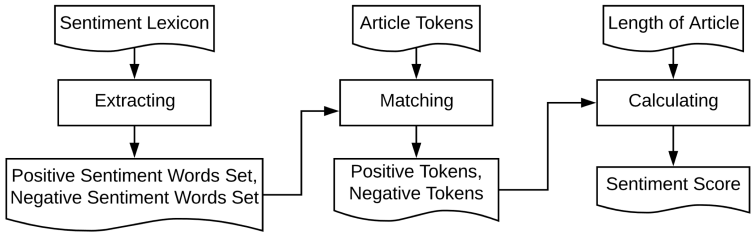


Figure 5.1.: Process of sentiment analysis for Wikipedia articles

With respect to lexicons which only label either positive or negative to words (i.e. polarity lexicon), such as OL, MPQA[8], and LIWC, we take the frequency of sentiment words as weight. With respect to lexicons which rate words with continuous valence, such as ANEW, we take both frequency and valence as factors for the weight. Equation 5.1 is the formula to calculate positive score for article based on polarity lexicon:

$$pos.score = (f_{pos}/n) * 100 \tag{5.1}$$

where f_{pos} denotes the number of positive tokens, n denotes the length of the article, and 100 is used for normalisation[9].

Analogously, negative score could be calculated with Equation 5.2:

$$neg.score = (f_{neg}/n) * 100 \tag{5.2}$$

where f_{neg} denotes the number of negative tokens.

And the total score for an article has been shown in Equation 5.3, which is the same as the sum of the positive score and negative score (differentiate to the neutralised score, which is the score of an entity calculated by subtracting the negative score from the positive score).

$$total.score = \frac{f_{pos} + f_{neg}}{n} * 100 \tag{5.3}$$

In contrast, ANEW (i.e. valence lexicon) gives a valence to each word ranging from 1 to 9, with 5 as neutral. In order to reflect the positive and negative feature for words better, we normalise the valence and make it ranging from -1 to 1 by Equation 5.4:

$$vnorm_i = (v_i - 5)/4 \tag{5.4}$$

[8]MPQA labels words with "neutral" or "both" sometimes, but we only take "positive" and "negative" into consideration.

[9]In this thesis, we normalise the sentiment score for all entities into 0 to 100.

where v_i is the original valence score for word i and $vnorm_i$ is the normalised valence score for word i. In this way, the score with positive value indicates positive words, and the the score with negative value indicates negative words.

With the normalised valence score for each word, we build the formula for the positive score of each article regarding valence lexicon, as shown in Equation 5.5:

$$pos.score = (\sum_{article} pos\_vnorm_i)/n * 100 \tag{5.5}$$

where pos_vnorm_i is the normalised valence score for word i which is positive, $(\sum_{article} pos\_vnorm_i)$ is the sum of positive valence score for all words in this article (valence score for word absent in ANEW will be regarded as 0), and n is the length of article.

Analogously, the negative score for the article can be calculated with Equation 5.6. The difference between the calculation of positive score and negative score is that we use the absolute value of obtained score as the negative score.

$$neg.score = |(\sum_{article} neg\_vnorm_i)/n * 100| \tag{5.6}$$

The total score for an article with ANEW is calculated with Equation 5.7, which is the sum of positive score and negative score.

$$total.score = pos.score + neg.score \tag{5.7}$$

An example of scoring Wikipedia articles has been shown in Appendix B.1.

5.2.2. Sentiment Analysis for Wikipedia Talks

Editors exchange opinions about the article on the corresponding talk page. Comparing to Wikipedia articles, which can be regarded as written language, talks are more related to verbal language.

Since editors are discussing the corresponding article page on Wikipedia talks, they would probably frequently refer to terms presented in the article. The more the term appears in the article, the more likely it appears on the corresponding talk page (and the frequency of it tends to be high). However, these terms do not contain any opinions from editors, but act as an reference from article page or object of discussion. Therefore, it is supposed to decrease the weight of a word depending on its occurrence in the article. In this thesis, we decrease the weight of a word by subtracting the number of occurrence in the article from the number of occurrence in the talk. In other words, the frequency of a word in the text of the talk page has been corrected by subtracting the frequency of this word in the article.

Accordingly, the formula to calculate sentiment score for one sentiment token with polarity lexicon has been shown in Equation 5.8:

$$score_i = \frac{max(ft_i - fa_i, 0)}{n_t} \tag{5.8}$$

where $score_i$ is the sentiment score for token i (i is one of the sentiment tokens extracted from text), ft_i denotes the number of times i appears on the talk page, n_t denotes the length of the current talk page, fa_i denotes the number of i appearing on the corresponding article page. If fa_i larger than ft_i, we take 0 as the corrected frequency (i.e. pick the maximum number from $ft_i - fa_i$ and 0 as numerator), since it is unreasonable for a token to have a negative frequency. It is important to note that, scores of tokens calculated with polarity lexicon are all valued with positive numbers (in contrast to scores of tokens calculated with valence lexicon).

A limitation of Equation 5.8 is that we do not take the length of the corresponding article page into consideration, which might lead to either overestimation or underestimation of the weights for tokens in an article[10]. If the length of an article shorter than the length of the corresponding talk by a significant degree, the result will underestimate the weight of tokens in the article and therefore overestimate the sentiment scores of tokens. On the contrary, if the length of a talk shorter than the length of the corresponding article by a significant degree, the result will overestimate the weight of tokens in the article and therefore underestimate the sentiment score of tokens.

After building the sentiment score for singular sentiment token, we calculate sentiment score for each talk page by summing up the sentiment scores of all tokens in the talk. The positive sentiment score for one talk page is calculated with Equation 5.9:

$$pos.score = (\sum_{i \in pos} score_i) * 100 \tag{5.9}$$

where pos denotes a set of unique positive tokens on the talk page, and 100 is used for normalisation, as mentioned in the previous section.

Similarly, the negative sentiment score is calculated in a similar way by replacing the set of unique positive tokens with the set of unique negative tokens. Accordingly, the total score of each talk is the sum of its positive score and negative score.

With regard to ANEW (valence lexicon), we use the following formula to get the sentiment score for each sentiment token:

$$score_i = \frac{max(ft_i - fa_i, 0) * vnorm_i}{n_t} \tag{5.10}$$

where the denotation of ft_i, fa_i, and n_t are consistent with Equation 5.8, and $vnorm_i$ is defined by Equation 5.4.

To calculate sentiment score for a talk page, we apply Equation 5.9 for the positive score, and Equation 5.11 for the negative score.

[10] *Weight* here indicates frequency of the token. In this Chapter, it refers to the measurement of importance of a singular token to the text. In terms of the calculation with polarity lexicon, *weight* indicates the frequency of the token, as mentioned in Chapter 3.

$$neg.score = |(\sum_{i \in neg} score_i) * 100| \tag{5.11}$$

It is important to note that, we use the same formula to calculate the positive score for polarity lexicon and valence lexicon regarding talks, but the meaning of $score_i$ in the formula of each of them are different. $score_i$ in the formula to calculate sentiment score with valence lexicon (including Equation 5.11) is defined by Equation 5.10, while in the formula to calculate sentiment score with polarity lexicon it is defined by Equation 5.8. The total score of each talk page, as set out earlier, is the sum of the positive score and negative score.

An example of how to score Wikipedia talks is shown in Appendix B.2.

6. Analysis of Sentiment Distributions

In the previous chapter, the sentiment scores for Wikipedia entities regarding articles and talks have been calculated with a lexicon-based method of sentiment analysis based on four different lexicons.

In this chapter, we will look into the sentiment distributions of Wikipedia entities from a variety of perspectives, such as the domain of Wikipedia[1], the category of entities, the sentiment lexicon, and the attributes.

By the end of this chapter, the sentiment distributions regarding people and events on temporal dimension will be explored. Before that, the data characteristics and sentiment distributions for overall Wikipedia entities will be explained.

To include both positive and negative sentiment, the total score of each entity will be used in the following analysis. It is necessary here to clarify exactly what is meant by total score. As described in Section 5.2, the total score used for individual entities, is defined as the sum of the positive score and negative score. Unless otherwise stated, all terms *total score* used in this chapter shall comply with the above definition (to differentiate another frequently used definition: the sum of the score of various entities).

6.1. Data Characteristics

In this section, the data characteristics collected from previous chapters will be presented, which include the statistics of involved entities and obtained data for each entity.

The statistics of entities with article or talk page and entities with certain attributes have been shown in Table 4.7. However, a possibility exists that an entity from DBpedia has no corresponding Wikipedia article or talk page. Also, the possibility that entities from Wikipedia have no attributes extracted from DBpedia exists. For instance, *Redirect pages* do not have their own article pages but they act as entities on DBpedia. Another example would be the article pages, which created after April 2016, do not have attributes stored in DBpedia, since the version of DBpedia dataset being used is 04.2016 (i.e. the structured data stored in DBpedia database is extracted from Wikipedia by date 04.2016). Similarly, article pages which changed their titles after April 2016 would fail to match the corresponding entities in DBpedia.

To analyse sentiment distributions of Wikipedia entities with the attributes of entities, both the text and attributes are needed. Table 6.1 shows the statistics about

[1]Domain of Wikipedia in our thesis refers to articles or talks.

entities that possess both Wikipedia text and attributes.

	Articles	**Talks**
Entities	5,669,867	1,906,375
Entities with Geolocation	922,870	274,406
Entities with Geolocation (including Country)	866,067	255,457
Entities of People	1,146,257	415,124
People with Geolocation	1,816	1,191
People with Date	775,637	289,082
People with Geolocation and Date	1,256	824
People with Geolocation (including Country) and Date	1,156	739
Entities of Events	54,071	21,621
Events with Geolocation	7,477	4,442
Events with Date	22,549	10,256
Events with Geolocation and Date	4,887	3,152
Events with Geolocation (including Country) and Date	4,108	2,609

Table 6.1.: **Statistics of Wikipedia entities combining attributes:** It shows the number of Wikipedia entities possessing various attributes, such as geolocation and date. "Geolocation (including country)" has been separated from "Geolocation" since a considerable number of geolocations are fail to obtain corresponding country, and most of those geolocations are incorrect (as set out in Section 4.3.2). Furthermore, we have specifically listed statistics on people and events, since they play an instrumental role in analysis with varying time due to their date attribute. The date for people refers to the birthday, and the date for events refers to the occurrence date. Additionally, all the dates of entities in this table are after Christ.

From the table, one can infer from both entities of people and events that the number of entities with date attribute (a total of 775,637 entities for articles describing people, 22,549 entities for articles describing events, 289,082 entities for talks describing people, and 10,256 entities for talks describing events) is more than the number of entities with geolocation attribute (a total of 1,816 entities for articles describing people, 7,477 entities for articles describing events, 1,191 entities for talks describing people, and 4,442 entities for talks describing events). The number of events that have geolocation attribute (7,477 regarding articles, and 4,442 regarding talks) is larger than the number of people (1,816 regarding articles, and 1,191 regarding talks). With respect to domain, in spite of the overall quantity gap between articles and talks, they have consistent quantities regarding each group of attributes (by comparing the number among the rows in Table 6.1 in terms of articles and talks respectively).

In order to explain further regarding the information for each of the Wikipedia

entities, Table 6.2 shows obtained data using three entities as examples. Specifically, it uses *Alfred Nobel* and *Battle of Xuzhou* as examples of entities of people and events respectively, and uses *Koblenz* as an example of entities except people and events.

	Alfred Nobel	**Battle of Xuzhou**	**Koblenz**
Birth Date	1833-10-21		
Occurrence Date		1938-05-01	
Geographical Coordinates	[18.02, 59.36]	[117.17,34.27]	[7.60, 50.36]
Article Length	1180	298	1225
Article P-Score with OL	8.4746	2.3490	3.2653
Article N-Score with OL	4.9153	3.0201	2.3673
Article P-Score with MPQA	9.3220	3.3557	4.6531
Article N-Score with MPQA	4.7458	10.0671	3.3469
Article P-Score with LIWC	8.0508	1.0067	1.7143
Article N-Score with LIWC	1.6949	8.0537	1.6327
Article P-Score with ANEW	3.4066	1.2517	2.7061
Article N-Score with ANEW	1.2803	1.4690	0.7080
Talk Length	1466	30	200
Talk P-Score with OL	4.0246	0	3.0000
Talk N-Score with OL	3.0696	0	5.5000
Talk P-Score with MPQA	6.5484	0	8.0000
Talk N-Score with MPQA	3.6835	3.3333	4.0000
Talk P-Score with LIWC	4.8431	0	9.0000
Talk N-Score with LIWC	2.1828	0	1.5000
Talk P-Score with ANEW	1.8631	0.5667	0.8588
Talk N-Score with ANEW	0.4480	0	0.0225

Table 6.2.: **Examples of obtained data for Wikipedia entities:** It uses *Alfred Nobel*, *Battle of Xuzhou*, *Koblenz* three entities as examples to show detailed data for these entities. They belong to people, events, and others[2]three categories respectively. In the table, P-Score means Positive Score, and N-Score means Negative Score.

As shown in Table 6.2, birth date is extracted for entities of people, occurrence date is extracted for entities of events, and geolocation is extracted for all entities. Besides, scores for article page and talk page are determined for all entities as described in Chapter 5.

This data will be used to analyse the sentiment distributions for Wikipedia entities in the rest of this Chapter.

[2]The others here indicates entities except for people and events.

6.2. Sentiment Distributions for Overall Wikipedia Entities

Before delving into sentiment distributions with certain categories or attributes, it is necessary to examine the sentiment of overall Wikipedia entities.

This section is combined into two parts, one of which is regarding Wikipedia articles, and the other one is regarding Wikipedia talks. Both of them are investigating the sentiment distributions based on the four mentioned lexicons.

6.2.1. Sentiment Distributions for Wikipedia Articles

To compare the sentiments of Wikipedia articles based on the four different lexicons, a series of investigations have been carried out in this part. First, the histogram (i.e. distribution) of sentiment scores for Wikipedia articles with each of the four lexicons has been presented, respectively. Second, Jensen-Shannon Divergence (JSD) has been calculated on each pair of the four distributions, to assess the similarity of these distributions. Finally, a direct comparison for each pair of lexicons has been visualised according to the score of individual entities. The entire 5,669,867 Wikipedia articles have been involved in this part.

Figure 6.1 shows four distributions, each of which is the histogram of total scores (which measures a total of positive sentiment and negative sentiment for each entity, as mentioned earlier) for the whole Wikipedia articles based on one of the four lexicons.

From the figure, it can be seen that a general pattern is present in all lexicons: the most frequent score is 0, the trend of frequency goes up at the start on score axis, and goes down from the score around the median score. The histogram of ANEW is much steeper than the other lexicons, and it is also more regular (with less number of spikes). For each subplot, the value of mean is close to the value of median, and rank of the mean in terms of the four lexicons is same to the rank of the mean. Comparing the four distributions, MPQA has the highest value with both median and mean (median=6.4516, mean=6.8168), OL has the second highest value for median and mean(median=4.2553, mean=4.7468), LIWC followed after OL is in the third place (median=3.2258, mean=3.8546), and ANEW has the least value for median and mean (median=3.1687, mean=3.5386) with a little lower than LIWC. **This result indicates that MPQA has the highest sensitivity for emotional expressions among the four lexicons, and ANEW has the least sensitivity with regard to Wikipedia articles.**

As shown in the histograms, especially for OL, MPQA, and LIWC, there are a few spikes particularly at score 20, 25, 34, 50, and 100. After looking into the details, we found that the main reason for this phenomena is the articles with extremely short lengths. Since there are only a few words in those pages, the probability of holding sentiment score would be 0, 1, 1/2, 1/3, 2/3, 1/4, etc. Accordingly, the determined score will be 0, 100, 50, 33.3, 67.6, 25, etc., and the large amount of those articles cause these spikes. One typical example is the *Disambiguation page*, as described in Section 4.1.1. Another typical example is articles about the award. The content of

this kind of articles usually contains more terms such as “winner” or “prize” than other words. Since “winner” and “prize” belong to positive words according to OL, MPQA or LIWC, these articles are, therefore, be scored with extremely high values. Examples of these cases have been listed in Appendix B.3. Compared with these three lexicons, ANEW has relatively less spikes, which seems reveal that ANEW has more restrictions of collecting sentiment words than the other three lexicons (words such as might, stand, and winner are not included in ANEW).

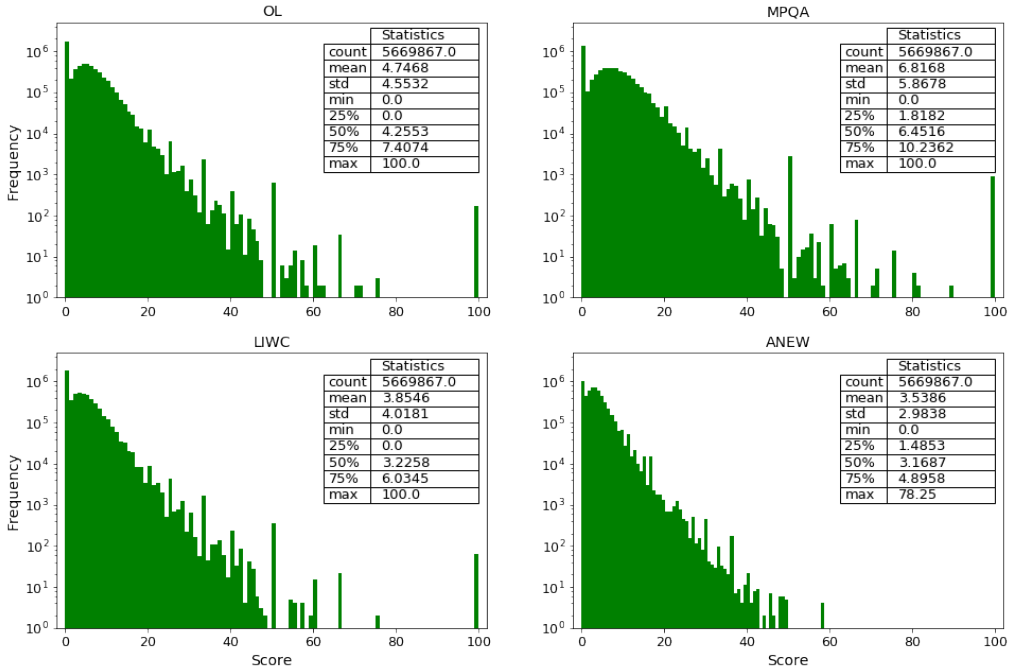


Figure 6.1.: **Distributions of total scores for the whole Wikipedia articles (5,669,867) based on OL, MPQA, LIWC, ANEW four lexicons:** The total score, ranging from 0 to 100, has been divided into 100 bins in the plot. More specifically, the frequency for the ith bin, where i larger than 1, indicates the frequency of score larger than $i-1$ and less or equal than i (mathematical notation: $i-1 < score \leqslant i$). And for the first bin, the interval of score is larger or equal to 0 and less or equal to 1 (mathematical notation: $0 \leqslant score \leqslant 1$). For example, the value on y-axis is 223,128 for the second bin of the plot with OL, which means there are 223,128 articles whose total score calculated based on OL is in the interval (1,2]. Because of the distribution characteristic, log scale has been used on y-axis.

To visualise the distinction among the four distributions, Figure 6.2 compares the four fit lines, each of which relates to one of the above four distributions.

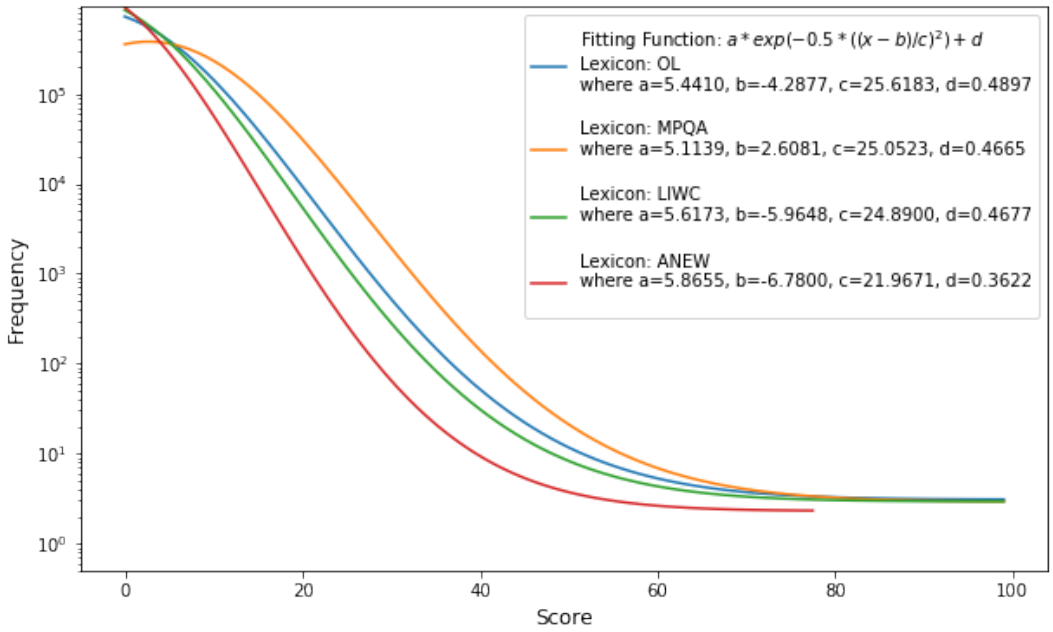


Figure 6.2.: Fitting lines of distributions for the whole Wikipedia articles (5,669,867) based on lexicon OL, MPQA, LIWC, and ANEW

From Figure 6.2, it can be observed that the fit line of ANEW goes down the fastest with the score increasing, and the fit line of MPQA goes down the slowest. The fit line of LIWC is close to the fit line of OL, and they lie between the lines of MPQA and ANEW. Moreover, the frequency of MPQA is lower than others in where the score is low ($0 \leqslant score \leqslant 5$), and higher than others in where the score is high. This result provides further evidence that MPQA tends to give entities relatively high score and ANEW tends to give entities relatively low score. OL and LIWC act moderately with the avoidance of extremes. It seems that these results are due to both the number of sentiment words in lexicon and frequencies of those words being used, considering the fact that MPQA has more sentiment words than other lexicons except for OL. ANEW has the overall lower score than the other three lexicons, it is almost certainly due to the normalised valence, which ranges the weight of each word for one appearance from -1 to 1 with mostly decimal, while other lexicons assigning 1 as the weight every time the sentiment word appears.

To measure the similarity between the sentiment distributions based on the four lexicons (from which the difference between these four lexicons can be revealed), Jensen-Shannon Divergence (JSD) of each pair of distributions has been calculated, as shown in Table 6.3. The lower the value of JSD, the stronger the similarity of distributions. For example, the JSD value of two identical distributions is 0.

	OL	MPQA	LIWC	ANEW
OL	0.0	0.0256 (3)	0.0085 (6)	0.0239 (4)
MPQA		0.0	0.0579 (2)	0.0729 (1)
LIWC			0.0	0.0163 (5)
ANEW				0.0

Table 6.3.: **Jensen-Shannon Divergence of each pair of sentiment distributions regarding the total score of the whole Wikipedia articles:** The sentiment of each distribution is calculated with one of OL, MPQA, LIWC, and ANEW lexicon. The values inside parentheses are the rank of JSDs in order of highest to lowest. Accordingly, it presents the order of similarity from the strongest to the weakest.

From the data in Table 6.3 we confirm some of the general observations that we can make from the fit lines in Figure 6.2. **Namely, the distributions of OL and LIWC similar with each other the most (with the lowest JSD value), and the distributions of ANEW and MPQA differ from each other the most (with the highest JSD value).** In addition, the most similar lexicon to MPQA is OL (JSD=0.0256) according to the sentiment distributions they worked out, and the most similar lexicon to ANEW is LIWC (JSD=0.0163).

The histograms present the distributions on score regarding different lexicons, and the JSD values reveal the similarity between the distributions. However, it is not enough to determine the similarity of different lexicons with the above methods, since the similarity on histograms concerns only score but not the entities the score relating to. As an example, assuming there are two lexicons with identical histograms, it is a possibility that the scores assigning to the same entity from different lexicons are very different. In other words, the rank of entities based on their scores is supposed to be considered as well while delving into the difference between lexicons. Consequently, a comparison of scores regarding each entity has been carried out.

Figure 6.3 shows a direct comparison between the total scores of each pair of lexicons in terms of Wikipedia articles. Each dot in the subplot represents an article, the x-axis value of the dot indicates the total score of the article based on lexicon marked on the bottom of the plot, while the y-axis value indicates the total score based on lexicon marked on the left side of the plot[3]. Since most of the scores lie in the range of 0 to 50 while the full range is from 0 to 100, the cut-off is set at 50 for both axes in each subplot, and the percentage of remaining points has been marked as ρ in each legend.

[3]The log-log scale has not been applied since it will lose data of entities who scored 0 with lexicon on either x-axis or y-axis, and the number of this kind of entities is considerable.

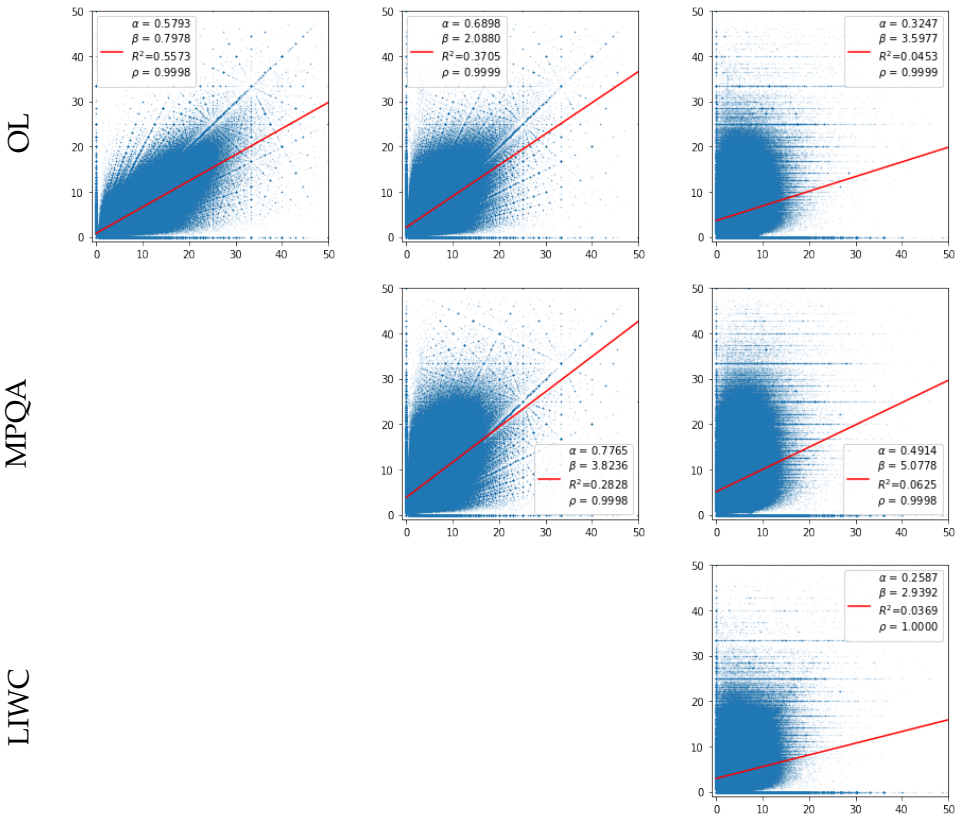


Figure 6.3.: **Direct comparison of each pair of lexicons regarding the total score of individual Wikipedia articles:** It shows scores of entities by displaying the score for one lexicon on the x-axis and the score for the other lexicon on the y-axis. The described four lexicons are involved in this figure, and each of them has been compared with the other three lexicons. After excluding the repeated ones, six scatter plots are displayed. The cut-off is set at 50 for both axes, and the percentage of remaining points are presented as ρ in the form of decimal number. The red line is the linear regression fitting to the observed points with α as slope, β as intercept, and R^2 as the coefficient of determination.

For each scatter plot, a linear least-squares regression has been presented as a red line, which is generated by the function *stats.linregress* from Scipy library, and the corresponding R^2 is shown in the legend. The closer the value of R^2 is to 1, the better the linear regression model can fit the observed points. The more likely the observed points are linear relationship, the stronger the scores of the specific two lexicons are similar with each other. As shown in Figure 6.3, the closest lexicon to OL is MPQA (R^2=0.5573), to MPQA is OL vice versa, to LIWC is OL (R^2=0.3705), and to ANEW is MPQA (R^2=0.0625). Generally, the most similar lexicons are be-

tween OL and MPQA with the highest value of R^2, and the least similar lexicons are between LIWC and ANEW with the lowest value of R^2 (R^2=0.0369). ANEW has a vary low value of R^2 to any other lexicons, which suggests the significant difference between ANEW and other lexicons. As it can be seen from the relatively low value of R^2 in Figure 6.3, there is no evidence of linear relationship in all the subplots, **which indicates an insufficient relationship between the four lexicons according to the produced score for written text**.

Besides, the result from Figure 6.3 confirms our previous statement that JSD value is not enough to measure the similarity between the distributions with different lexicons. As it can be seen that the pair of distributions with the highest JSD value is OL and LIWC, but the R^2 of the subplot in terms of OL and LIWC is not the highest one, which indicates that the rank of entities in terms of lexicon OL is not the most similar to the rank of entities in terms of LIWC.

As a supplementary explanation of Figure 6.3, the conspicuous lines inside the scatter plots are caused by a large number of short articles.

6.2.2. Sentiment Distributions for Wikipedia Talks

Similar to Wikipedia articles, an analysis has been carried out on Wikipedia talks with their total scores. This study focuses on the difference between the distributions of sentiment scores for the mentioned four lexicons. A total of 1,906,375 Wikipedia talks are involved in the analysis.

The histograms of total score for Wikipedia talks based on different lexicons can be compared in Figure 6.4.

It can be seen that the general pattern of the histograms for Wikipedia talks (as shown in Figure 6.4) is the same as the pattern of Wikipedia articles (see Figure 6.1, and the description of general pattern can be seen in Section 6.2.1). The distance between the lexicon with the highest median and the lexicon with the lowest median is larger than that of Wikipedia articles (see Figure 6.1). MPQA has the highest value for both median and mean (median=9.1429, mean=9.6561), LIWC has the second highest value (median=5.9701, mean=5.9641), the value of both median and mean for OL is close to LIWC which ranked in the third place (median=5.1724, mean=5.7299), and ANEW has the lowest value (median=1.4722, mean=1.5738), which is much lesser than MPQA.

In addition, the spikes in Figure 6.4 are more noticeable than in Figure 6.1. To provide an evidence of the statement that these spikes are caused by the large number of short text, a figure plotting the histograms of the total score of entities excluding short text has been shown in Appendix D.1.

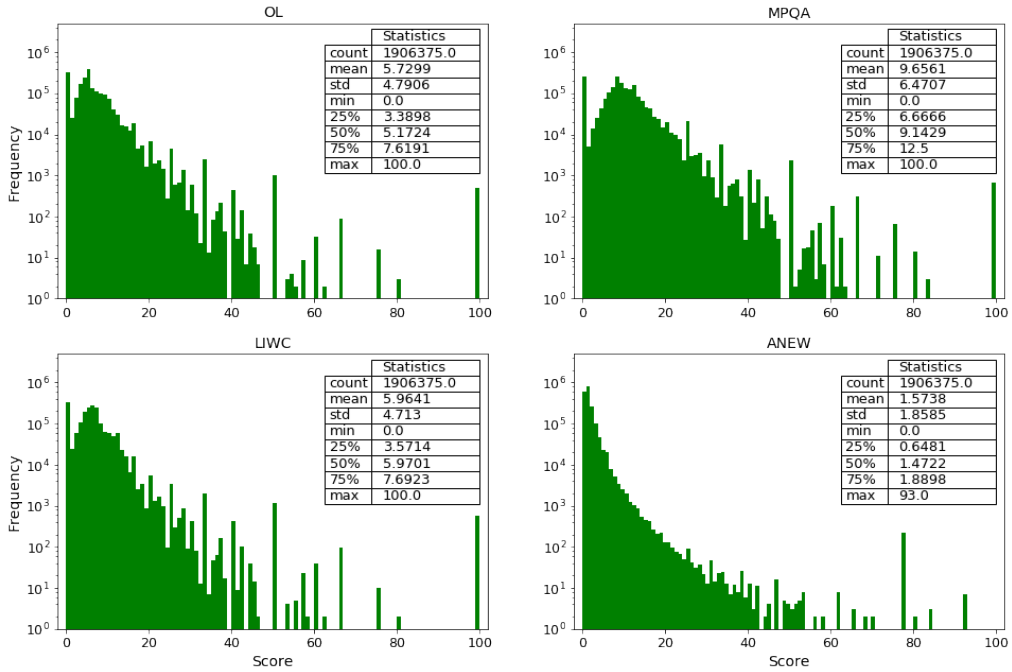


Figure 6.4.: **Distributions of total scores for the whole Wikipedia talks (1,906,375) based on OL, MPQA, LIWC, ANEW four lexicons:** The detailed description of the bins can be seen in Figure 6.2.

To look the difference between articles and talks closer, the score in Figure 6.4 has been compared with Figure 6.1 further. What is interesting in the comparison is that the median score for talks concerning OL, MPQA, and LIWC, is higher than that of the articles, whereas the median score for talks concerning ANEW is lower than that of articles. People usually use different expressions in writing and speaking. In Wikipedia text, articles correspond to written language, and talks correspond to verbal language. **Taken together, the results probably indicate that lexicons such as OL, MPQA, and LIWC are more sensitive for verbal language than written language. On the contrary, lexicons such as ANEW are much less sensitive for verbal language.**

This findings provide a suggestion about lexicon selection on sentiment analysis. If researchers want to analyse the emotion inside verbal text and want to dig out emotional expressions as much as possible, lexicon MPQA is recommended in the first place. If researchers want to exclude subjective emotions as far as possible, ANEW is the most recommended lexicon (ANEW has the lowest value for the score of verbal text). However, MPQA shows a considerable sensitivity for emotional words, for example, it regards *might* as a positive word, whereas ANEW shows a considerable insensitivity for emotional words. Both of them are likely to give extreme results. Therefore, if researchers want to find a respectively

balanced lexicon, OL and LIWC can be considered. The difference between these two lexicons is that OL is more sensitive than LIWC for written language, while LIWC is a bit more sensitive than OL for verbal language (according to the median scores of these two lexicons between articles and talks).

It is important to note that the measurement of the sentiment for talks is a bit different from that of the articles. While calculating the sentiment scores for talks, the frequency of each word has been corrected by the frequency of the word in the corresponding article. In order to examine the impact of this adjustment, we calculate the sentiment for talks again by using the same method as that of the articles (i.e. without correcting the frequency). The formulas used to replace Equation 5.8 and Equation 5.10 have been shown accordingly in Equation 6.1 and Equation 6.2. To differentiate between these two methods, we name the score with the correction of the frequency as corrected score, and name the score without correction of the frequency as pure score in the following text. An direct comparison between the pure score and corrected score has been presented in Figure 6.5. Each subplot presents score for entities with one lexicon by showing corrected score on the x-axis and pure score on the y-axis. The red line in each subplot denotes linear function $y = x$. Points located in the red line indicates that the corrected score and pure score of this entity are the same. In order to obtain the proportion of the difference, the percentage of entities, which have the identical corrected score to pure score, has been calculated and shown as p_1. The percentage of entities, which have a lower corrected score than the pure score, has been calculated and shown as p_2. Accordingly, the percentage of entities, which have 0 as the corrected score, can be obtained by subtracting p_1 and p_2 from 1.

$$score_i = \frac{ft_i}{n_t} \tag{6.1}$$

$$score_i = \frac{ft_i * vnorm_i}{n_t} \tag{6.2}$$

It can be observed from p_1 and p_2 in Figure 6.5 that, there are more than half of the entities that have identical corrected score to pure score (p_1=0.6548, 0.6080, 0.6892, 0.5754 for OL, MPQA, LIWC, and ANEW, respectively). From the cluster in the scatter plots, we can also see that, there is a substantial number of entities which have a corrected score near the pure score. Altogether, MPQA has the lowest value of R^2 (R^2=0.3646), and ANEW has the lowest value of p_1 (p_1=0.5754), which suggests that lexicon MPQA and ANEW have the least similarity between the corrected score and the pure score. To compare the distribution of pure score to the distribution of corrected score in Figure 6.4, Appendix D.2 presents the histograms of pure scores of Wikipedia talks with each of the above four lexicons. It shows that the distributions of pure scores are similar to the corrected scores, but with a lower value of the median score. **Taken together, the results suggest that there is no significant difference between the corrected score and the pure score.** The rest of the analysis

regarding Wikipedia talks will use the corrected score as we did before.

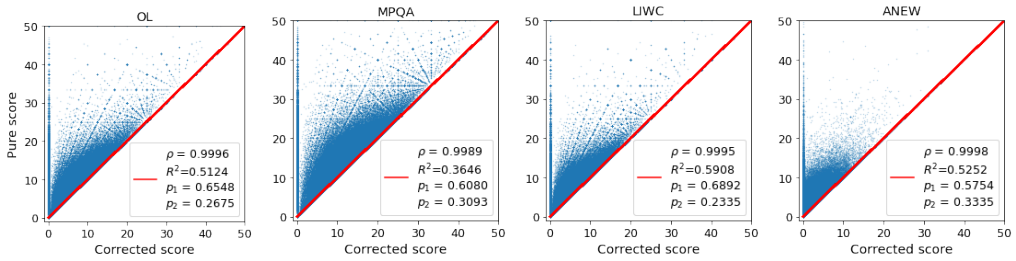


Figure 6.5.: **Direct comparison between corrected score and pure score for Wikipedia talks:** Each dot in the figure indicates an entity. The value on x-axis is the corrected score of this entity, and the value on y-axis is the pure score of this entity. Both axes of plots are with a cut-off at 50, and the percentage of remaining dots are presented with ρ in the legend. The red line in each plot indicates function y=x, and the corresponding coefficient of determination between overall data and the red line has been shown as R^2. Besides, p_1 denotes the percentage of dots whose corrected score is equal to pure score, and p_2 denotes the percentage of dots whose corrected score less than pure score and larger than 0.

To access the difference among the four distributions in Figure 6.4, the JSD values of each pair of them have been calculated and shown in Table 6.4.

	OL	**MPQA**	**LIWC**	**ANEW**
OL	0.0	0.1392 (4)	0.0266 (6)	0.3082 (3)
MPQA		0.0	0.1166 (5)	0.4381 (1)
LIWC			0.0	0.3232 (2)
ANEW				0.0

Table 6.4.: **Jensen-Shannon Divergence of each pair of sentiment distributions regarding the total score of the whole Wikipedia talks:** The sentiment of each distribution is calculated with one of OL, MPQA, LIWC, and ANEW lexicon. The values inside parentheses are the rank of JSDs in order of highest to lowest. Accordingly, it presents the order of similarity from the strongest to the weakest.

It can be seen from the table that, in terms of Wikipedia talks, the pair of lexicons has the lowest value of JSD is OL and LIWC (JSD=0.0266), and the pair of lexicons has the highest value of JSD is MPQA and ANEW (JSD=0.4381). **This result suggests, regarding verbal language, OL and MPQA are similar to each other the most, according to the sentiment score they produced, and ANEW and MPQA are**

different from each other the most. In addition, the most similar lexicon to MPQA is LIWC (JSD=0.1166), and the most similar lexicon to ANEW is OL (JSD=0.3082).

Data from this table can be compared with the data in Table 6.3, which shows a generally higher value of JSD of distributions for Wikipedia talks than articles. It seems to suggest that talks express greater difference among the four lexicons than articles. It is important to note that, this discrepancy could also be contributed by the difference in quantity of the entities, since the number of articles is around five times larger than the talks.

In summary, for both written language and verbal language, MPQA and ANEW are different from each other the most, and OL and LIWC are similar to each other the most. Lexicon OL is closer to MPQA for written language, but closer to ANEW for verbal language. LIWC is just the opposite.

As a further exploration of the similarity between the scores produced by different lexicons, Figure 6.6 shows a direct comparison between the total scores of each pair of lexicons in terms of Wikipedia talks. The more information of how the scatter plot represents the similarity between lexicons can be seen in the explanation of Figure 6.3.

It can be observed from the figure that, the highest value of R^2 is between OL and MPQA (R^2=0.6005), which indicates that the two lexicons sharing the nearest ranking of entities are OL and MPQA. On the contrary, the lowest value of R^2 is between MPQA and ANEW (R^2=0.1557), which indicates that the two lexicons sharing the least ranking of entities are MPQA and ANEW. Again, there is evidence of linear relationship found in Figure 6.6, **which suggests none of the pair of lexicons share a substantial similarity in the ranking of entities based on their scores for verbal text.**

Taken together, these results show that there is no significant association between each pair of these four lexicons. Relatively, lexicon OL and MPQA seem to assign entities the most similar scores, and lexicon OL and LIWC seem to generate the most similar distributions of scores. Lexicon MPQA is likely to have the highest sensitivity for both verbal and written text, and lexicon ANEW is likely to have the least sensitivity among the four lexicons, especially for verbal text.

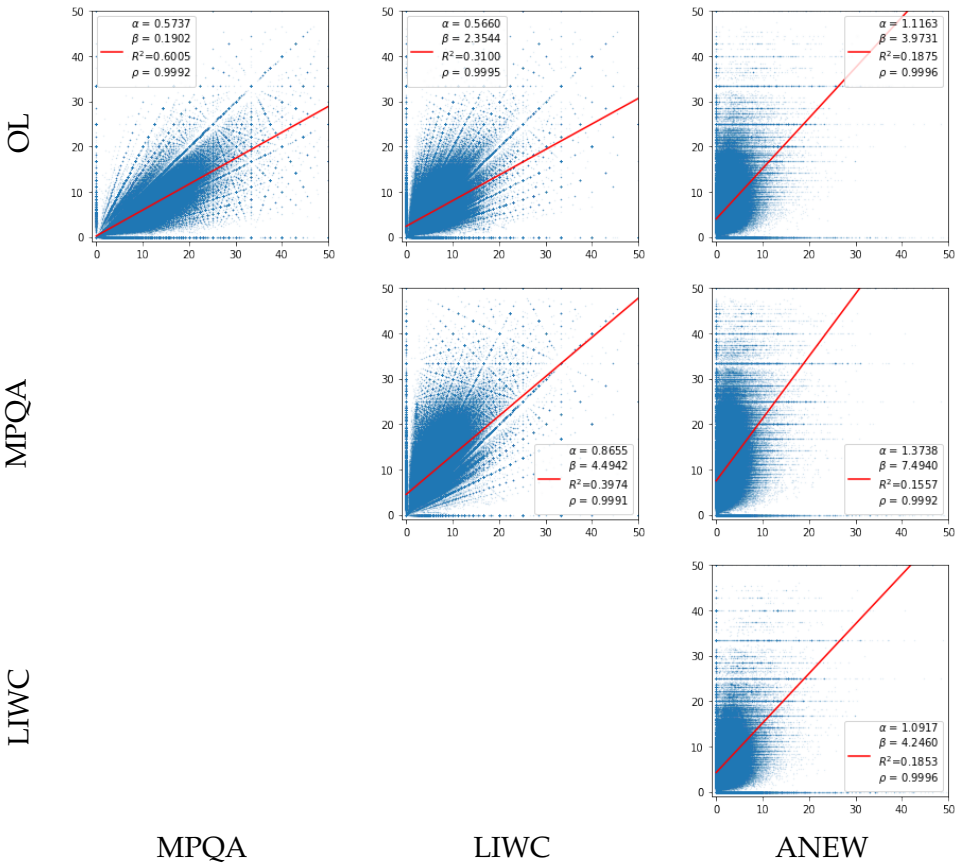


Figure 6.6.: **Direct comparison of each pair of lexicons regarding the total score of individual Wikipedia talks:** It shows scores of entities by displaying the score for one lexicon on the x-axis and the score for the other lexicon on the y-axis. The described four lexicons are involved in this figure, and each of them has been compared with the other three lexicons. After excluding the repeated ones, six scatter plots are displayed. The cut-off is set at 50 for both axes, and the percentage of remaining points are presented as ρ in the form of decimal number. The red line is the linear regression fitting to the observed points with α as slope, β as intercept, and R^2 as the coefficient of determination.

6.3. Sentiment Distributions with Varying Time

In this section, the sentiment distribution with varying time will be explored. In particular, the analysed entities will be focused on two categories: people and events. And the range of time has been set to around 100 years. By connecting people to their birth date, and connecting events to their occurrence date, a series of sentiment rivers on temporal dimension will be presented to see if there are any temporal

changes of sentiments for people and events on Wikipedia.

6.3.1. Number of Entities about People and Events with Varying Time

Before exploring the changes of general sentiments over time, it is necessary to examine the number of involved entities for each month. Since entities from talks are roughly a subset of entities from articles[4], we focus on the number of articles as the number of entities. Again, we take entities of people and events as an instance. Moreover, the number of subgroups of people classified by occupation and the number of subgroups of events classified by type are further investigated.

Figure 6.7 shows the number of articles describing people (i.e. biography) by month and Figure 6.8 shows the numbers of articles describing events.

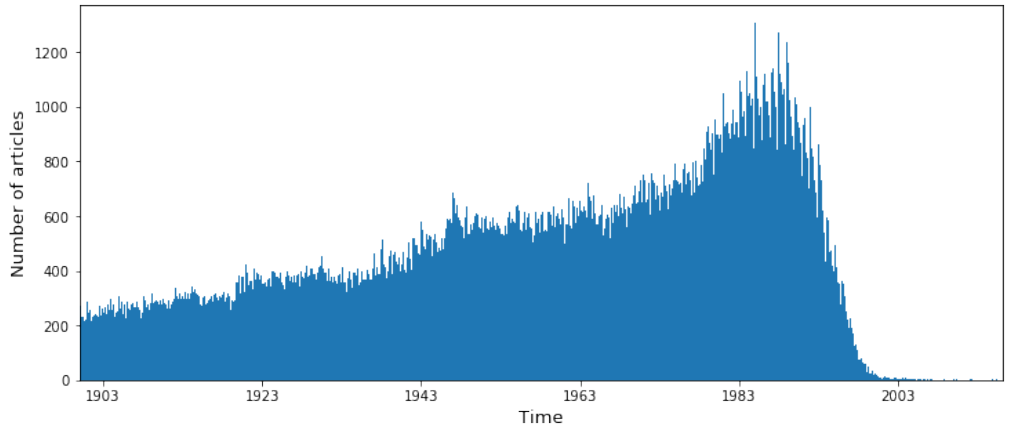


Figure 6.7.: **Number of articles describing people on Wikipedia who born from 1900 to 2016:** It shows the number of people on Wikipedia who born between 01.01.1990 and 30.04.2016. There is a total of 625,054 people included. The time is grouped by month.

The time is range from 01.01.1900 to 01.04.2016. The end date is decided by the version of used DBpedia dataset (April 2016) to increase the accuracy (the possibility exists that date has been assigned in error, especially for the date after April 2016 related to people). As we mentioned earlier, the timestamp for people is the birthday, and the timestamp for events is the occurrence date.

In Figure 6.7, we see that the growth in articles describing people levelled off around 1990, and transformed into a sharp decline until 2000, when the number of people became very less. As compared, the growth in articles describing events, as shown in Figure 6.8, is much slower than the growth of people. The number

[4]Basically, each article page has a corresponding talk page, but two special cases exist. First, talk pages which are empty have been excluded in this thesis. Second, some article pages have been removed (being merged or being redirected), while the talk pages are retained.

of events is generally stable with a slightly climb from 1900 to 2000, and it has an marked upswing from 2000. Besides, the trend in Figure 6.8 has three peaks and a series of regular spikes. The first peak is located around 1918, the second one is located around 1945, and the third one is the upswing in recent years.

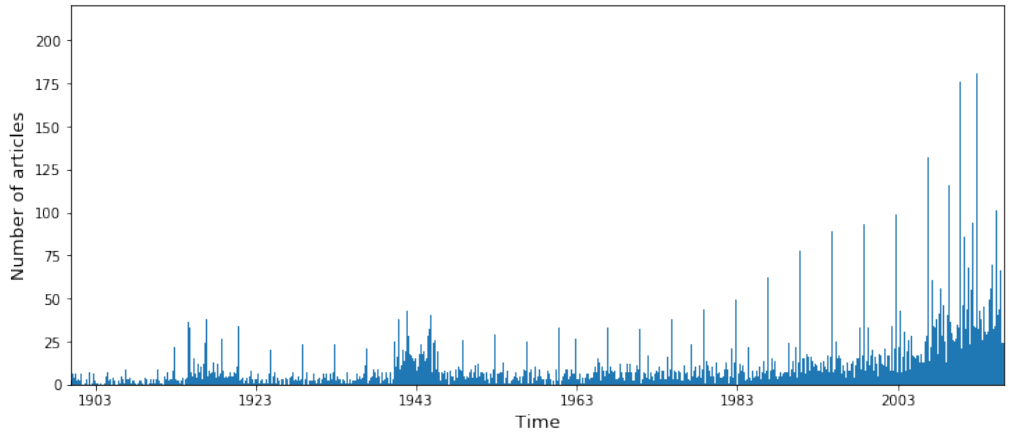


Figure 6.8.: **Number of articles describing events from 1900 to 2016:** It shows the number of events on Wikipedia between 01.01.1900 and 30.04.2016. There is a total of 16,355 events included. The time is grouped by month.

The sharp decline of the number of people in recent year might due to the fact that people with young age have less opportunity to be published on Wikipedia. The increasing of the number of events in the recent year might due to fact that events would be published on Wikipedia as soon as Wiki editors know it, even if it has not happened yet (e.g. annual activities). Also, the more and more activities nowadays are likely to contribute to the increase in events on Wikipedia. The peaks and the regular spikes will be investigated later by examining subgroups of events by type.

To validate the results of Figure 6.7 and 6.8, we investigate the noise inside these two figures. Noise here refers to dates which included in these two figures but do not match the dates on Wikipedia. It stems from the imperfect extraction of attributes from Wikipedia by DBpedia. With the noise, entities are incorrectly located in the date they do not belong to. While randomly checking the entities, most of people and events show the identical date to Wikipedia, except for some people born in the last several years of the plot. By looking into the details, we found that there exists a considerable number of erroneous date in terms of people born after 2006. The detailed analysis has been presented in Appendix C, in which we examine the entities of people between 2006 and 2017 by hand to identify correct entities from erroneous entities. Moreover, both correct and erroneous entities have been classified into finer categories for the sake of better understanding. The result concludes that, after excluding erroneous data, there are only a few people after 2000. It

makes no sense to carry out the analysis on dataset which contains substantial noise (e.g. person entities from 2007 to 2016), or dataset with insufficient size (e.g. person entities from 2000 to 2006). In other words, it is reasonable to ignore the sentiment analysis after 2000.

In order to gain an insight of the distribution of the number with varying time in terms of people and events, we take a closer look at the subgroup of people (grouped by occupation, such as politicians, writers, and actors) and events (grouped by type, such as wars, games, and elections).

Figure 6.9 compares the number of articles describing politicians, writers, and actors.

Apart from these three occupations, we also examined the number of people with other occupations, including scientists, players (e.g. football players), singers, directors, and producers. The result shows that, among people on Wikipedia who born from 1900 to 2016 with the above eight occupations, the number of actors is the highest, and the number of writers is the second highest. Singers, directors, and producers are ranked from third to fifth in similar numbers. The number of politicians is the sixth highest. Lastly, scientists and all kind of players are the two occupations with the least number of people.

As it can be seen in Figure 6.9, the number of actors is the highest, followed by the number of writers and politicians. The peak years for these three occupations are different. The number of politicians starts to rise from 1940, and reaches a peak around 1960. Then, the number drops steadily. From 1983, the number of politicians becomes considerably less. The latest born year of politicians is 1995. The number of writers is increasing until 1975, and then transformed into a decline. The peak year of it is around 15 years later than the peak year of politicians. The latest born year of writers is 2010. As compared, the peak year for actors is around 1985, which is 10 years later than the peak year of writers. The latest born year of actors is 2013. We can see that there has been a marked increase from 1963 in the number of actors, which might due to the development and spread of television and film in the late 20th century. From the peak year and the latest born year of these three subplots, it can be seen that actor could be someone with very young age, and writer could be young but not as young as an actor (e.g. a baby can be an actor, but it cannot be a writer). In contrast, politicians are usually elder people.

As we described earlier, several trends stand out in Figure 6.8: a peak near World War I (1914-1918), a peak near World War II (1939-1945), a upswing in recent years, and a series of regular spikes. From Figure 6.10, we can moreover glean that some specific types of events (wars, games, and elections) contribute to the these trends to a large extent. Specifically, war-related events contribute to the peaks near the two world wars, game-related events contribute to the marked rise in the number of events in recent years, and election-related events contribute to the regular spikes.

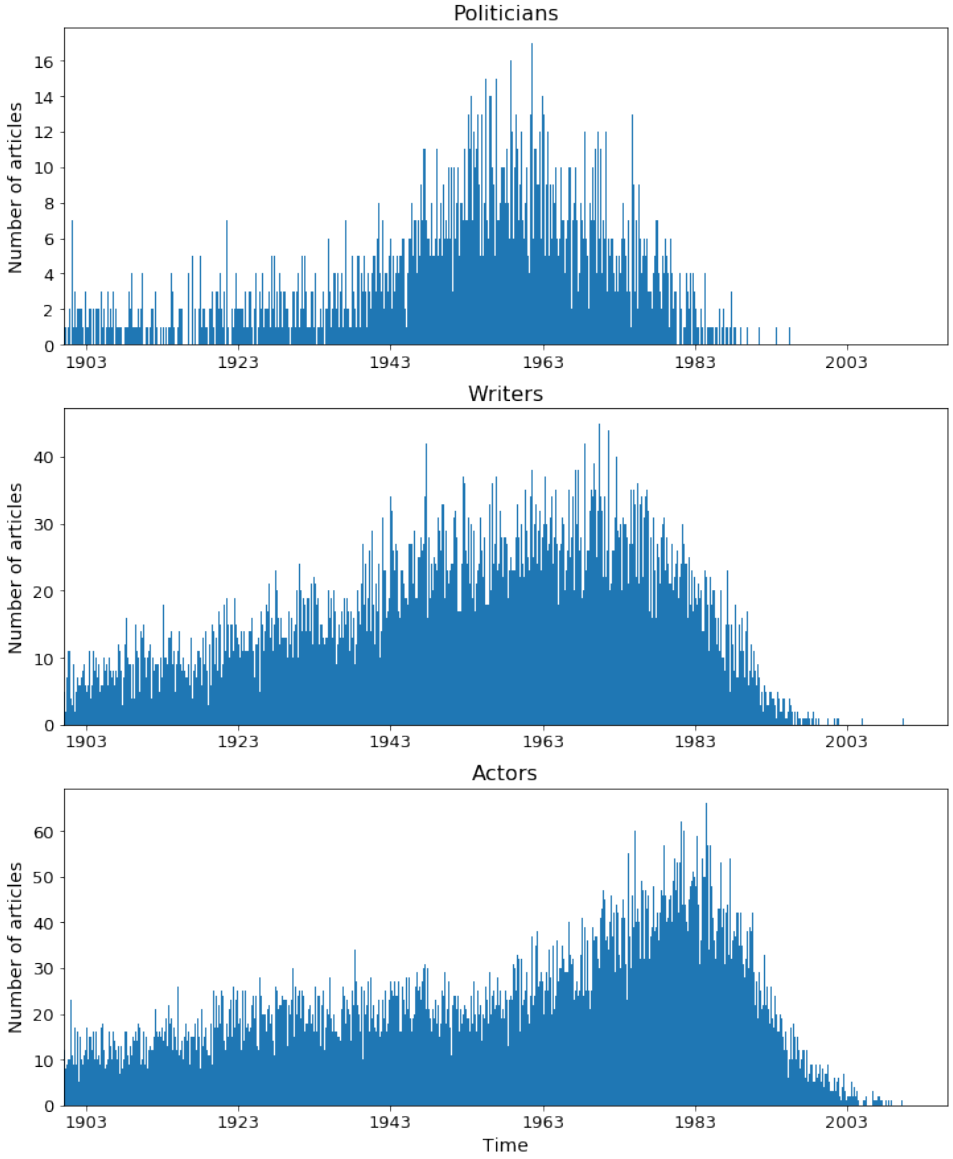


Figure 6.9.: **Comparison of the number of biographies describing politicians, writers, and actors by their born year:** There is a total of 4170 politicians, 20,382 writers, and 28,800 actors in the figure. All people are born between 01.01.1900 and 30.04.2016. The information of occupations is obtained from Infobox on Wikipedia, which means the data might be incomplete. In terms of *Actors*, both actors and actresses are included. In terms of *Writers*, all of writers, authors, novelists, and journalists are included.

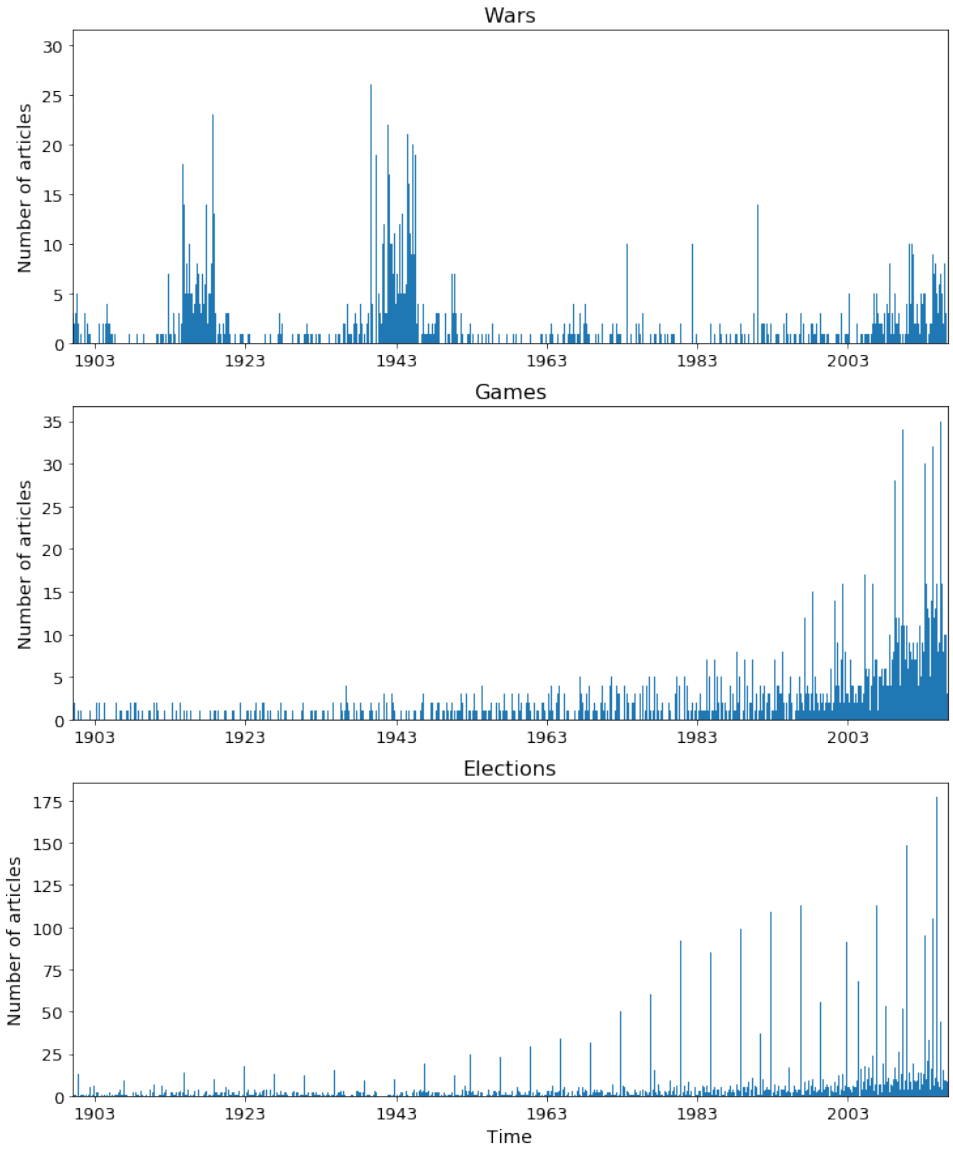


Figure 6.10.: **Comparison of the number of articles describing wars, games, and elections by their occurrence year:** There is a total of 2303 wars, 2806 games, and 7188 elections in the figure. All events are from 01.01.1900 to 30.04.2016. The information of types is obtained from the title of articles. Specifically, articles containing the term election or democratic (or republican) primary (or caucuses) in their title will be regarded as elections[5]. Articles containing one of the terms battle, war, offensive, attack, campaign, raid, occupation in their title will be regarded as wars. Articles containing one of the terms football, cup, champion, UFC (Ultimate Fighting Championship), final in their title will be regarded as games.

6.3.2. Changes of Sentiment Over Time for Wikipedia Articles

In this section, we will look into the sentiment distributions of Wikipedia concepts regarding articles with varying time, and try to find out the factors contributing to the results. The categories of concepts will focus on people and events as described in the previous section.

Figure 6.11 compares the sentiment rivers[6] regarding people and events against four lexicons. It reveals the changes of sentiment over time by plotting the median score together with 25th and 75th percentile as the sides of the river for each month from 01.01.1990 to 30.04.2016. In order to decrease the fluctuation of the river, the rolling average has been used on the original median river. Since the median score for each month will explain better together with the total number of involved entities to this month, the counting of involved entities has been shown in front of the sentiment river.

To state specifically about the sentiment river that, the upper and lower side of the river indicate the rolling average of 75th and 25th percentile respectively, and the red line inside the river indicates the rolling average of the median. The dotted line in green is a baseline meant to observe the rise and fall of the trend. The horizontal value of this line represents the median of scores for all entities involved in the current river. The rolling average takes 31 as window size, and sets the label at the center of the window. In other words, the median score for each month has been calculated first, and the horizontal value in the red line is the average of the medians from the nearest 31 months.

As shown in the figure, there is a significant difference between the movement of the sentiment rivers for people and events. The movement of the river for events is more fluctuated, whereas the movement of the river for people is steady except for the rightmost part. The sentiment river for people becomes more and more narrow and fluctuating after 2000. However, this part could be ignored in our analysis considering two aspects as discussed in Section 6.3.1 that, on the one hand, the number of people is very less from 2000, on the other hand, from 2006 there is a considerable number of people with an erroneous birth date.

Comparing the sentiment rivers against the four lexicons, it can be seen that the river of MPQA is the widest, and the horizontal value of it is the highest. In contrast, the river of ANEW is the narrowest, and the horizontal value of it is the lowest. Relatively, lexicons OL and LIWC have the similar widths and horizontal values.

Analysis of Person Entities

From the trends of rivers for people, we can see that there has been a slight fall in the sentiment from 1900 to 2000 regarding MPQA, OL, and ANEW, whereas a slight rise in the sentiment regarding LIWC.

[5] Democratic (or Republican) primary (or caucuses) is usually held to choose delegates for states for the sake of later presidential elections.

[6] It uses a river metaphor to represent sentiment changes over time in this thesis.

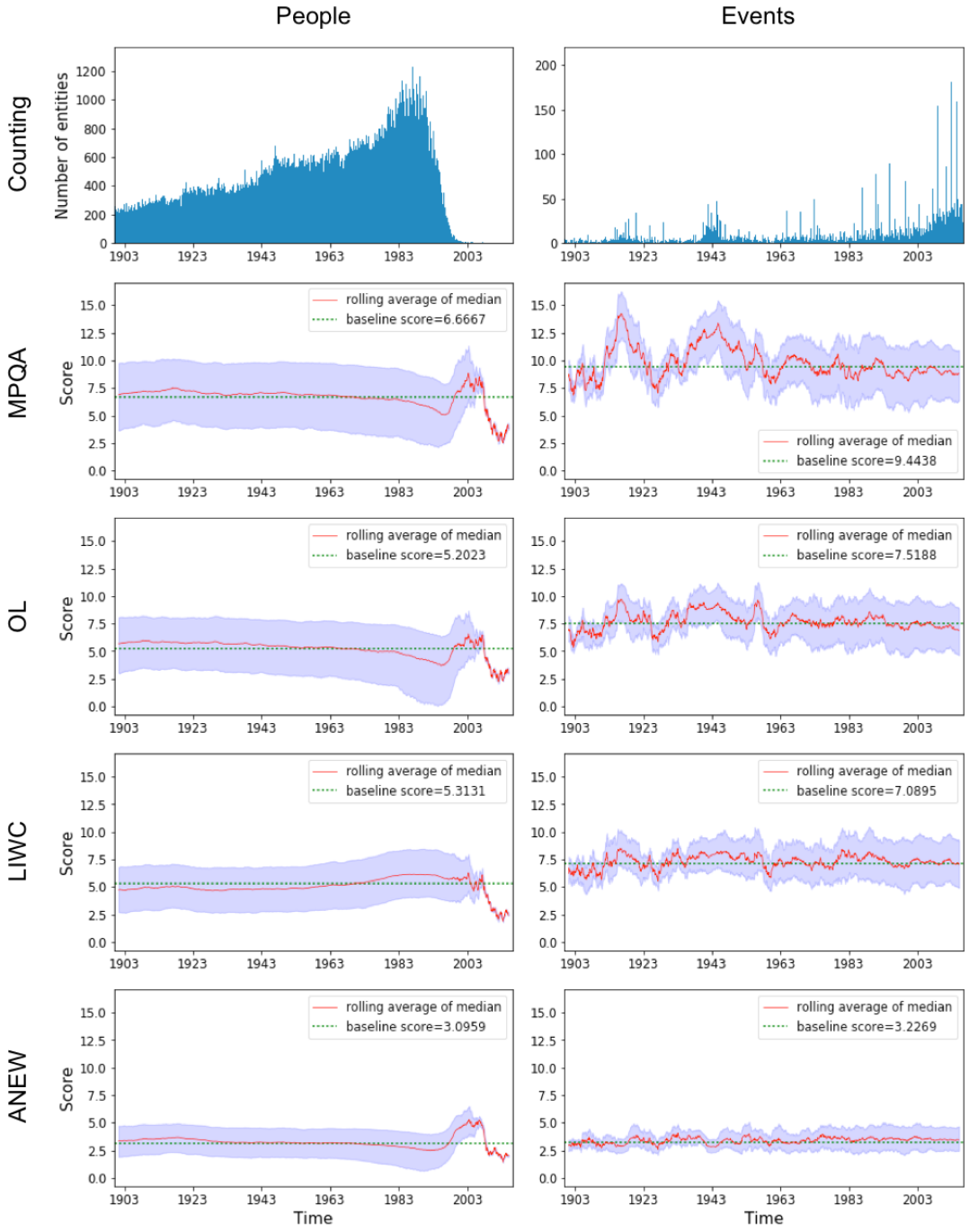


Figure 6.11.: **Comparison of sentiment rivers for Wikipedia articles regarding people and events with different lexicons:** Each sentiment river shows the rolling average of 25th, 50th, and 75th percentile of the total score for articles by month.

In order to find out the factors contributing to the sentiment, Multiple Linear Regression Model (MLR) has been built for each of the sentiment rivers. In terms of entities of people, four models have been built, each of which relates to one of the four lexicons. Table 6.5 shows the main parameters of each linear regression model.

	MPQA	OL	LIWC	ANEW
Time	-2.824e-05	-3.616e-05	5.403e-05	-2.157e-05
Length	0.0036	0.0027	0.0009	0.0011
Male by Gender	0.4045	0.0747	0.0602	0.1780
Female by Gender	0.8039	0.4182	-0.7237	0.8704
Writer by Occupation	1.6918	1.3909	-0.0044	0.8672
Actor by Occupation	1.3976	1.2660	0.5284	0.6913
Politician by Occupation	-0.9703	-0.8516	-0.0839	0.9197
Musician by Occupation	0.9298	0.9923	-0.3415	1.0071
Scientist by Occupation	0.1502	0.2134	-1.0874	-0.4848
Asia by Geolocation	-0.6185	-0.7184	-0.5924	-0.0570
Europe by Geolocation	0.2240	0.2641	0.1441	0.3923
Africa by Geolocation	-0.1732	0.1881	-0.5425	-0.1452
South America by Geolocation	-0.1419	0.2516	-1.5788	-0.1855
North America by Geolocation	-0.1291	-0.1723	-0.3633	0.8502
Oceania by Geolocation	0.3522	0.8798	1.0981	0.6991
Intercept	25.9262	30.3903	-33.3447	18.2676
R^2 **of the Model**	0.1120	0.0970	0.0290	0.0610

Table 6.5.: **Coefficients of predictors and other related parameters for MLR models built for sentiment score with the four lexicons in terms of the biography on Wikipedia:** The parameters of each model reveals to what extent the predictors contribute to the total score for each article of biography. The total score of individual entities is regarded as the response for the model. For each model, it shows R^2, intercept, and coefficients for all predictors. There are five aspects considered as predictors, which are time (i.e. birth date), length (i.e. the number of tokens in the text), gender, occupation, and geolocation. It is important to note that, the extracted attributes are incomplete. There is a total of 623,054 biographies fed to each model, all of them have information about birth date and length, 1,551 of them have information of gender, 57,727 of them have information about occupation, and only 393 of them have information about geolocation. Before building the model with those predictors, the correlation between predictors has been calculated, and no significant relationships were observed. All the MLR models are built with Statsmodels library.

In each model, five aspects have been considered as predictors, which are time (i.e. birth date), text length, gender, occupation, and geolocation. With regard to

occupation, there are five occupations being considered: writer, actor, politician, scientist, and musician. With regard to geolocation, there are six continents included, which are Africa, Asia, Europe, North America, South America, and Oceania. The categorical variables such as gender, occupation, and geolocation, are processed By using the dummy variable.

From the R^2 of the model it can be seen that the MRL model is unable to fit the sentiments in a good way. With respect to the predictors, almost all of the coefficients are in a low value. The predictor having the highest value of the coefficient for the model of MPQA is *Writer by Occupation* (1.6918), and for the model of OL it is *Writer by Occupation* as well (1.3909). Regarding the model of ANEW, the predictor having the highest coefficient is Politician by Occupation (0.9197). The level of coefficients regarding different predictors for the above three lexicons is roughly consistent, which means if a predictor has a relatively higher coefficient in one lexicon, it will be relatively higher in the other two lexicons as well. Differ from the above three lexicons, the model of LIWC shows a relatively low coefficient for *Writer by Occupation* (-0.0044), and shows a relatively high coefficient for *South America by Geolocation* (-1.5788) that other lexicons give a really low coefficient to it (MPQA: -0.1419, OL: 0.2516, ANEW: -0.1855). Besides, the coefficients regarding time also reveal a significant difference between LIWC and the other three lexicons, by showing a positive value for LIWC (5.403e-05) and negative values for the other three lexicons (MPQA: -2.824e-05, OL: -3.616e-05, ANEW: -2.157e-05). The value of the coefficient for time is much smaller than the other coefficients because we replaced the variables of time with a series of ordinal numbers, and each of those ordinal numbers is a six-digit number. This result is consistent with our observations from Figure 6.11 that **the trend of score regarding LIWC is slightly downward over time, while the trend of score regarding the other three lexicons is slightly upward over time.**

Analysis of Event Entities

In terms of the sentiment rivers of events, it is apparent that the sentiment with MPQA reaches noticeable peaks during both World War I and World War II while the sentiments with the other three lexicons only have small peaks at the same time (see Figure 6.11). In order to look up the reason for this difference, the positive and negative sentiments have been examined separately. The result shows that all the four lexicons present noticeable peaks during the two world wars in terms of negative sentiment. However, there are events with positive sentiment in where the negative sentiment is low. In the case of MPQA, the negative score for wars is much higher than the positive score for other events. As a result, it shows a relatively high peak during each of the two world wars. In the case of the other three lexicons, the negative score for wars is comparable with the positive score for other events. As a result, the peaks during the two world wars are much smaller. In other words, the contrast of the peaks between different lexicons is mainly due to their divergent sensitivity to events.

Except for the two world wars, there are some other small peaks shown in the plot. By looking into the details, most of the peaks generated by negative sentiments are because of wars, such as Philippine-American War (1899-1902), Russo-Japanese War (1904-1905), First Balkan War (1912-1913), and Cyprus Emergency (1955-1959). Peaks generated by positive sentiments are mostly due to festival-related and election-related events. Berlin International Film Festival[7] is a typical example. This festival was founded in 1951, and it contributes to some small peaks between 1954 and 1957 because of its relatively high positive score together with the absence of other events in the same month[8]. Similarly, elections tend to be positive events, and contribute to small peaks occasionally especially when no other events presented in the same month. Furthermore, by exploring some certain periods, such as the period from 1932 to 1934, we found that different lexicons behave differently. Specifically, while working with ANEW, the scores for elections are usually higher than the scores of other types of events such as battles[9] or games (e.g. Football League Championship). While working with MPQA, the scores for battles are usually higher than other types of events. While working with OL, all of elections, games, and earthquakes are valued with high score. While working with LIWC, all of elections, games, and battles are valued with high score.

The above findings suggest that ANEW is more sensitive for elections, MPQA is more sensitive for battles, OL is more sensitive for elections, games, and earthquakes, and LIWC is more sensitive for elections, games, and battles. In order to get more reliable evidence instead of proposing suggestions from the sample of a randomly picked period, the sum of scores for wars, games, elections, earthquakes, and festivals has been calculated separately from 1900 to 2016, and their ranking has been compared in Table 6.6.

It can be seen from the table that MPQA is more sensitive than the other three lexicons regarding events of wars, elections, and earthquakes, while LIWC is the most sensitive lexicon regarding events of games and festivals. However, the sum presented above relates to the number of involved entities, which is insufficient to compare the sensitivity for a certain lexicon to different types of events. In order to see if a certain lexicon is more sensitive for some specific types of events rather than others (e.g. seeing whether MPQA expresses more emotions to wars rather than other types), the score of certain types of events has been calculated together with the score of overall events by their average, as shown in Table 6.7.

[7] `https://en.wikipedia.org/wiki/Berlin_International_Film_Festival`, as seen on Aug. 14, 2018

[8] According to the description of sentiment river, if there is only one entity located in a certain month, the score of this entity will be the median score of this month and thereby being plotted on the river.

[9] We regard *battles* and *wars* the same in this thesis while discussing the type of events.

	Wars	Games	Elections	Earthquakes	Festivals
No. 1	MPQA	LIWC	MPQA	MPQA	LIWC
	(31216.9166)	(21881.1666)	(58029.7169)	(5995.5772)	(3555.2358)
No. 2	OL	OL	OL	OL	OL
	(21403.0316)	(20962.3900)	(43759.1276)	(5396.4347)	(2759.8018)
No. 3	LIWC	MPQA	LIWC	LIWC	MPQA
	(20900.3769)	(20764.9499)	(43482.7120)	(4189.8783)	(2678.6329)
No. 4	ANEW	ANEW	ANEW	ANEW	ANEW
	(7057.3651)	(9369.6413)	(24664.9681)	(2629.0737)	(771.0742)

Table 6.6.: **Ranking of lexicons according to the sum of scores for each type of events regarding Wikipedia articles:** The entities involved in this table are events occurred from 01.01.1900 to 30.04.2016 with five specific types. In total, 16,355 entities are included. Amongst them, there are 2303 war-related events, 2806 game-related events, 7188 election-related events, 591 earthquake-related events, and 370 festival-related events. The type of entities is identified by the keywords in their titles. The sum is written inside parentheses.

	Overall	Wars	Games	Elections	Earthquakes	Festivals
MPQA	9.2467	13.5549	7.4002	8.0731	10.1448	7.2395
OL	7.2753	9.2935	7.4706	6.0878	9.1310	7.4589
LIWC	7.1820	9.0753	7.7980	6.0493	7.0895	9.6087
ANEW	3.4387	3.0644	3.3391	3.4314	4.4485	2.0840

Table 6.7.: **Average of scores for both overall events and some certain types of events in Wikipedia articles in terms of four lexicons**: The involved entities are the same as entities in Figure 6.6

Table 6.6 and 6.7 provide an insight into the sensitivity of the four lexicons regarding different types of events. From Table 6.6 it can be seen that ANEW is in the last place of the ranking for all types of events according to the sum of scores. From Table 6.7 it can be seen that, regarding ANEW, the average score of earthquakes is higher than the average score for overall events. It indicates that **ANEW is more sensitive to earthquakes than other types of events**, even though it is much less sensitive to events than other lexicons. With regard to elections, the score with ANEW is close to the average score for overall events, whereas the scores with the other three lexicons are all lower than their average score for overall events. It indicates that **ANEW has a general sensitivity to election-related text, while the sensitivity of the other three lexicons is relatively low. With regard to war-related text, ANEW seems to be the least sensitive lexicon**, depending on the fact that the average score with ANEW is lower than its average score for overall events, while the scores with

the other lexicons are all higher than their average score for all events, **especially with respect to MPQA, which has an extremely high score for wars. With regard to festivals, both OL and LIWC show higher sensitivity than other events, and particularly the score with LIWC is much higher than its average score for overall events.** On the contrary, both MPQA and ANEW show a lower score for festival-related events than other events.

Besides, four MLR models have been built to explore the factors contributing to the sentiment of events shown in Figure 6.11 with respect to the four lexicons.

	MPQA	OL	LIWC	ANEW
Time	-7.499e-05	-3.841e-05	-2.803e-05	-3.188e-06
Length	0.0011	0.0006	0.0002	-3.867e-05
Earthquake by Type	-0.3483	-0.1629	-0.1832	-0.0276
Election by Type	0.0685	0.1131	0.0081	-0.0160
Festival by Type	0.2726	0.0323	0.0357	0.0763
Game by Type	-0.0241	0.0071	0.0121	-0.0040
War by Type	0.0962	0.1619	0.1177	0.0489
Asia by Geolocation	2.4935	1.7761	1.0439	0.1923
Europe by Geolocation	2.6403	1.5077	0.7936	-0.3893
Africa by Geolocation	3.1660	1.7076	1.4283	-0.1885
South America by Geolocation	1.0116	1.2121	0.1115	0.5747
North America by Geolocation	1.3739	1.7540	-0.0330	0.9051
Oceania by Geolocation	0.5502	0.8629	0.0272	-0.3395
Intercept	62.7094	34.5374	27.2467	5.7668
R^2 **of the Model**	0.1520	0.0750	0.0040	0.0610

Table 6.8.: **Coefficients of predictors and other related parameters for MLR models built for sentiment score with the four lexicons in terms of events on Wikipedia:** The parameters of each model reveals to what extent the predictors contribute to the total score for each article of events. The total score of individual entities is regarded as the response for the model. For each model, it shows R^2, intercept, and coefficients for all predictors. There are four aspects considered as predictors, which are time (i.e. occurrence date), length (i.e. the number of tokens in the text), type, and geolocation. It is important to note that, the information of type and geolocation is incomplete. There is a total of 16,355 events fed to each model, all of them have information about occurrence date and length, 13,168 of them have information of type, and 1,687 of them have information about geolocation. Before building the model with those predictors, the correlation between predictors has been calculated, and no significant relationships were observed.

In each model, four aspects have been considered as predictors, which are time,

length, type, and geolocation. Time refers to the occurrence date of the event. The explanation of length and geolocation are identical to the predictors in the MLR models for entities of people. With regard to type, it includes earthquake, game, war, election, and festival. By using dummy variables to replace type and geolocation, the linear regression model has been built, and the coefficients of predictors and other related parameters are shown in Table 6.8.

It can be seen from both the coefficients and R^2 that there is no notable linear relationship between predictors and the score. Most coefficients for predictors are too low to analyse. The coefficients for geolocation are higher than coefficients for other predictors, which might suggest a higher score for events occurred in certain locations, such as events occurred in Africa seem to get a higher score with a value of coefficient as 3.1660. **With regard to the time, coefficients for the four lexicons are all negative, which indicates the slight decrease of the score over time.**

So far the sentiment distributions with varying time of both people and events on Wikipedia have been explored according to their article pages. The following analysis will proceed to talk pages to see if there is any different characters between verbal language and written language.

6.3.3. Changes of Sentiment Over Time for Wikipedia Talks

In order to carry out the analysis for talks, Figure 6.12 shows the sentiment rivers for people and events separately with four different lexicons, to present the 25th, 50th, and 75th percentile of sentiment score as same as Figure 6.11. As comparison, it use the same setting as Figure 6.11 regarding rolling average and time range. Detailed explanation of how the river presents sentiment can be seen in the explanation of Figure 6.11.

Analysis of Person Entities

From the figure, we can see that the trends of sentiment rivers for people are steady before 1995. As shown in the subplot of counting of people, 1995 is about the time that the number of people starts to less than 100. Thus the exceptional value of sentiment after 1995 might be due to the small number of entities. In addition, the 25th percentile of sentiment has a new low between 1913 to 1918.

By probing into the details, we found that two facts might be contributing to the result. First, there is a considerable number of talk pages about people having very short text. As we described earlier, a short text might lead to extremely high or low sentiment score. It shows that the low becomes lighter after filtering out short text[10].

[10]It has been examined by using WikiSentiFlow, one of the proposed visualisation widgets in this thesis.

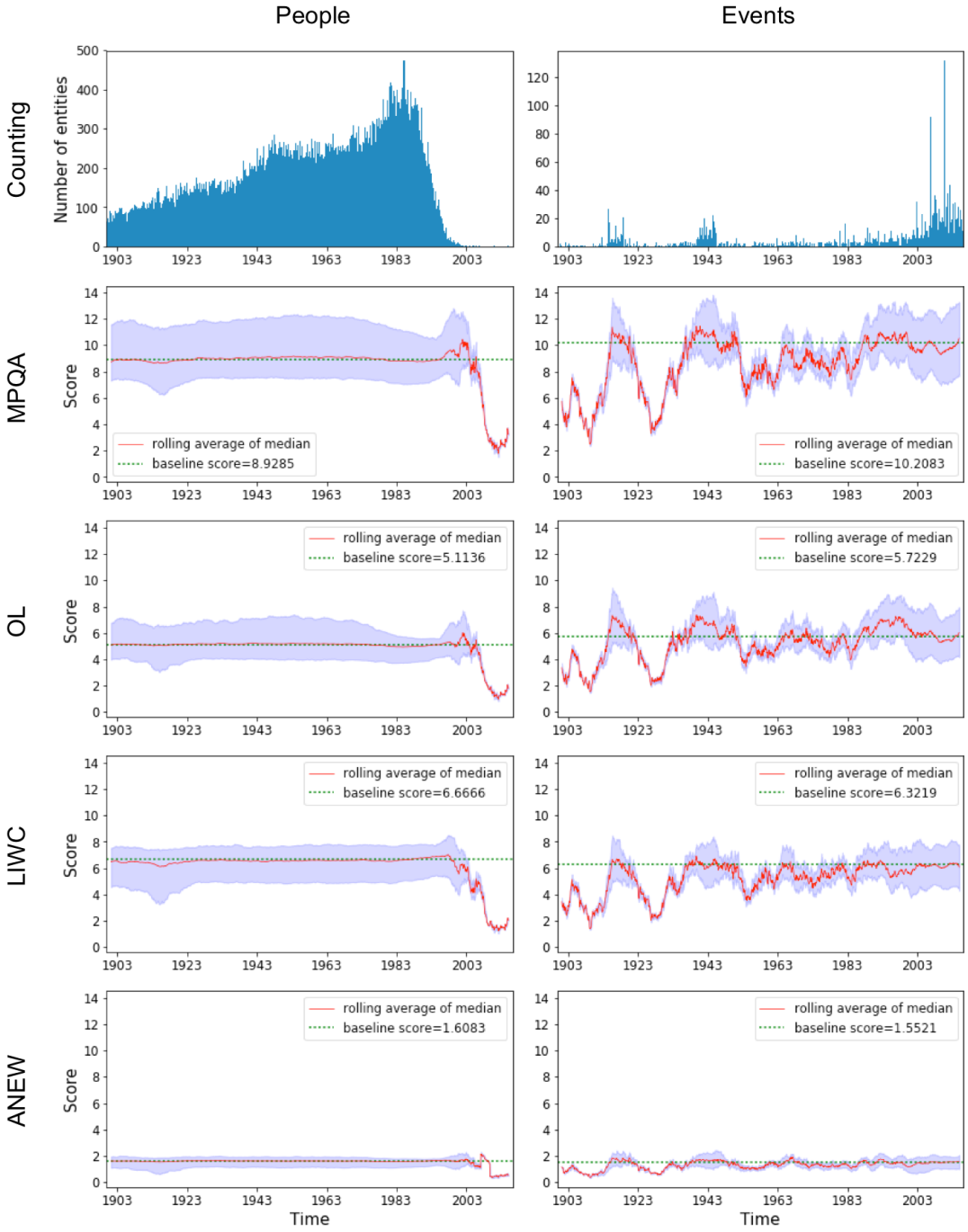


Figure 6.12.: **Comparison of sentiment rivers for Wikipedia talks regarding people and events with different lexicons:** Each sentiment river shows the rolling average of 25th, 50th, and 75th percentile of the total score for talks by month.

Second, a substantial amount of people have been changed or merged to another item, thus the article of which become a redirected page, and the talk page has accordingly been removed. However, the URL of their original talk pages have remained. Since most of the time the content of this kind of talk pages makes less sense than the content of the talk page of the new article page the original one redirecting to, we skip the calculation of this kind of talk pages and assign 0 as their score. A typical example is the recipients of the Knight's Cross. Almost all recipients of Knight's Cross have been merged to entity "List of Knight's Cross of the Iron Cross recipients". Therefore the recipients born from 1913 to 1918 pulled down the overall sentiment of that time.

Analysis of Event Entities

With regard to events, the peaks of the plots can be seen more noticeably for all the four lexicons, especially during the two world wars. It indicates the proportion of sentiment in talks is higher than the proportion in articles. In order to see the sensitivity of lexicons to different types of events is the same as articles, the sum of total scores for events regarding wars, games, elections, earthquakes, and festivals from 1900 to 2016 has been calculated separately. Figure 6.9 shows the result of ranking.

	Wars	**Games**	**Elections**	**Earthquakes**	**Festivals**
No. 1	MPQA	MPQA	MPQA	MPQA	MPQA
	(17835.3385)	(8318.8383)	(27298.4557)	(3581.1930)	(1357.6463)
No. 2	OL	LIWC	LIWC	OL	LIWC
	(11035.3466)	(5501.0746)	(16740.3346)	(2346.1384)	(1202.2275)
No. 3	LIWC	OL	OL	LIWC	OL
	(10509.2530)	(4841.3683)	(15803.1777)	(2223.7888)	(816.5931)
No. 4	ANEW	ANEW	ANEW	ANEW	ANEW
	(2812.5085)	(1337.1381)	(4113.6058)	(617.4157)	(213.4432)

Table 6.9.: **Ranking of lexicons according to the sum of scores for each type of events regarding Wikipedia talks:** The entities involved in this table are events occurred from 01.01.1900 to 30.04.2016 with five specific types. In total, 7323 entities are included. Amongst them, there are 1547 war-related events, 823 game-related events, 2562 election-related events, 374 earthquake-related events, and 139 festival-related events. The type of entities is identified by the keywords in their titles. The sum is written inside parentheses.

Differ from the result in Table 6.6, MPQA is the top-ranking lexicon for all the five types of events. OL and LIWC are ranked the second or third place, and ANEW are ranked the fourth place.

Furthermore, the average scores of both certain type of events and overall events have been calculated in Table 6.10.

	Overall	Wars	Games	Elections	Earthquakes	Festivals
MPQA	10.7518	11.5290	10.1079	10.6551	9.5754	9.7672
OL	6.5110	7.1334	5.8826	6.1683	6.2731	5.8748
LIWC	6.5495	6.7933	6.6842	6.5341	5.9460	8.6491
ANEW	1.6681	1.8180	1.6247	1.6056	1.6508	1.5356

Table 6.10.: **Average of scores for both overall events and some certain types of events in Wikipedia talks in terms of four lexicons**: The involved entities are the same as entities in Figure 6.6

It can be seen from the table that these four lexicons locate their emphasis on emotions of different type of events. With regard to wars, the average scores of all the four lexicons are higher than the average score of overall events, which indicates all the four lexicons express stronger sensitivity on text describing wars. With regard to games, the score is lower than the average score of overall events for most of the lexicons, except for lexicon LIWC. It seems possible that most lexicons express less sensitivity while describing games except for lexicon LIWC. With regard to elections, the average scores for all the four lexicons are only slightly lower than the average score of overall events, which indicates the emotions of verbal text describing elections is moderate. With regard to earthquakes, the average scores for most of the lexicons are lower than the overall one, except for ANEW that the score for ANEW is almost the same as the overall one. It is likely to suggest that most lexicons express less sensitivity to verbal text describing earthquakes. With regard to Festivals, the score for most lexicons are lower than the overall one, except for lexicon LIWC with a considerably higher score than the overall one. This result probably indicates that LIWC is more sensitive to verbal text describing festivals.

The finding from Table 6.7 can be compared and merged with the finding from table 6.10, which reveals the sensitivity for the mentioned four lexicons regarding certain types of events. **While processing text describing wars, MPQA expresses the highest sensitivity for both verbal and written text, whereas ANEW expresses relatively low sensitivity especially for written text. Both OL and LIWC express a moderate sensitivity. While processing text describing games, LIWC shows a relatively high sensitivity for both written and verbal text, and OL shows a slightly higher sensitivity than average only for written text. While processing text describing elections, there is no significant sensitivity for all the four lexicons. While processing text describing earthquakes, MPQA, OL, and ANEW express relatively higher sensitivity only for the written text. While processing text describing festivals, LIWC presents the highest sensitivity for both verbal and written text, and OL presents a relatively high sensitivity only for written text.** Those findings could provide further suggestions for selecting lexicons in sentiment

analysis.

Analysis of the High Frequency Sentiment Words

Considering that verbal text is very different from written text, we extract the top 100 entities with the highest sentiment score in Wikipedia talks, and count the occurrence of each sentiment word from them, to see the most frequent words presented on the talk page. It helps to reveal the characteristics of the high-frequency words which contributing to the sentiment score the most. Table 6.11 shows the top 10 positive words and Table 6.12 shows the top 10 negative words by the number of their occurrence.

	MPQA	**OL**	**LIWC**	**ANEW**
No. 1	right (121)	right (122)	please (66)	party (280)
No. 2	please (43)	victory (38)	thank (29)	leader (261)
No. 3	thank (27)	like (22)	thanks (16)	victory (81)
No. 4	progress (24)	work (19)	party (11)	win (60)
No. 5	like (12)	good (18)	like (11)	people (50)
No. 6	white (9)	reliable (15)	better (10)	green (48)
No. 7	correct (9)	defeat (14)	good (8)	time (46)
No. 8	reason (8)	correct (13)	honour (6)	event (30)
No. 9	good (8)	clear (11)	sure (5)	present (27)
No. 10	sure (7)	enough (9)	free (4)	opinion (27)

Table 6.11.: Top 10 positive tokens collected from the top 100 talks with the highest sentiment score

	MPQA	**OL**	**LIWC**	**ANEW**
No. 1	need (21)	failed (18)	battle (17)	war (208)
No. 2	massacre (8)	offensive (17)	problem (7)	seat (109)
No. 3	wrong (7)	loss (13)	war (5)	massacre (58)
No. 4	battle (7)	error (13)	sorry (5)	lost (24)
No. 5	little (6)	casualty (12)	wrong (5)	death (20)
No. 6	casualty (6)	suffered (9)	disagree (5)	army (19)
No. 7	war (5)	dead (8)	fighting (5)	bomb (15)
No. 8	sorry (5)	massacre (8)	loss (3)	failure (12)
No. 9	death (5)	lost (8)	mistake (3)	crime (11)
No. 10	rumor (5)	wrong (8)	forbidden (3)	alone (8)

Table 6.12.: Top 10 negative tokens collected from the top 100 talks with the highest sentiment score

From Table 6.11 it can be seen that the common used words in ANEW are quite

different from the other three lexicons. More specifically, it shows a large proportion of verbal words (such as thank, please, and sure) in MPQA, OL, and LIWC. These words are probably used as a response to the comments of others, and barely relate to the entity this text is describing for. Similarly, there is a considerable number of this kind of words being included in the top 10 negative tokens (such as wrong, sorry, and disagree), as shown in Table 6.12. These results suggest that the text of talk pages mixes a considerable number of conversation words (words used to complete the conversation) with opinions.

Comparison of Sentiment Distributions for Entities of People and Events

Figure 6.13 shows the heatmap of the JSD values between eight sentiment distributions for four lexicons regarding two domains respectively. It aims to measure the similarity of sentiment distributions for entities of both people and events.

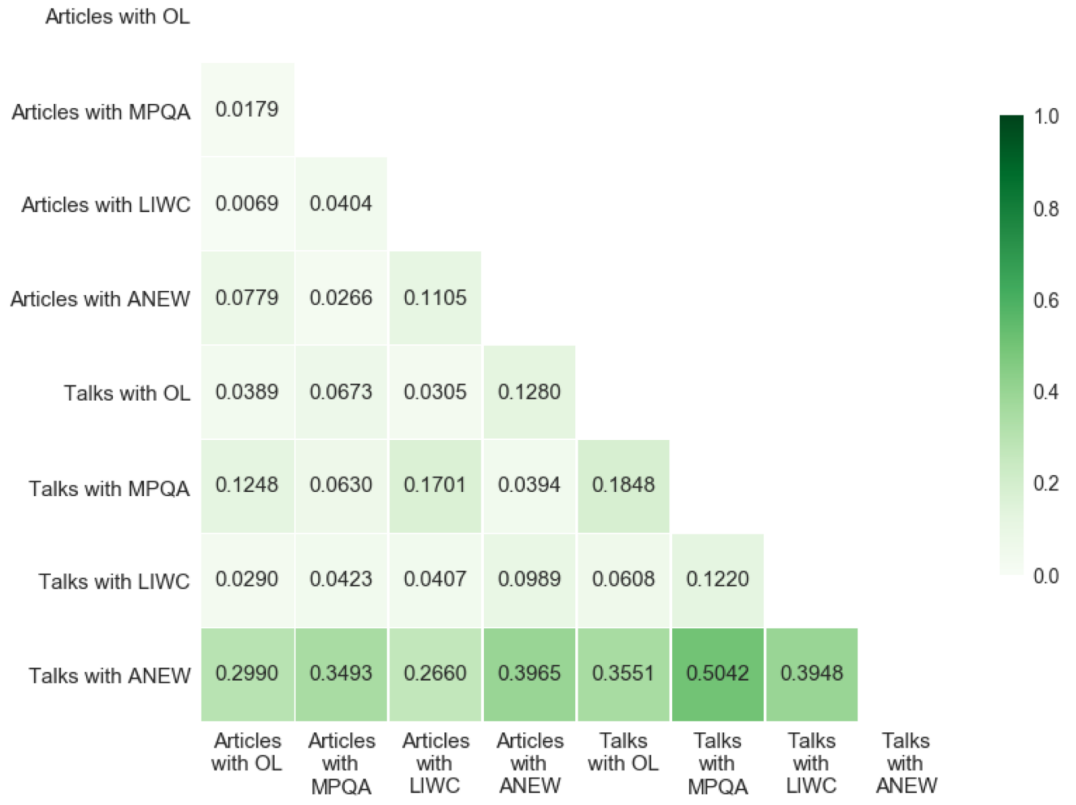


Figure 6.13.: Heatmap of Jensen-Shannon Divergence values between sentiment distributions with different lexicons for people and events on Wikipedia.

It can be seen from the heatmap that the lowest value of JSD is for OL and LIWC both about article pages (JSD=0.0069), which indicates that the distributions of the

score of written text produced by these two lexicons are similar with each other the most. The highest value of JSD in the figure is for lexicon MPQA and ANEW about talk pages (JSD=0.5042), which indicates the distributions of the score of verbal text produced by these two lexicons are similar to each other the least. Moreover, the color between distributions of articles is lighter than the color between distributions of talks, which indicates the distributions between the four lexicons for articles is more similar to each other than the distributions between the four lexicons for talks. In addition, the row of *Talks with ANEW* has the darkest color than others, which indicates that sentiment distribution of talks with ANEW has the least similarity to other distributions. What is surprising is that the most similar distribution of talks to articles with certain lexicon is not the one with the same lexicon. For example, comparing distributions of talks with the four lexicons to the distribution of articles with OL, the most similar distribution is with LIWC. Similarly, the most similar distribution of talks to articles with MPQA is with LIWC, to articles with LIWC is the distribution of talks with OL, and to articles with ANEW is the distribution of talks with MPQA. An implication of this is the possibility that there is a significant difference between verbal text and written text while carrying out sentiment analysis with some certain lexicon.

This chapter presented the sentiment distributions regarding Wikipedia articles and talks, including the distribution for overall entities and entities with certain attributes. By relating the date of birth to people, and relating the date of occurrence to events, the changes of sentiment on temporal dimension taking people and events as objects have been explored. In addition, by comparing the different behaviours of lexicons to various situations, some suggestions of lexicon selection in sentiment analysis have been given.

In the next chapter, three interactive visualisation widgets will be presented, which can be used to explore the sentiment of Wikipedia entities in various ways.

7. Visualisation Widgets

In order to visualise the sentiment of Wikipedia entities interactively and explore untrivial patterns inside, we propose three interactive visualisation widgets in this thesis. In this chapter, we will introduce these three widgets by showing an instance for each of them. All these three widgets can be used to visualise the specified group of Wikipedia entities by setting certain attributes.

These three widgets named as WikiSentiFlow, WikiSentiScatter, and WikiSentiViewer. Specifically, WikiSentiFlow is a widget mainly used to visualise the sentiment distribution of a large number of entities on the temporal dimension as a sentiment river. WikiSentiScatter is a widget mainly used to visualise the sentiment for individual entities as a scatter plot on a timeline. WikiSentiViewer is a widget mainly used to visualise the sentiment distribution of individual entities on a world map according to their geolocations.

All these three widgets are edited with Jupyter Notebook, and can be accessed on Github or online interactive notebooks[1] via Binder service[2]. The interactive functions have been implemented by Ipywidgets library. Besides, Appmode library helps the widgets add a mode that hiding all codes and showing only the interactive interface, which is helpful for general users to get rid of codes. Meanwhile, programmers could access the edit mode and interact with codes. The following sections will introduce the three visualisation widgets separately.

7.1. WikiSentiFlow

WikiSentiFlow as an interactive widget aims to show the overview of sentiment distributions of Wikipedia concepts on temporal dimension. It is specific for entities which are people or events, by connecting people to their birth date and connecting events to their occurrence date. It grouped the entities relating to the same month and sought out the 25th percentile, median, and 75th percentile of their scores for each month to draw a sentiment river on a time axis where time flow from left to right. It can be used to quickly locate untrivial patterns from the sentiment river merged with a large number of entities across hundreds of years.

To run this widget, users can select the lexicon being used, the domain of Wikipedia, the target category of entities, the target sentiment, the covered geographies and times of entities, and the minimum length of the text. Besides, the rolling average can be chosen to be used or not. The setting of minimum length can be used to

[1] All the related links can be found on `https://github.com/qianhongYe/WikiSentiment`.
[2] `https://mybinder.org`, as seen on Oct. 14, 2018

filter out extremely short articles or talks, since most of the time these articles or talks make no sense but affect the result. The rolling average is recommended when the sentiment fluctuates to a great extent. The larger the windows size of rolling average, the smoother the river turning to.

After completing all settings, the corresponding sentiment river will be generated dynamically, together with the statistics of entities being involved for each month. The horizontal value of the sentiment river covers from 25th to 75th percentile of the scores in terms of the corresponding month on the x-axis, and the line in red inside the river indicates the median of scores. These plots are implemented with Matplotlib library. The statistics attached to plots show the number of entities before and after each filtering according to the settings.

Figure 7.1 shows the user interface of WikiSentiFlow, by using an instance of exploring the sentiment of events in North America occurred during the 19th century.

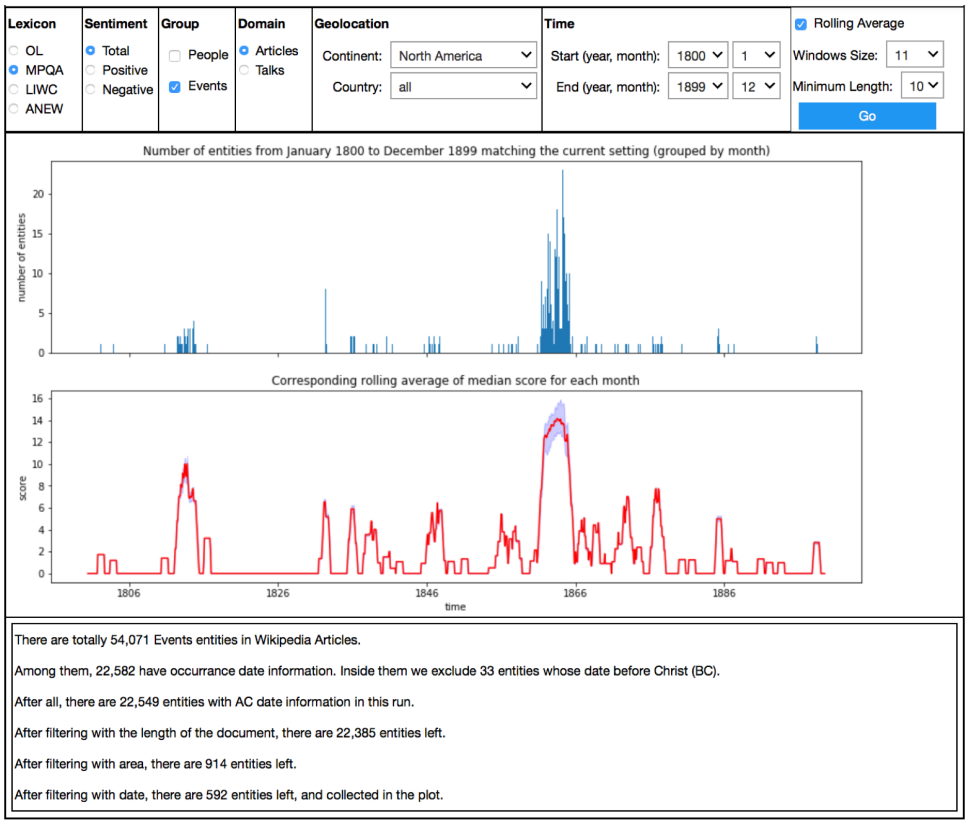


Figure 7.1.: Interface of WikiSentiFlow by taking the exploration of sentiments for events in North America during 19th century as an instance

The top of the interface positions the input parameters that users can set according to their own needs. In this case, we select total sentiment for events in terms of

Wikipedia articles. Since MPQA shows the highest sensitivity for events than other lexicons as discussed in the previous chapter, we use MPQA in this case. Following this, the continent has been set to North America for geolocation, and the time has been set from January 1800 to December 1899. In order to smooth the fluctuation slightly, the window size has been set to 11. Besides, the minimum length has been set to 10 to exclude extremely short articles. After clicking "Go", the corresponding sentiment river together with the bar plot of the number of calculated entities for each month is generated in the middle of the interface. In the meantime, the statistics of the displayed entities are printed on the bottom of the interface, as shown in Figure 7.1, which shows the number of overall events in Wikipedia articles, and the number of excluded events by filtering with the length, geolocation, and time.

The number of entities being included for each month has been shown together because it plays an instrumental role in the explanation of the results. For example, a small number of entities usually result in a fluctuating river since the median of scores will be determined by the limited individuals.

From the plot, it is apparent that the median of the sentiment reaches a peak around 1863, as well as the number of entities being involved. To further glean the insights, we looked up the sentiment river with positive and negative sentiment separately for Wikipedia articles and talks respectively (see Figure E.1). By doing this, we can see if the sentiment is dominated by positive or negative sentiment, and the different sentiments between written text and verbal text. The results show that the most prominent peak for either positive or negative sentiment regarding either articles or talks is located in the same period of time. In detail, the median score of negative sentiment is higher than positive sentiment with regard to articles, whereas the median score of positive sentiment is higher than negative sentiment with regard to talks. This finding can be further explored using WikiSentiScatter by inspecting the entities contributing to the results (Bying using WikiSentiScatter, we can see that the peak stems from American Civil War. More details will be described in the next section).

In summary, with WikiSentiFlow it can be easy to not only find out exceptional sentiments or untrivial patterns from the overview of large-scope entities, but also compare the sentiment distributions between different lexicons, sentiments, domains, etc.

7.2. WikiSentiScatter

WikiSentiScatter, as the name suggests, is a widget used to show sentiment score with an interactive scatter plot for individual entities. The same as WikiSentiFlow, WikiSentiScatter focuses on people and events, and displays the sentiment on a time axis. Specifically, it presents the positive score, negative score, total score, and neutralised score for each entity satisfying the requirements of the settings. Neutralised score is the score calculated by subtracting the negative score from the positive

score of an entity as mentioned earlier (see Section 5.2.1). It is recommended to use WikiSentiFlow in particular after finding exceptional sentiments with WikiSentiFlow, since it is capable of examining the detailed entities and their scores contributing to those exceptional sentiments. However, the number of displayed entities is limited for each run. In other words, it is unable for this widget to display the sentiment of all entities across hundreds of years. In this case, users need to shrink the number of displayed entities by narrowing the range of attributes.

Similar to WikiSentiFlow, users can select the lexicon, domain, category, and set minimum length, geolocation, and time for each run to generate the interactive scatter plot. Each dot on the scatter plot represents one entity. The position on the horizontal axis indicates the related time of this entity, and the position on the vertical axis indicates the neutralised score of this entity. The size and color of each dot indicate the total score of this entity. The larger the size, the higher the total score. The darker the color, the higher the total score. While hovering the mouse on the dot, it will show the name (i.e. title) this dot representing to, as well as its positive score and negative score. The scatter plot is implemented with Plotly library. Attached to the plot, there will be a text describing statistics for the entities being displayed.

By using the same instance as the one used for WikiSentiFlow, i.e. exploring the sentiment of events in North America during the 19th century, we will demonstrate how WikiSentiScatter works. Figure7.2 shows the interface based on this instance. The top of the interface positions the similar parameters as WikiSentiFlow displayed. The only difference is the absence of *Sentiment* and *Rolling Average*. In this case, all the parameters are set to the same as the instance for WikiSentiFlow, and the explanation can be seen in the previous section. After triggering the "Go" button, the corresponding interactive scatter plot will be shown in the middle of the interface, together with a statistics of displayed entities shown on the bottom of the interface as described in WikiSentiFlow.

It can be seen from the figure that there are three noticeable clusters of entities during the 19th century, and the largest one is between 1860 to 1866. There are a total of 592 entities collected based on the setting as shown in the statistics. By narrowing down the setting of geolocation from North America to America, the statistics show a total of 536 entities. These two numbers indicate that most of the events in Figure 7.2 have occurred in the United States. By changing the range of time from 1860 to 1866 (or zooming in the largest cluster) and examining the entities each dot representing for, it shows that most of them are battles belonging to American Civil War. This result explained the peak in Figure 7.1. Besides, while exploring further by changing the value of *Domain* from *Articles* to *Talks*, it shows that a considerable number of battles are owning a higher positive score than negative score. These results would seem to suggest that there is a substantial amount of content of Wikipedia talks either being used to organise conversations or relating to the opinions about articles, rather than relating to the entities directly.

Figure 7.2.: Interface of WikiSentiScatter by taking the exploration of sentiments for events in North America during 19th century as an instance

In summary, WikiSentiScatter can be used to gain an insight into the sentiment distributions, as well as locate entities with exceptional scores quickly and visually.

7.3. WikiSentiViewer

WikiSentiViewer is a widget designed to show sentiment for individual entities on a world map based on their geolocations. It can be used for all entities with geolocation attribute. By using circle marker to represent each entity, and using the size and color of the circle marker to indicate the sentiment of the entity, WikiSentiViewer is able to visualise the sentiment distributions regarding the certain geographical area. This widget can be used to compare the sentiment distributions among different areas, quickly locate the area abundant in sentiments, and identify entity with the noticeable sentiment.

To generate a new sentimental map, users can select lexicon, domain, group, geolocation, time, and minimum length as described in the other two widgets. Besides, users can set the threshold for sentiment scores. As mentioned earlier, the entities have been represented by circle markers. The size of the circle indicates the total

score of the entity this circle on behalf of. The larger the circle, the higher the total score. The color of the circle indicates the balance between positive and negative sentiment. An entity carrying only positive sentiment shows blue, and an entity carrying only negative sentiment shows red. If entities carrying both positive and negative sentiment, the proportion of blue to red determined by the proportion of positive sentiment to negative sentiment. For example, if the ratio of positive sentiment and negative sentiment for some entity is the same, the corresponding circle will be purple. By clicking a certain circle marker, the name of the circle (i.e. title of the entity) will be shown on the plot, together with its detailed positive score and negative score. The interactive map is implemented by Folumn library. Attached to the plot, a text description about statistics of displayed entities will be shown, as described in the other two widgets.

Figure 7.3 shows an instance of visualising the sentiment distributions of American events during American Civil War. Regarding the layout of the interface, most of the settings have been positioning on the right side panel. For the instance in the figure, OL has been used as lexicon, domain has been set to articles, category has been set to events, geolocation has been set to America, time has been set to from April 1861 to May 1865, minimum length has been set to 10, and all the three sentiment scores have been set to from 1 to 100. As it can be seen from the figure, the sentimental map has been shown in the center of the interface, and the detailed statistics have been set out on the bottom of the interface.

It can be seen from the figure that there are a large number of red circles, and most of them are gathered in the lower right corner. It indicates a considerable number of negative events. By examining each circle, we confirmed that most of them are battles. From the map, we can see that most of the battles took place in the southeast part of America.

While exploring the sentiment distributions of events with more periods of time, we found that a majority proportion of displayed events are negative events, and nearly all of them relate to battles or earthquakes. One possible explanation for this might be that events relating to battles or earthquakes on Wikipedia are more likely being tagged with geographic coordinates. Differ from events, the positive people and negative people are generally in similar proportions. Besides, most of the time the number of entities in terms of people is less than the number of entities in terms of events, which seems suggests that the event entities on Wikipedia are more likely to get geographic coordinates than the person entities. As an example, Figure 7.4 shows the sentimental map of American people who born in the 19th century. In detail, the lexicon has been set to *OL*, and the domain has been set to *Articles*. It can be seen from the figure that there are comparable positive entities to negative entities.

In summary, WikiSentiViewer can be used to visualise the geographical distribution of the sentiment on Wikipedia.

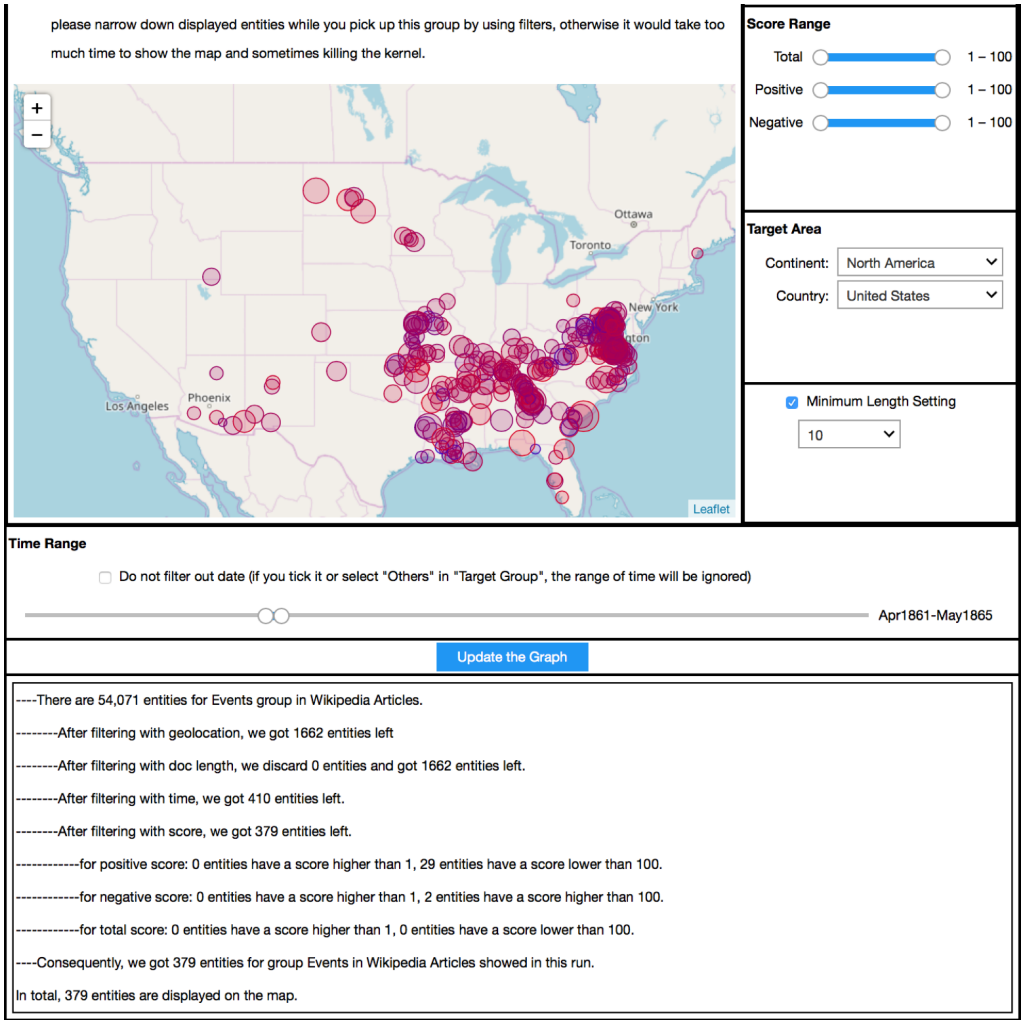


Figure 7.3.: Interface of WikiSentiViewer by taking the exploration of sentiments for American people who born in 19th century as an instance

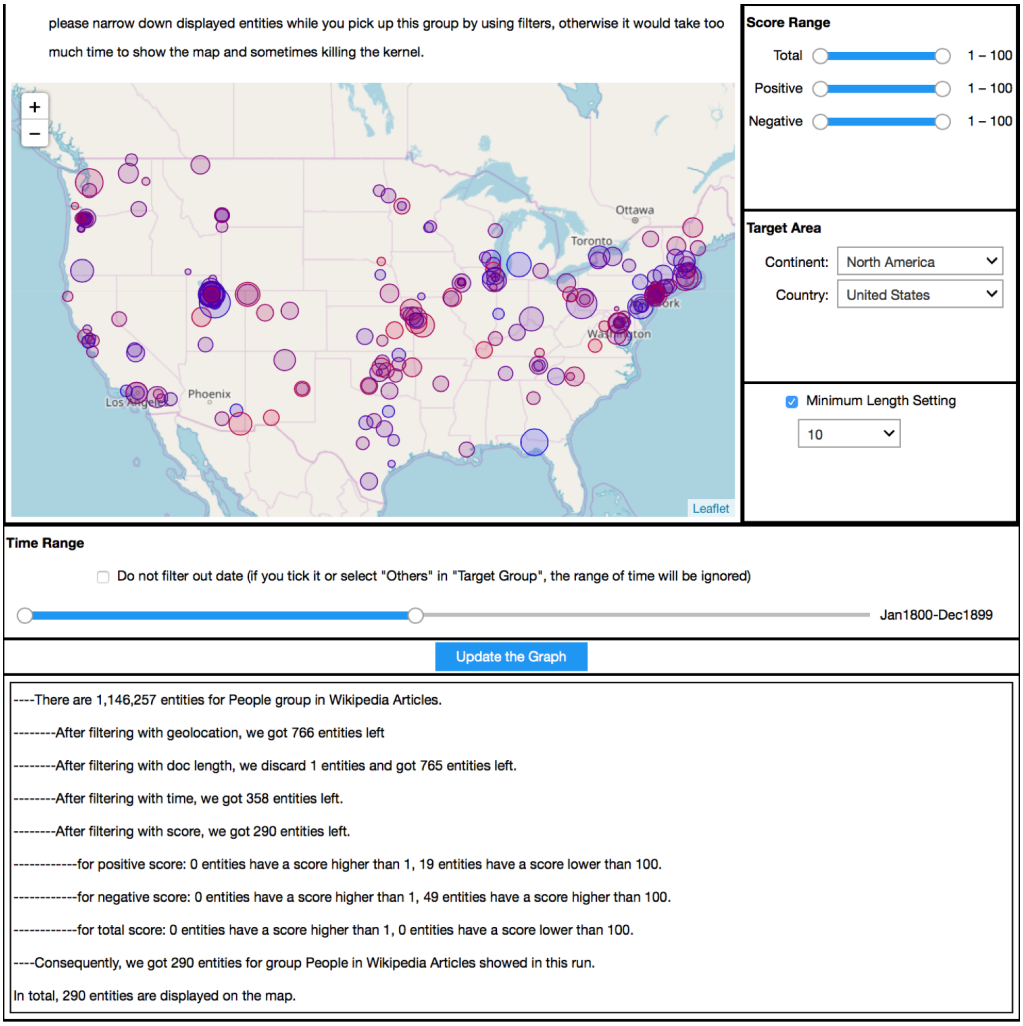


Figure 7.4.: Interface of WikiSentiViewer while visualising the sentiment of American people who born in 19th century

8. Conclusion and Future Work

8.1. Conclusion

In this thesis, we analysed the sentiment for entire English Wikipedia concepts with a lexicon-based method, and gave sentiment scores for each concept. Specifically, we analysed Wikipedia articles and Wikipedia talks separately. By comparing the sentiments with OL, MPQA, LIWC, ANEW four lexicons, we found that the median of the sentiment score with MPQA is the highest, and the median with ANEW is the lowest. The medians with OL and LIWC are close, ranked in the middle place. By comparing the sentiments between Wikipedia articles and talks, it turned out that the scores for talks are higher than the scores for articles in terms of lexicon OL, MPQA, and LIWC, whereas the scores for articles are higher than the scores for talks in terms of lexicon ANEW. This result probably suggests that lexicons such as OL, MPQA, and LIWC are more sensitive for verbal language, while lexicons such as ANEW are more sensitive for written language.

Besides, we explored the changes of sentiment over time with regard to Wikipedia concepts about people and events, and investigated the discovered exceptional sentiment. Regarding entities of people, the sentiment is slightly increasing over time according to OL, MPQA, and ANEW, whereas it is decreasing over time according to LIWC. Regarding entities of events, the sentiment is slightly decreasing over time for all lexicons. Further, we picked five representative types of events and examined the sentiment for each of them with each of the four lexicons. The result through analysing Wikipedia articles shows that text describing wars gets more attention of MPQA, text describing festivals gets more attention of LIWC, and text describing earthquakes gets more attention of ANEW. Taken together, each lexicon has its own bias regarding the text describing different types of events.

Furthermore, we proposed three interactive visualisation widgets for effectively visualising the sentiment distributions of Wikipedia concepts with varying time and geolocation attributes. Moreover, the source codes of these widgets are visible and editable due to the property of Jupyter Notebook, which means users can directly modify the widgets according to their own needs. These widgets are able to assist in discovering either untrivial patterns of sentiments on Wikipedia or Wikipedia entities with exceptional sentiments.

8.2. Limitations

However, there are several limitations in our study. First, the version of Wikipedia articles and talks we used is July 2018, whereas the version of Wikipedia extracted by DBpedia is April 2016, therefore, there is a considerable number of entities on DBpedia unable to match the corresponding entities on Wikipedia. It happens when the title of entity has been changed, or the date that the entity joined to Wikipedia is later than April 2016. Accordingly, a fair number of attributes offered by DBpedia, such as category, date, and geolocation, fail to connect to the Wikipedia entities.

Second, there is a number of geolocations which are incorrect. In general, DBpedia manages the geographic coordinates by using the positive and negative number to denote the east and west respectively in terms of longitude, and using the positive and negative number to denote the north and south respectively in terms of latitude. However, it sometimes missed the minus sign out. Consequently, it caused incorrect geolocations, which reflected on WikiSentiViewer the most.

Third, since we used a lexicon-based method to do the sentiment analysis, the sentiment words in the text which are authenticated by lexicons become the only basis of measurement for the sentiment. However, words authenticated by lexicons could carry no emotions in the context, and words absent in the lexicon might bear emotions. It influences the accuracy of the final results.

Fourth, the employed approaches to document-level sentiment analysis make no distinction between different domains (i.e. articles and talks) and categories (e.g. people or events) of text.

8.3. Future works

As the future work, we plan to improve our approach by dealing with the issues described above. First of all, we will try to employ the newest version of DBpedia dataset or build an extra connection between entities on DBpedia and entities on Wikipedia. An existing connection is offered by DBpedia, which is a list of links between the URI of entities on DBpedia and the URI of entities on Wikipedia[1].

Second, we plan to apply tools or datasets from the third-party to examine the correctness of geolocation information from DBpedia and correct the erroneous part of it.

Third, we will analyse the text combining semantic analysis to process the natural language more precisely. Besides, machine learning techniques could be used to generate the results of sentiment more meticulously.

Finally, field knowledge can be applied to analyse text describing specific area in our future work, to improve the precision of the result. And text for article pages and talk pages can be analysed differently according to their own characteristics.

[1]The solution of using this connection is limited to resolving the entities whose name has been changed before April 2016.

Appendices

A. Set of Stop Words

Since the set of stop words from NLTK package has been updated occasionally, we listed the set being used as follows.

will, does, the, just, ourselves, his, himself, t, against, further, up, out, is, than, weren, your, being, an, above, no, my, y, any, during, more, into, from, isn, as, for, o, most, they, where, aren, yours, when, so, between, them, me, hers, our, am, hasn, were, it, ma, having, on, be, shouldn, but, a, that, ve, him, very, needn, too, before, its, itself, didn, m, have, mustn, myself, here, was, she, shan, why, because, re, had, same, and, now, he, who, d, until, about, wouldn, own, by, of, under, doesn, if, in, once, can, below, off, should, been, there, ours, hadn, theirs, couldn, whom, other, down, at, yourself, her, after, how, then, s, again, their, do, ll, some, doing, herself, we, only, not, don, ain, did, all, won, with, both, what, to, nor, themselves, i, these, few, this, while, which, such, those, has, yourselves, or, through, each, over, wasn, you, are, haven, mightn

Table A.1.: NLTK's set of English stop words

B. Example of Document-Level Sentiment Analysis

B.1. Example of Scoring Wikipedia Articles

Here is an example of calculating sentiment score for a short article with *Opinion Lexicon*.

Title: Sleeping Betty

Original text:

Sleeping Betty () is a Canadian animated short film that humorously reinterprets the classic fairy tale, Sleeping Beauty. Awards for the film include Best Animated Short at the 29th Genie Awards, the Audience Award at the Etiuda&Anima International Film Festival, the Audience Award and Judges Award at the Melbourne International Animation Festival, Best Animation at the Jutra Award, as well as the Public Prize and the Best Canadian Animation Award at the Ottawa International Animation Festival.

First we change all words to lower-case and remove all title-words and get the new text as below.

() is a canadian animated short film that humorously reinterprets the classic fairy tale, sleeping beauty. awards for the film include best animated short at the 29th genie awards, the audience award at the etiuda&anima international film festival, the audience award and judges award at the melbourne international animation festival, best animation at the jutra award, as well as the public prize and the best canadian animation award at the ottawa international animation festival.

And then we do tokenization and exclude all punctuation except for hyphen within hyphenated words.

is, a, canadian, animated, short, film, that, humorously, reinterprets, the, classic, fairy, tale, sleeping, beauty, awards, for, the, film, include, best, animated, short, at, the, 29th, genie, awards, the, audience, award, at, the, etiuda, anima, international, film, festival, the, audience, award, and, judges, award, at, the, melbourne, international, animation, festival, best, animation, at, the, jutra, award, as, well, as, the, public, prize, and, the, best, canadian, animation, award, at, the, ottawa, international, animation, festival

After that we remove stop-word, digital words, do the lemmatisation for tokens and get the result with 49 tokens as below.

canadian, animated, short, film, humorously, reinterprets, classic, fairy, tale, sleeping, beauty, award, film, include, best, animated, short, genie, award, audi-

ence, award, etiuda, anima, international, film, festival, audience, award, judge, award, melbourne, international, animation, festival, best, animation, jutra, award, well, public, prize, best, canadian, animation, award, ottawa, international, animation, festival

According to lexicon OL, we get 15 positive tokens shown as below and 0 negative tokens.

humorously, classic, beauty, award, best, award, award, award, award, best, award, well, prize, best, award

Therefore, we get $(15/49) * 100 = 30.6122$ for its positive sentiment score, $(0/49) * 100 = 0$ for its negative sentiment score, and $30.6122 + 0 = 30.6122$ for its total score based on OL.

B.2. Example of scoring Wikipedia talks

In this part, we use talk page "Koblenz" as an example to show how we calculate sentiment score for talk pages. In this example we use ANEW as lexicon to calculate positive score for "Koblenz".

First, we extract tokens of the target talk page with the way we did for Wikipedia articles. Then we extract positive tokens according to ANEW lexicon. After that, for each positive token we obtain the valence and term frequency in article "Koblenz". With formula 5.10, we calculate the positive score for each positive token. All positive tokens and corresponding information are shown in table B.1. The length of talk page "Koblenz" is 200.

Positive Token	Valence	Frequency in Talk Page	Frequency in Article Page	Positive Score
history	0.0600	1	1	0.0
flag	0.2550	1	2	0.0
good	0.6175	1	0	0.0030875
village	0.2300	1	1	0.0
city	0.2575	4	43	0.0
moment	0.1900	1	0	0.0009500
museum	0.1350	2	1	0.0006750
cheer	0.7750	1	0	0.0038750

Table B.1.: The necessary information to calculate positive score for each positive token in talk page "Koblenz", and the corresponding positive score.

By summing up their positive score and normalised with 100, we get the final positive score 0.8588 for talk page "Koblenz".

B.3. Example of Typical Articles with Extremely High Score

	Example 1	Example 2
Title	BSAC	Te Papa
Article Text	BSAC can stand for:	Te Papa can signify:
Tokens	stand	signify
Number of Tokens	1	1
Positive Tokens with MPQA	stand	signify
Number of Positive Tokens	1	1
Positive Score	100	100

Table B.2.: Example of disambiguation pages with extremely high score based on MPQA

	Example
Title	Odisha State Film Award for Best Singer
Article Text	Winners of Odisha State Film Award for Best Singer:
Tokens	winner
Number of Tokens	1
Positive Tokens with MPQA	winner
Number of Positive Tokens	1
Positive Score	100

Table B.3.: Example of articles about the award with extremely high score based on MPQA

C. Noise Analysis regarding People in Articles

By looking into entities of people, we found a considerable number of noise in the last few years. In order to dig out the proportion of the noise, the entities of people during 01.01.2006 to 31.12.2017 have been identified one by one manually. The statistics have been shown in Table C.1.

There is a total of 124 people whose birth date is between 01.01.2006 and 31.12.2017 (collected in the table), and 8 people whose birth date is during 04.2016 and 01.01.2018 (excluded in our analysis). From the table, it can be seen that the number of people is less and less while the date near 2018. A possible explanation for these results may be that rare people will be written into Wikipedia in their young age. As shown in the table, most of people born after 2009 are assigned erroneous birth information, and all the people born after 2013 except for royals are assigned erroneous birth date. It is likely that only princes or princesses will be written into Wikipedia in their baby age. In addition to royal, child actors or actresses play a instrumental role as well. In fact, they take up a half proportion of the total number of people between 2006 and 2008. Besides, the other entities are relating to criminal cases (such as death cases or disappearance cases), medicine cases (such as child with rare disorder), or child prodigy.

Year	People	N	Ambi	P	Actor	Royal	Case	Prodigy	Others
2017	5	5	1	0	0	0	0	0	0
2016	8	7	1	1	0	1	0	0	0
2015	13	11	1	2	0	2	0	0	0
2014	7	4	2	3	0	3	0	0	0
2013	7	5	0	2	0	2	0	0	0
2012	8	4	1	4	1	2	1	0	0
2011	10	5	0	5	0	3	1	0	1
2010	9	4	0	5	2	0	3	0	0
2009	4	3	0	1	0	0	1	0	0
2008	9	1	0	8	4	1	1	2	0
2007	23	2	0	21	11	7	2	1	0
2006	21	0	0	21	14	1	3	2	1

Notes:

- N (negative) indicates the number of people whose birth date extracted from DBpedia does not match the information on Wikipedia. It could be a different date from the date on Wikipedia, or a created one while there is no birth date on Wikipedia.
- Ambi indicates a special situation that for the same entity there are two different birth date simultaneously recorded on DBpedia, and the erroneous one has been extracted.
- P (positive) indicates the number of people whose birth date does match the birth date on Wikipedia.

The following columns are the finer categories for P (positive) entities.

- Actor indicates the number of people which are child actors, actresses or models.
- Royal indicates the number of people which are princes, princesses or other royals.
- Case includes the number of people which are involved in death case, disappearance case or special medicine case.
- Prodigy indicates the number of people which are child prodigy, such as artist, singer, or chess player.
- Others indicates the number of people which are actually collectives instead of real people. DBpedia sometimes assigns the same entity with multiple categories, for example assigning an company with people, company, and collective three categories. In our thesis, entities with people as one of their categories will be all involved, even through they are not real people.

Table C.1.: Statistics about people on Wikipedia between 01.01.2006 and 31.12.2017

D. Extra Sentiment Distributions for Wikipedia Talks

D.1. Sentiment Distributions by Filtering with Text Lengths

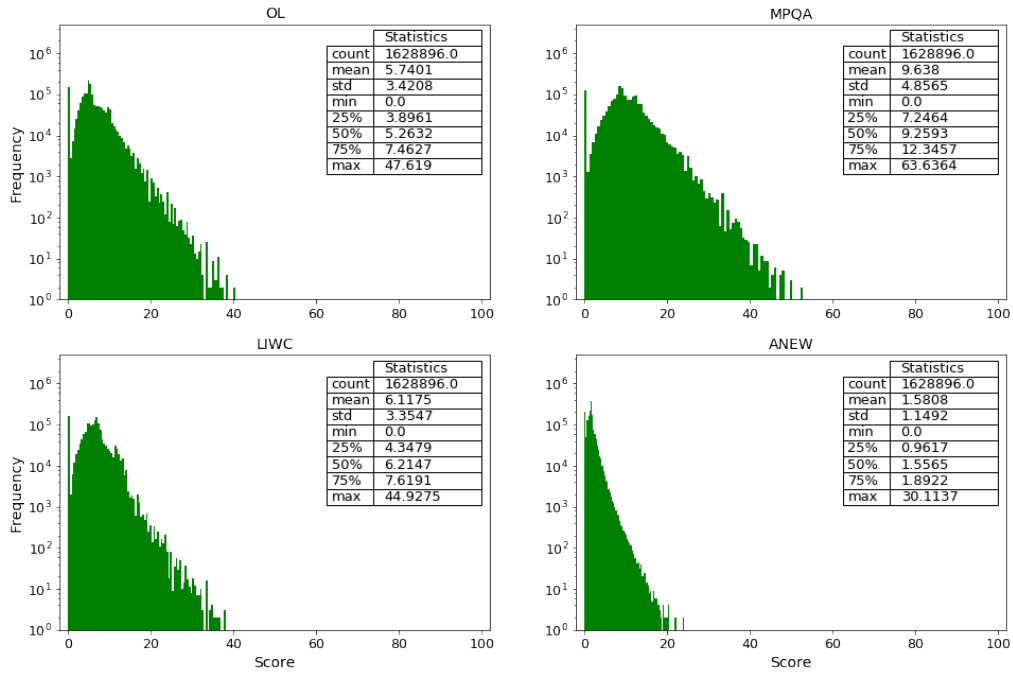


Figure D.1.: Distributions of total score for Wikipedia talks except for talks whose length less than 20

D.2. Sentiment Distributions by Using Pure Score

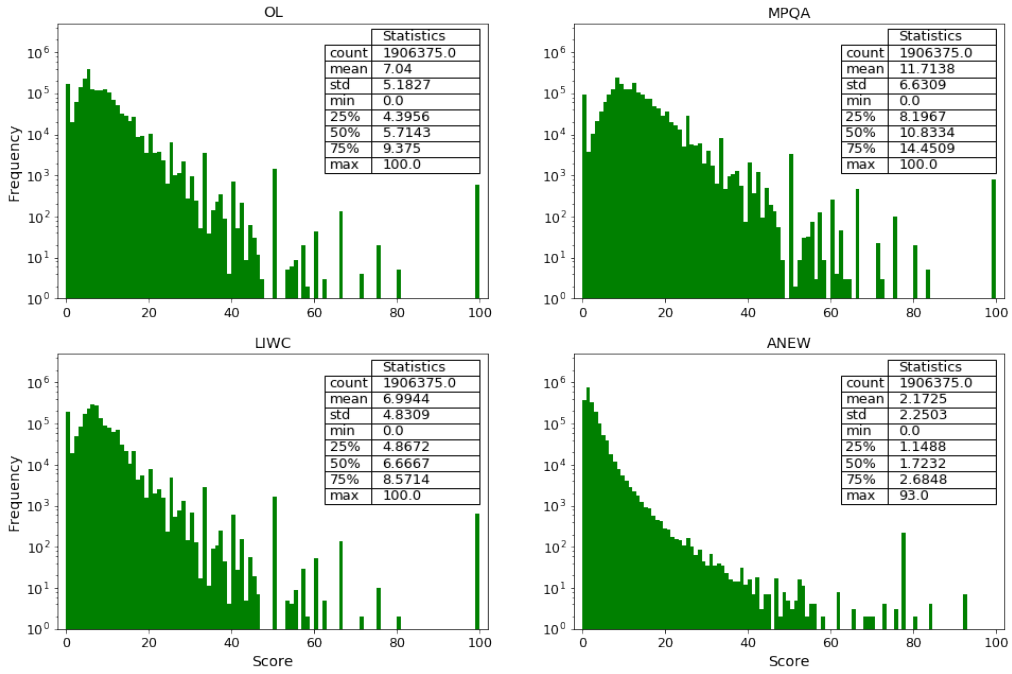
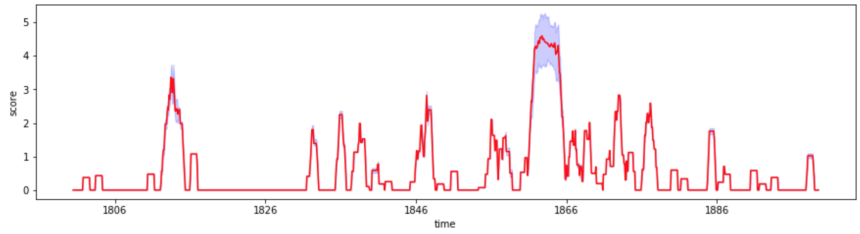
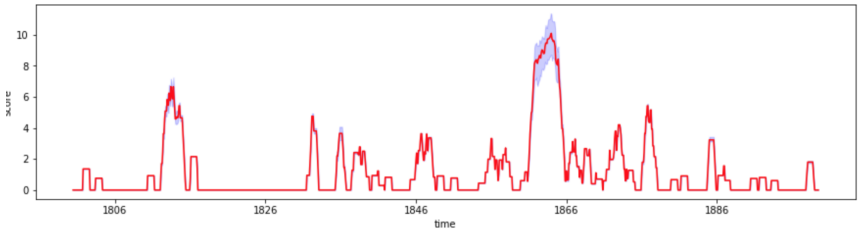


Figure D.2.: Distributions of pure total score for entire Wikipedia talks

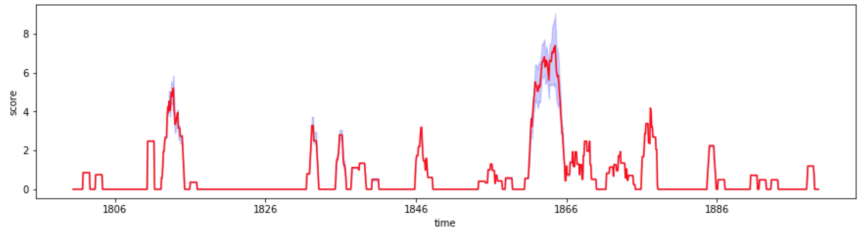
E. Example of Using WikiSentiFlow to Explore the Sentiment on Wikipedia



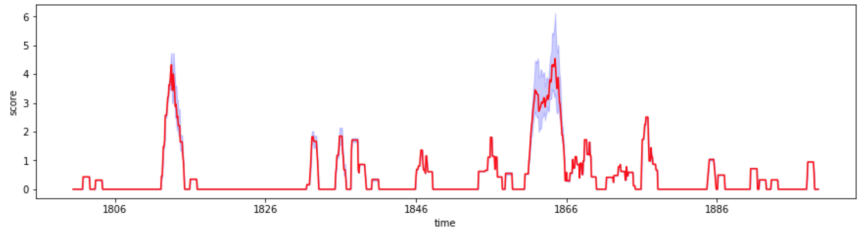
(a) Articles, Positive



(b) Articles, Negative



(c) Talks, Positive



(d) Talks, Negative

Table E.1.: **Comparison of sentiments for events in North America during 19th century between different polarities and domains by using WikiSentiFlow:** Each plot shows the rolling average of median score for specific entities. It aims to compare the positive and negative sentiment regarding Wikipedia articles and talks.

Bibliography

[AYHK11] Basak Alper, Huahai Yang, Eben Haber, and Eser Kandogan. Opinionblocks: Visualizing consumer reviews. In *IEEE VisWeek 2011 Workshop on Interactive Visual Text Analytics for Decision Making*, 2011.

[BHBL11] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global, 2011.

[BL99] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer, 1999.

[BL06] Tim Berners-Lee. Linked data, 2006.

[BLK+09] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia-a crystallization point for the web of data. *Web Semantics: science, services and agents on the world wide web*, 7(3):154–165, 2009.

[Bou15] Ameni Boumaiza. A survey on sentiment analysis and visualization. *Journal of Emerging Technologies in Web Intelligence Vol*, 7(1), 2015.

[BVS08] Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. International sentiment analysis for news and blogs. In *ICWSM*, 2008.

[Cha08] Rishi Chandy. Wikiganda: Identifying propaganda through text analysis. *Caltech Undergraduate Research Journal. Winter*, 2009:6–11, 2008.

[CLT+11] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE transactions on visualization and computer graphics*, 17(12):2412–2421, 2011.

[CNP06] Giuseppe Carenini, Raymond T Ng, and Adam Pauls. Interactive multimedia summaries of evaluative text. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 124–131. ACM, 2006.

[CSR17] Benjamin Cabrera, Laura Steinert, and Björn Ross. Grawitas: a grammar-based wikipedia talk page parser. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, 2017.

[GACOR05] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. Pulse: Mining customer opinions from free text. In *international symposium on intelligent data analysis*, pages 121–132. Springer, 2005.

[GCW+06] Michelle L Gregory, Nancy Chinchor, Paul Whitney, Richard Carter, Elizabeth Hetzler, and Alan Turner. User-directed sentiment analysis: Visualizing the affective content of documents. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 23–30. Association for Computational Linguistics, 2006.

[GGLBY16] Eduardo Graells-Garrido, Mounia Lalmas, and Ricardo Baeza-Yates. Sentiment visualisation widgets for exploratory search. *arXiv preprint arXiv:1601.02071*, 2016.

[GR11] Mihai Grigore and Christoph Rosenkranz. Increasing the willingness to collaborate online: An analysis of sentiment-driven interactions in peer content production. 2011.

[GS06] Matthew Gentzkow and Jesse M Shapiro. Media bias and reputation. *Journal of political Economy*, 114(2):280–316, 2006.

[GZ12] Shane Greenstein and Feng Zhu. Is wikipedia biased? *American Economic Review*, 102(3):343–48, 2012.

[HHN00] Susan Havre, Beth Hetzler, and Lucy Nowell. Themeriver: Visualizing theme changes over time. In *Information visualization, 2000. InfoVis 2000. IEEE symposium on*, pages 115–123. IEEE, 2000.

[HL04] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[ILC+14] Daniela Iosub, David Laniado, Carlos Castillo, Mayo Fuster Morell, and Andreas Kaltenbrunner. Emotions under discussion: Gender, status and communication in online collaboration. *PloS one*, 9(8):e104880, 2014.

[INPG10] Takashi Iba, Keiichi Nemoto, Bernd Peters, and Peter A Gloor. Analyzing the creative editing behavior of wikipedia editors: Through dynamic social network analysis. *Procedia-Social and Behavioral Sciences*, 2(4):6441–6456, 2010.

[KCM04] Graham Klyne, Jeremy J Carroll, and Brian McBride. Resource description framework (rdf): Concepts and abstract syntax, 2004.

[LIJ+15] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey,

Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.

[Liu10] Bing Liu. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666, 2010.

[Liu12] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[LKCM12] David Laniado, Andreas Kaltenbrunner, Carlos Castillo, and Mayo Fuster Morell. Emotions and dialogue in a peer-production community: the case of wikipedia. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, page 9. ACM, 2012.

[LYK+12] Dongning Luo, Jing Yang, Milos Krstajic, William Ribarsky, and Daniel Keim. Eventriver: Visually exploring text collections with temporal references. *IEEE Transactions on Visualization and Computer Graphics*, 18(1):93–105, 2012.

[MDR+06] Gilad Mishne, Maarten De Rijke, et al. Moodviews: Tools for blog mood analysis. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 153–154, 2006.

[MKGD10] Thelwall Mike, Buckley Kevan, Paltoglou Georgios, and Cai Di. Sentiment in short strength detection informal text. *JASIST*, 61(12):2544–2558, 2010.

[MMLW09] Olena Medelyan, David Milne, Catherine Legg, and Ian H Witten. Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, 67(9):716–754, 2009.

[MZL12] Andrius Mudinas, Dell Zhang, and Mark Levene. Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*, page 5. ACM, 2012.

[NEH13] Finn Årup Nielsen, Michael Etter, and Lars Kai Hansen. Real-time monitoring of sentiment in business related wikipedia articles. 2013.

[Nie11] Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[PCI+] James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. The development and psychometric properties of liwc2007.

[PL+08] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[RHF05] Patrick Riehmann, Manfred Hanfler, and Bernd Froehlich. Interactive sankey diagrams. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 233–240. IEEE, 2005.

[RTW+15] Andrew J Reagan, Brian Tivnan, Jake Ryland Williams, Christopher M Danforth, and Peter Sheridan Dodds. Benchmarking sentiment analysis methods for large-scale texts: a case for using continuum-scored words and word shift graphs. *arXiv preprint arXiv:1512.00531*, 2015.

[RW03] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112. Association for Computational Linguistics, 2003.

[Shn92] Ben Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on graphics (TOG)*, 11(1):92–99, 1992.

[SM86] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1986.

[SPUW13] Vivek Kumar Singh, Rajesh Piryani, Ashraf Uddin, and Pranav Waila. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *Automation, computing, communication, control and compressed sensing (iMac4s), 2013 international multi-conference on*, pages 712–717. IEEE, 2013.

[TBP11] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.

[WLY+14] Yingcai Wu, Shixia Liu, Kai Yan, Mengchen Liu, and Fangzhao Wu. Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE transactions on visualization and computer graphics*, 20(12):1763–1772, 2014.

[WWH05] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[ZCR15] Yiwei Zhou, Alexandra Cristea, and Zachary 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*, pages 160–168, 2015.

[ZJZ10] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.