UNIVERSITÄT KOBLENZ · LANDAU
Fachbereich 4: Informatik

WeST People and Knowledge Networks
Institute for Web Science and Technologies

# Subjectivity Detection through Quotation Analysis in Wikipedia

## Masterarbeit

zur Erlangung des Grades einer Master of Science (M.Sc.)
im Studiengang Informatik

vorgelegt von

Wee Pang Wayne

| | |
|---|---|
| Erstgutachter: | JProf. Dr. Claudia Wagner<br>GESIS - Leibniz Institute for the Social Sciences |
| Zweitgutachter: | Dr. Fabian Flöck<br>GESIS - Leibniz Institute for the Social Sciences |

Koblenz, im Mai 2018

# Erklärung

Hiermit bestätige ich, dass die vorliegende Arbeit von mir selbstständig verfasst wurde und ich keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe und die Arbeit von mir vorher nicht in einem anderen Prüfungsverfahren eingereicht wurde. Die eingereichte schriftliche Fassung entspricht der auf dem elektronischen Speichermedium (CD-Rom).

| | Ja | Nein |
|---|---|---|
| Mit der Einstellung dieser Arbeit in die Bibliothek bin ich einverstanden. | ☐ | ☐ |
| Der Veröffentlichung dieser Arbeit im Internet stimme ich zu. | ☐ | ☐ |
| Der Text dieser Arbeit ist unter einer Creative Commons Lizenz (CC BY-SA 4.0) verfügbar. | ☐ | ☐ |
| Der Quellcode ist unter einer GNU General Public License (GPLv3) verfügbar. | ☐ | ☐ |
| Die erhobenen Daten sind unter einer Creative Commons Lizenz (CC BY-SA 4.0) verfügbar. | ☐ | ☐ |

.....

(Ort, Datum) (Unterschrift)

## Zusammenfassung

## Abstract

This Master Thesis is an exploratory research to determine whether it is feasible to construct a subjectivity lexicon using Wikipedia. The key hypothesis is that that all quotes in Wikipedia are subjective and all regular text are objective. The degree of subjectivity of a word, also known as "Quote Score" is determined based on the ratio of word frequency in quotations to its frequency outside quotations. The proportion of words in the English Wikipedia which are within quotations is found to be much smaller as compared to those which are not in quotes, resulting in a right-skewed distribution and low mean value of Quote Scores.

The methodology used to generate the subjectivity lexicon from text corpus in English Wikipedia is designed in such a way that it can be scaled and reused to produce similar subjectivity lexica of other languages. This is achieved by abstaining from domain and language-specific methods, apart from using only readily-available English dictionary packages to detect and exclude stopwords and non-English words in the Wikipedia text corpus.

The subjectivity lexicon generated from English Wikipedia is compared against other lexica; namely MPQA and SentiWordNet. It is found that words which are strongly subjective tend to have high Quote Scores in the subjectivity lexicon generated from English Wikipedia. There is a large observable difference between distribution of Quote Scores for words classified as strongly subjective versus distribution of Quote Scores for words classified as weakly subjective and objective. However, weakly subjective and objective words cannot be differentiated clearly based on Quote Score. In addition to that, a questionnaire is commissioned as an exploratory approach to investigate whether subjectivity lexicon generated from Wikipedia could be used to extend the coverage of words of existing lexica.

# Contents

## List of Figures



## List of Tables

# 1 Introduction

## 1.1 Motivation

User-generated text is ubiquitous on the Internet and is present in various forms, such as comments in social media, blogs and reviews in e-commerce websites. While engaged in a common topic, different users may have varying degrees of motivations, emotions and experience, which could result in different ways of formulating a textual response. For example, an online review for a restaurant could contain additional useful information for other users such as nearby landmarks, recommended dishes and prices, or simply personal opinion and criticism of the quality of food and service. Consequently, the large number of variations in user-generated text has stimulated interest in the field of opinion mining and sentiment analysis, where textual opinions are classified into either positive or negative categories.

Subjectivity detection is the precursor to opinion mining and sentiment analysis. This is because not all user-generated text constitute an opinion; it is possible that a given text has neither positive nor negative sentiment. In the aforementioned example, factual information about the restaurant resembles an objective review, whereas a more descriptive and adjective-rich critique resembles a subjective review. When performing opinion mining and sentiment analysis, it would be practical to first apply subjectivity detection to the corpus to extract only the subjective content. On the same token, subjectivity detection enables factual words or phrases within a corpus to be extracted when looking for unbiased information.

Existing techniques of subjectivity detection include usage of human evaluators and annotators as well as analysis of sentence structures and features based on small text corpora. However, relying exclusively on these methods may not be very scalable because they require a certain degree of understanding of the language; therefore the findings are often applicable to only one language. The approach used in this Master Thesis takes advantage of the large corpus of Wikipedia, which is an open-source resource that spans multiple topics and languages, in order to perform subjectivity detection. Regular text in Wikipedia is hypothesised to be objective as articles are usually verified for impartiality, whereas quotations, which are cited directly, tend to be subjective. A relatively simple and language-independent method is adopted to generate a subjectivity lexicon to evaluate the viability of this hypothesis. If successful, this technique could be scaled across multiple languages and appeal to the commercial world due to its ease of implementation.

## 1.2 Research Scope

This research encompasses the usage of the Wikipedia text corpus to develop a subjectivity lexicon. It is an exploratory research to determine whether the large quantity of words as well as the variety of topics in Wikipedia could be leveraged upon to compute subjectivity of words, instead of relying on language-specific knowledge of grammar and sentence structures. In this Master Thesis, the language that is used to develop and test the subjectivity lexicon is English. The end result would be a subjectivity lexicon consisting of English words and their corresponding "Quote Scores".

Validation of results of this Master Thesis is performed by comparing against lexica from previous researches, namely MPQA and SentiWordNet. As the lexicon generated from Wikipedia contains words which are not present in these two lexica, human evaluators are also employed via online questionnaire to validate the subjectivity of the additional words. The potential use case of the subjectivity lexicon developed from Wikipedia in this research is explored, but its detailed implementation could be a future work for another research project.

## 1.3 Research Question

Are texts within quotations significantly more subjective as compared to regular text in Wikipedia? The key hypothesis is that all texts within quotations in Wikipedia are subjective, whereas other all texts outside quotations are considered objective. The reason behind this hypothesis is that existing text in Wikipedia have gone through multiple revisions to ensure its objectivity. Quotations are usually cited directly hence subjectivity is retained.

Could Wikipedia be used as ground truth for subjectivity detection? While Wikipedia covers a large range of topics and languages, it must be determined whether the Quote Scores assigned to words in Wikipedia are applicable to the commercial world. The subjectivity lexicon generated from Wikipedia should be more comprehensive than existing lexica and also include words which are less common.

Last but not least, it would be interesting to investigate whether subjectivity detection algorithm could be immediately scaled to include other languages in the world. Most existing researches on subjectivity detection are based on the English language; however it is also essential to be able to apply similar scale of subjectivity detection to other languages as well.

# 2 Related Work

The publications and results of previous researches are one of the key driving forces of this Master Thesis. The potential use cases in the commercial world are explored, followed by techniques developed in other researches in the field of subjectivity detection which are used to inspire and evaluate the results of this Master Thesis.

### 2.0.1 Product ratings in e-commerce platforms

Online marketplaces provide a platform for buyers and sellers to transact goods and services. In order to enable consumers to make informed decisions before purchasing a product, average star rating is displayed beside every product. The average star rating of a product is an indication of customer satisfaction; for example, in Amazon, a 5-star rating is the best possible score whereas a 1-star rating is the worst possible score. The final rating for a product is derived from the average of ratings given by users for that product.

However, as product reviews are written by Internet users who are usually hidden behind a pseudonym, this calls into question the trustworthiness of user ratings. For example, Danescu-Niculescu-Mizil C. et al. (2009) pointed out the existence of ”plagiarized reviews” in Amazon, which are reviews with almost identical text, but only with certain words changed to suit the product being reviewed. While it is possible for different individuals to have the same opinion about the product, user reviews which are plagiarised are not so helpful to other users as a user review is intended to be based on personal, first-hand experience of using the product.

Susan M. Mudambi, S. M. and Schuff, D. (2010) analysed 1587 reviews in Amazon to determine whether there are certain factors which could influence the helpfulness of reviews. It is found that the perceived helpfulness of reviews depends on review depth, review extremity as well as product type. Review depth is the length and extensiveness of the review, whereas review extremity resembles the sentiment (positive, negative or neutral). This is where subjectivity detection could play a role in the evaluation of helpfulness of reviews, as determining the degree of subjectivity is a precursor to sentiment analysis of reviews.

Furthermore, Ghose, A. and Ipeirotis, P. G. (2009) have investigated the relationship between product reviews and sales performance. One of the key findings is that for products which are feature-oriented, such as electronic goods, a higher degree of helpfulness is perceived for reviews which contain primarily objective information as compared to subjective expressions. A subjectivity detection and scoring system could be used to enhance the way in which product reviews are displayed to users. Reviews are usually sorted by either top positive, top critical or most recent reviews, but with the incorporation of subjectivity detection in reviews, the reviews could also be ranked by helpfulness score, calculated based on their subjectivity scores

rather than only relying on other users marking a review as helpful. This option could be useful for a user who prefers to find out more factual information about a product from the review section, who would then sort reviews by subjectivity score in ascending order.

Deriving from the aforementioned related works, another potential use case for subjectivity detection in the context of product reviews in the e-commerce environment is to complement the existing review system, especially for products with low number of reviews. The review system in e-commerce tends to scale well for products with large number of reviews because the average rating value would tend to be a reasonable representation of the product quality. However, this poses a problem for new products as well as low-volume products due to low number of reviews. For example, if a product only has 2 reviews in which one is 1-star and the other is 5-star, the average rating would tend to be 3-star. Subjectivity detection could help in ambiguous cases by assigning a certain weightage to ratings based on subjectivity values of the accompanying text, in which ratings that are accompanied by text which are less subjective could be assigned greater weightage.

### 2.0.2 Detection of strongly-biased or fake news

Internet users of today have access to a large selection of news portals; however due to lack of regulation on online journalism, it is becoming increasingly difficult to determine whether information contained within a news article is reliable and accurate. The phenomenon of ”fake news” is becoming prevalent in the Internet, which is intended to deceive and sway opinions of readers.

The approach taken by Shu, K. et. al. (2017) is to first define features of news content, namely; source, headline, body text as well as image or video, if any. Subsequently, fake news is identified through ”knowledge-based” and ”style-based” approach. A knowledge-based approach includes the use of human experts, crowdsourcing or computational fact-checking whereas style-based approach focuses on capturing intent, such as deception. While there are many approaches available to tackle the fake news phenomenon, Tschiatschek, S. et. al. (2018) warns that some of the key issues in identifying fake news are the scarcity of text corpus, risk of bias when labelling ground truth as well as a large variance among the sources of fake news. This would particularly affect human experts and crowdsourcing which could be influenced by personal biases and emotions when attempting to decipher the validity of news.

Moreover, Rubin V. et al. (2015) state that there are 3 types of fake news, namely; serious fabrications, large-scale hoaxes and humourous fakes. A common trait that is observed in the various types of fake news is the usage of exaggeration and sensationalism in text so as to gain traction among readers. Words which are used to

generate such an impression on readers of fake news could be subjective in nature, therefore subjectivity detection can play an important role in fake news detection. As fake news can span across multiple domains, this presents an opportunity to leverage on Wikipedia as a large and openly-available resource to develop a subjectivity lexicon to identify highly-subjective words which may also be commonly used in fake news.

Simm, W. et. al. applied four methods of sentiment analysis on short text comments, namely; tagger analysis, ReadMe analysis, Naïve Bayes classifier as well as lexicon and rule-based method. The lexicon and rule-based method, which is based on a subjectivity lexicon, performed the best in terms of classifying the sentiments of comments. This shows that the subjectivity detection can not only be applied to news articles but also for gaining a better understanding of user comments within the social media environment. In addition to detecting fake news, subjectivity detection could be used to assist in moderation of comments by capturing strongly-worded and inflammatory user comments.

### 2.0.3 Subjectivity Detection

Literature pertaining to other subjectivity detection methodologies is studied so as to draw ideas from existing approaches and identify those which could be used to evaluate the results of this Master Thesis. Lin, C., He, Y. and Everson, R. (2011) have implemented sentence subjectivity detection using weakly-supervised learning. This is achieved using the ”SubjLDA” model, which is defined in the paper as ”a four-layer Bayesian model”. A sentence-level subjectivity label generation layer is added and word prior sentiment information is encoded to the process of drawing word from per-corpus distribution. It was found that this improved version of the LDA model performed better in terms of accuracy of detecting subjectivity as compared to the base LDA model.

Besides that, Janyce Wiebe, J., Wilson, T. and Cardie, C. (2005) discuss the annotation of various ”private state frames” in text, which consist of text anchor, source, target and properties. Text anchor is the key phrase in the text, which is mentioned by the source individual or entity to the target individual or entity. Certain properties describe the ”private state frames” in greater detail, such as its intensity and attitude type. The intensity could be low, medium, high or extreme, whereas the attitude type could be positive, negative, other or none. Based on this concept, annotators are trained and agreement study is conducted in order to produce the Multi-perspective Question Answering (MPQA) opinion corpus (Wilson, T. A., 2008). The MPQA lexicon is subsequently improved from version 2.0 to 3.0 by Deng, L. and Wiebe, J (2015) with the addition of entity-target and event-target annotations.

In addition to that, Esuli, A. and Sebastiani, F. (2006) eloborate on the blueprint

of the SentiWordNet lexicon for opinion mining. A classifier is constructed using semi-supervised method, using 3 initial small sets of positive, negative and objective synsets, which are then used to train other unlabelled data iteratively. The classifier is used to determine to ”PN-polarity” as well as ”SO-polarity” of words. ”PN-polarity” represents the degree of which a word is positive or negative, which closely resembles sentiment analysis. ”SO-polarity” is a binary classification of whether the text is factual, which is closely-related to subjectivity categorisation. The proportion of positive, negative and objective components is also investigated for various parts of speech such as adjectives, names, verbs and adverbs. The algorithm to generate scores in the SentiWordNet lexicon is further refined by Baccianella, S., Esuli, A., and Sebastiani, F. (2010) by using random walk.

Moreover, Khanna S. and Shiwani S. (2013) compared two methods to perform subjectivity detection, namely; developing a learning model for subjectivity detection based on Naïve Bayes classifier and using the SentiWordNet method. Five different rules for detecting and extracting phrases are defined. Both approaches use different formulae to calculate objectivity and subjectivity scores for unigrams and phrases but classification rule is the same; if the subjectivity score exceeds objectivity score, resulting classification is subjective and vice versa. It is determined that SentiWordNet outperformed Naïve Bayes classifier in terms of accuracy of subjectivity detection. Phrases are found to be more effective than unigrams in capturing sentiment, however it is yet to be determined whether this observation holds true for other languages.

Furthermore, Chenlo J. M. and Losada D. E. (2013) proposed an approach to subjectivity classification using sentence features. The features used to construct a subjectivity classifier are unigram and bigram, sentiment lexicon, rhetorical, length and positional features. Weight could also be applied to features, where a positive weight indicates subjectivity and a negative weight indicates objectivity. It is found that using unigram and bigram representation in combination with the aforementioned sentence features resulted in higher overall precision, recall and F1 measure as compared to OpinionFinder, which is based on Naïve Bayes classifier.

Riloff, E. and Wiebe, J. (2003) developed a bootstrapping process for subjectivity classification of sentences. Instead of relying on manual annotation of sentences to determine their subjectivity, a computational approach is taken instead. Subjectivity classifiers first attempt to label sentences as either subjective or objective with high confidence. Extraction patterns are generated from remaining unlabelled sentences to further determine the subjectivity of sentences. Overall, it is shown that the bootstrapping process is a viable alternative to using human evaluators for subjectivity detection.

Both the MPQA lexicon and SentiWordNet lexicon play an important role in the

evaluation of the subjectivity lexicon that is generated from Wikipedia in this Master Thesis. The MPQA lexicon provides categorical subjectivity classification, namely "strongsubj" (strongly subjective) and "weaksubj" (weakly subjective). The SentiWordNet lexicon offers a numerical scoring system for synsets, which can be compared with the scores obtained by using quotation detection in Wikipedia.

### 2.0.4 Human Evaluators

Although the subjectivity lexicon is generated using computational methods, human evaluators are still required to validate the subjectivity of words in Wikipedia, especially those which are not present in other subjectivity lexica like MPQA and SentiWordNet. Certain best practices when engaging human evaluators are adopted from literature.

Jansen, K. J., Corley, K. G. and Jansen, B. J. (2007) discuss the advantages and disadvantages of various survey approaches; namely web-based, email-based and point of contact. The key benefits of web-based survey, such as low turnaround time, scalability, confidentiality and ease of implementing multiple choice questions, highlighted in this paper justify the usage of web-based survey instead of using Email or face-to-face approach. Certain drawbacks of web-based approach mentioned in the paper, such as time-consuming development and issues with user accessibility, are eliminated with the use of Google Form, which is easy to use for both the creator and participants of the questionnaire. In addition to that, Franklin, S. & Walker, C. (2010) provide a comprehensive overview on various stages of conducting a survey, such as questionnaire design, data collection and processing methods as well as analysis of results. It is used as a reference guideline while designing the online questionnaire, especially with regards to the structure and wording of the questionnaire.

# 3 Research Design

The process of transforming raw text data in Wikipedia corpus into a subjectivity lexicon involves several key steps. A suitable Wikipedia corpus must be found, and then the text data must be ”cleaned” prior to the process of quotation detection and implementation of subjectivity lexicon. Furthermore, a questionnaire is prepared as part of the evaluation process of the subjectivity lexicon. The key procedures implemented in this research are detailed in the following subsections below.

## 3.1 Data Sourcing

The English Wikipedia is selected as the focal point of this research. An existing repository of the English Wikipedia (last updated August 2016) from f-squared (n.d.) is selected as the text corpus. Each Wikipedia article is stored as a HTML file within a compressed sub-folder, which in turn is stored in a major folder. A total of 485 major folders were extracted, each containing 10000 sub-folders (ie: 10000 HTML files). As such, the total data-set would constitute of 4.85 million HTML files. The HTML files have been named based on an assigned ID, whereas the major folders are named numerically. The contents of the major folders have been manually inspected in a random manner and there was no indication that the Wikipedia articles in the HTML files in each major folder were sorted alphabetically or classified by category. A total of 2500 Wikipedia articles are extracted randomly from each major folder (25 percent sample), resulting in a total of 1212500 English Wikipedia articles which constitute the text corpus used in this Master Thesis.

## 3.2 Data Cleaning

Data cleaning is performed on two different scales, firstly on a corpus level and subsequently on a word level. At corpus level, ”Beautiful Soup” Python library is used, whereas individual words are ”cleaned” before being included in the subjectivity lexicon.

### 3.2.1 Corpus-level cleaning

As the raw Wikipedia corpus is in HTML format, corpus-level cleaning is required to extract text data. In this Master Thesis, ”Beautiful Soup”, which is a Python library that functions as a parser, is used to extract only the paragraph sections of Wikipedia articles contained within the HTML files. This library is chosen due to its ease-of-use in achieving the desired text output. Moreover, according to Zheng, C., He, G. and Peng, Z. (2015), ”Beautiful Soup” is very flexible across multiple platforms, which is a favourable trait with regards to scalability of this research. By extracting only words that are contained within the paragraph tags, words in titles, tables, captions, bullet points and reference list are not considered. This eliminates non-relevant words such as names and those contained within website URLs.

### 3.2.2 Word-level cleaning

Each word that has been parsed in ”Beautiful Soup” and stored in a list undergoes a ”cleaning” process. This step is essential to ensure standardisation of words before they are counted and appended to the word-frequency dictionary. First and foremost, a Regex is applied to remove Wikipedia referencing numbers that may be attached to words. For example, the pair of square brackets and its content in the word ”sample[1]” are removed, resulting in ”sample”. Nevertheless, if a word still contains numbers even after the removal of Wikipedia referencing numbers, the word will be excluded as it is not meaningful in the terms of subjectivity analysis.

Besides that, words that correspond to names and places need to be removed because assigning a Quote Score to such words is also not meaningful. Since is it not possible to have an exhaustive list of all names and places, an assumption is made that all words that begin with a capital letter are most likely names and places. The drawback of applying such a rule is that the first word of every sentence will be excluded, however it is also not uncommon that sentences begin with either stopwords such as ”a”, ”an” and ”the”. The necessity of removing names and places outweighs the loss of some words as a result of excluding capitalised words. This method is effective for the English Language, however for other languages in which all nouns are capitalised, such as German, an alternative method is required. Once the capitalised words are eliminated, the remaining words are all set to lowercase to address cases where there are unexpected capitalised letters in the middle of the words.

While some stopwords may be already filtered out as a result of removing capitalised letters, it is also necessary to remove all other common stopwords in the text corpus. This is because stopwords tend to occur very frequently in sentences and may skew the overall distribution of Quote Scores. Each word is compared against a lexicon containing 119 common stopwords in English. If a word matches any of the 119 words in that lexicon, it will be excluded. Furthermore, all punctuation are also removed from words, with the exception of quotation marks. Words such as ”haPPy-go-lucKy!” and ”happy–golucky?” will be standardised as ”happy-golucky” and deemed as the same word.

Last but not least, all ”cleaned” words are validated against the ”Wordlist” corpus in the Natural Language Toolkit (NLTK). NLTK is selected due to its large selection of features to choose from; there are more than 50 corpora as well as various text processing libraries (Farkiya, A., Saini, P. and Sinha, S., 2015). The ”Wordlist” corpus in NLTK is selected as a basis of comparison in determining whether words extracted from Wikipedia are English words. It contains 236736 words which is observed to consist primarily of nouns and adjectives. For example, in addition to the stem word ”adapt”, there are many forms and variations of the word in NLTK:

- adaptability
- adaptable
- adaptation
- adaptational
- adaptationally
- adaptative
- adaptedness
- adapter
- adaption
- adaptional
- adaptionism
- adaptitude
- adaptive
- adaptively
- adaptiveness
- adaptometer
- adaptor
- adaptorial

However, it is observed that verb conjugations for the simple present (he/she/it ”adapts”), present continuous tense (”adapting”) as well as past tense (”adapted”) are often missing. However, this trade-off is essential to ensure that the words exist in the English dictionary and can be compared against words in other existing lexica.

### 3.3 Quotation Detection

Quote words are words which are enclosed within an opening and closing quotation mark. In the English Wikipedia, the opening and closing quotation marks are identical. Assuming that the text corpus is error-free, every odd number of occurrence of the quotation mark (first, third, fifth...) indicates the start of a quote, whereas every even number of occurrence of the quotation mark (second, fourth, sixth...) indicates the end of a quote. Therefore, the quotes would be in between the first and second, third and fourth, fifth and sixth quotation marks and so on respectively. However, in practice, the quotation marks are occasionally missing, resulting in significantly different outcome if such a method is used. An example is given below:

*She says, **”We should retreat!”**. However, the so-called **”all-knowing oracle”** shouts, **”I disagree!”** and storms out of the **”strategy room”**. She is shocked by his response. After that, she says, **”Maybe you are right... We should fight on!!!”***

Given that the text shown above is ”error-free”, the desired quote words can be extracted correctly by systematically looking for text between the first and second quotation marks, third and fourth quotation marks and so on. The list of quotations in the following page matches those which are boldfaced in the text.

['We should retreat!', 'all-knowing oracle', 'I disagree!', 'strategy room', 'Maybe you are right... We should fight on!!!']

Using the same example text as in the previous page, certain quotation marks are removed to simulate the issue encountered while trying to extract quotes from Wikipedia text. The phrase "all-knowing oracle" is lacking an opening quotation mark and "strategy room" is lacking a closing quotation mark.

*She says,* ***"We should retreat!"****. However, the so-called* ***all-knowing oracle"*** *shouts,* ***"I disagree!"*** *and storms out of the* ***"strategy room****. She is shocked by his response. After that, she says,* ***"Maybe you are right... We should fight on!!!"***

When some quotation marks are missing, using the same methodology as before will yield significantly different results. Certain phrases are incorrectly identified as quotations, as a result, several words which are actually not in quotes are incorrectly classified as quote words.

['We should retreat!', ' shouts, ', ' and storms out of the ', 'Maybe you are right... We should fight on!!!']

In cases where quotation marks are missing, there is no definite way of determining the intended start or end point of the quotation. The approach taken in this Master Thesis is to perform discovery of quotation marks one word at a time in a sequential manner. For a given word, if a quotation mark is present at either the first or second character position in the word, it is determined that an opening quotation mark has been found. On the other hand, if a quotation mark is present at either the last or second last character position in the word, it is determined that a closing quotation mark has been found. By identifying the type of quotation mark, only words within a complete pair of opening and closing quotation marks are deemed to be quotes. In the event where an opening or closing quotation mark is missing, resulting in 2 consecutive opening or closing quotation being encountered, the words in between them are treated as non-quote words. Applying this method onto the text with a few missing quotation marks above, the output would be as follows:

['We should retreat!', 'I disagree!', 'Maybe you are right... We should fight on!!!']

Note that not all quote phrases are detected using this method. However, this conservative approach is taken when attempting to extract quote words because it is preferred to exclude phrases with incomplete quotation marks than to incorrectly classify a large amount of non-quote text as quotes. This is particularly important in Wikipedia where quotations are expected to be relatively sparse; a missing quotation mark could result in a large portion of non-quote text being incorrectly classi-

fied as quotes. A conservative approach may result in fewer quotes being detected, but the risk of incorrectly classifying non-quote words as quotes is also lower.

### 3.4 Computation of Quote Score

After quotation detection is completed, the end result is two lists, one for words found within quotation marks (i.e: quote words), and the other for all other words in Wikipedia. Using the ”Counter” function in Python, the frequencies of words are represented as a dictionary, where each key corresponds to a word and the value corresponds to its frequency in the corpus.

The Quote Score is defined as the ratio of frequency of a particular word in quotes with respect to its frequency in the entire text corpus. Since the word frequencies are already stored as values in the dictionaries, the Quote Scores for all words can be calculated easily. The numerator is the frequency of word in quotes and the denominator is the sum of word frequencies in quotes as well as in normal text. The computation of Quote Score is depicted in the example below:

Dictionary for quote words - (happy:13 , holiday: 1)
Dictionary for normal words - (happy:7 , holiday: 4)

Quote Score for ”happy”:

$$\frac{13}{13+7} = 0.65$$

Quote Score for ”holiday”:

$$\frac{1}{1+4} = 0.2$$

Based on the calculations above, the word ”happy” has higher Quote Score than ”holiday”, hence it can be concluded that the word ”happy” is more subjective than ”holiday”. The Quote Score of a word lies between 0 and 1, where 0 represents a purely objective word and 1 represents a purely subjective word. By keeping the scores normalised between 0 and 1, it is possible to compare the Quote Scores with results from other related work, such as SentiWordNet.

An issue encountered when using this formula is that words which occur very rarely in the text corpus will be assigned either very high or very low Quote Scores. For example, a word which only occurs once in the whole corpus and within quotation marks will have a Quote Score of 1. Likewise, a non-quote word which occurs just once in the whole corpus will have a Quote Score of 0. As a result, rare words

would skew the Quote Score distribution. In order to mitigate this issue, a minimum threshold of 10 is set so that rare words will not be included in the lexicon. The selection of minimum threshold value is not trivial. The minimum threshold value should not be too low, or else there is no noticeable impact on the Quote Score distribution. On the other hand, setting a minimum threshold value that is too high would result in too many words being excluded from the Wikipedia lexicon. After taking into consideration the aforementioned requirements, it is decided that minimum threshold value of 10 would be used.

## 3.5 Human Evaluators - Word Subjectivity Questionnaire

In addition to lexica from previous researches, human evaluators are also necessary to supplement the validation of results of this Master Thesis. A questionnaire which consists of 100 words is prepared, which is approximately one percent of 10032 words in the lexicon generated from the English Wikipedia which are not present in MPQA and SentiWordNet lexica. Participants are tasked with determining whether a given word is strongly subjective, mildly subjective or objective. The sequence of questions is always randomised for each new attempt. Unknown to the participants, these 100 words are actually divided into 2 groups:

- Group 1: Quality control group (20 words). The subjectivity categorisation and scores of these words are already established, but are included in the survey so as to gauge the reliability of the human evaluator in rating assignment. These 20 words are randomly selected from a small pool of words which adhere to the following criteria:
  - Condition 1: Difference between SentiWordScore and Quote Score of the word is less than 0.05
  - Condition 2: If the word is classified as ”strongsubj”, both SentiWordScore and Quote Score of the word are above their respective means. If the word is classified as ”weaksubj”, both SentiWordScore and Quote Score of the word are below their respective means.
- Group 2: Words pending evaluation (80 words). These 80 words are randomly sampled from the 10032 words which have been neither evaluated by MPQA nor SentiWordNet lexicon.
  - ”High” bucket: Top 500 words with the highest Quote Scores
  - ”Low” bucket: Words which have Quote Score below the mean of Quote Score (6236 words)
  - ”Medium” bucket: The remainder words that are not in the other 2 buckets (3296 words)

The buckets of words in Group 2 are created with unequal sizes. 30 words are randomly selected from the High Bucket, which constitute highly-subjective words. 20

words are randomly selected from the Low Bucket, which represent the group of words with low degree of subjectivity. Likewise, 30 words are drawn randomly from the Medium Bucket, resulting in a total of 80 words to be evaluated.

With the 100 words ready, the next step was to create the questionnaire. The questionnaire is designed based on the reference guideline of Statistics Canada by Franklin, S. & Walker, C. (2010), which encompasses several key factors such as method of data collection, characteristics of respondents, response burden, complexity of the data to be collected, confidentiality and sensitivity of the information and consistency. Each and every factor that was taken into consideration during the design of the questionnaire will be elaborated in detail below.

### 3.5.1 Method of Data Collection

Data collection is conducted via an online survey form, which is easily shareable to participants using a URL. The platform of choice is Google Form because it is a free to use and has a simple user interface. Participants answer questions in the questionnaire via multiple choice, which will prevent any unintended input errors such as spelling mistakes. The 3 options to choose from are ”Strongly Subjective”, ”Mildly Subjective” and ”Objective”. An additional ”I don’t know” option is also included to discourage participants from randomly selecting an answer should they find a given word unfamiliar or too difficult.

### 3.5.2 Characteristics of Respondents

As the questionnaire consists of English words in which the subjectivity levels are evaluated, participants who either have English as their mother tongue or first language are selected to participate in this questionnaire. A high level of fluency in the English language is required as there are some uncommon words in the questionnaire which are most likely unfamiliar to beginner and intermediate English language users.

### 3.5.3 Response Burden

There is a total of 100 questions in the questionnaire in which each question is estimated to take between 6 to 9 seconds to answer. As a result, the total duration of the questionnaire should not exceed 15 minutes. In addition to that, the usage of multiple choice as a means of answering the questionnaire will not require the participants to use the keyboard. Furthermore, the option of selecting ”I don’t know” is given for each question just in case a participant does not understand a given word, so that he/she is not obliged to look up the meaning of the word in the dictionary.

### 3.5.4 Complexity of the Data to be Collected

While the intended participants of the questionnaire are fluent English users, it is not assumed that they are familiar with the terminologies related to subjectivity analysis. As such, simple definitions of the multiple choice options, ”Strongly Subjective”, ”Mildly Subjective” and ”Objective” are given, as well as examples of words for each corresponding option. In addition to that, a simple scenario is included to assist the participants in understanding the given multiple choice options clearly.

### 3.5.5 Confidentiality and Sensitivity of the Information

While the creation of questionnaire in Google Form requires a Google account, the participants are not required to have one. The link to the questionnaire is shared via social media and the participation is anonymous as there is no information collected from the participants other than their responses in the questionnaire.

### 3.5.6 Consistency

The question and answer format is standardised throughout the questionnaire. The question consists of only one word, and the multiple choice answers are always ”Strongly Subjective”, ”Mildly Subjective”, ”Objective” and ”I don’t know”. Examples are also included in the instructions at the top section of the questionnaire so that the task is clear right from the beginning.

# 4 Results

Several key findings of this Master Thesis are summarised as follows:

- The hypothesis that all quotes in Wikipedia are subjective and the other words are objective is tested. A subjectivity lexicon, which contains the list of words in Wikipedia along with their respective Quote Scores, is generated with moderate success. The proportion of words in the English Wikipedia which are within quotation marks is found to be much smaller as compared to those which are not in quotes. Mean value of Quote Score is low and the distribution of Quote Score of words is right-skewed.
- Words which are strongly subjective tend to have high Quote Scores in the subjectivity lexicon generated from English Wikipedia. There is a large observable difference between distribution of Quote Score for words classified as strongly subjective versus distribution of Quote Score for words classified as weakly subjective and objective. However, weakly subjective and objective words cannot be differentiated clearly based on Quote Score.
- The generated subjectivity lexicon contains more additional words which are not present in existing lexica such as MPQA and SentiWordNet. While this presents an opportunity to expand existing subjectivity lexica, the caveat is that the additional words are inferred to be rather complex or not commonly used. Human evaluators perceived a large majority of these words as objective although a mix of objective, mildly subjective and highly subjective words were included in the questionnaire.

## 4.1 Distribution of Quote Scores in English Wikipedia

The results of this Master Thesis are based on a random sample of 1212500 English Wikipedia articles. Total number of words contained in the generated lexicon is 86771. A minimum frequency threshold of 10 is set in order to eliminate infrequent words from the corpus, hence cutting down the corpus size to 44613 words. Elimination of infrequent words is crucial because Quote Score is calculated based on ratio of a word occurring in a quote versus its frequency in the corpus. As such, words which have very low frequencies in the corpus tend to have extreme Quote Score values of either 0 or 1, which are excluded with the threshold.

Once the data has been trimmed down in size, the first major step was to visualise the distribution of Quote Score and then compute the essential statistical measures. The distributions of Quote Scores before and after applying the threshold are represented in Figures 1 and 2 on the following page. Note that the y-axes of the histograms use a logarithmic scale, whereas the x-axes use a linear scale.

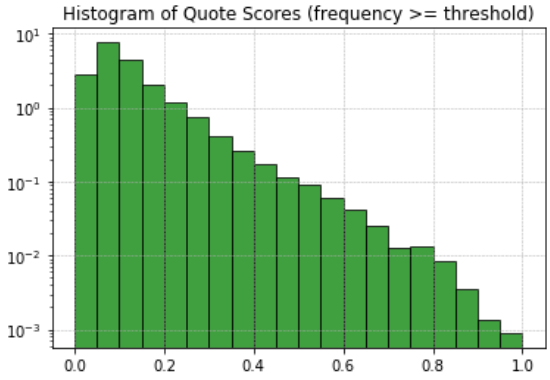


Figure 1: Histogram of Quote Scores (minimum threshold = 10)

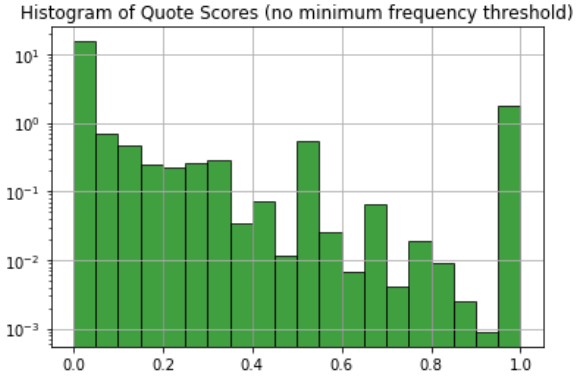


Figure 2: Histogram of Quote Scores (no threshold applied)

Based on Figure 1, it can be observed that the distribution of Quote Scores is skewed towards zero. The skewness value is 123.13, which shows that distribution is very right-skewed. This finding also indicates that there are fewer words which are within quotations as compared to regular text in the English Wikipedia.

On the other hand, it can be observed in Figure 2 that the furthest left and right columns, which represent Quote Score values close to 0 and 1, have significantly larger values than the rest. This demonstrates the impact of words with low frequencies on the distribution of Quote Scores, hence the minimum frequency threshold must be set in order to mitigate this effect. All results and discussions henceforth will be based on the trimmed group of 44613 words, where a minimum frequency threshold of 10 have been applied.

The mean, median and standard deviation of the Quote Score distribution with threshold applied are calculated as follows (rounded to 5 significant figures):

Mean: 0.12514
Median: 0.097192
Standard Deviation: 0.10188

As mentioned above, the distribution is right-skewed towards zero, hence it is not surprising that the mean and median values are relatively low.

## 4.2 Evaluation with MPQA Lexicon

MQPA is the acronym for ”Multi-Perspective Question Answering”, which is a source of various text lexica and corpora, such as the MQPA Opinion Corpus, Subjectivity Lexicon and OpinionFinder System. In this Master Thesis, the MPQA Subjectivity Lexicon is used to evaluate the Quote Score assignment to words with respect to their subjectivity classifications. The MPQA lexicon consists of several attributes as follows:

- Subjectivity classification - ”strongsubj” or ”weaksubj”
- Length of word - all entries are single words, hence the length is always 1
- Type of word - noun, adjective, verb, adverb or any position
- Stem flag - indicates whether the word is a stem word
- Prior polarity - positive or negative

Nevertheless, the only attribute of interest in the MPQA lexicon with respect to the Quote Scores in the lexicon generated from the English Wikipedia is the subjectivity classification. It is investigated whether words that are classified as ”strongsubj” in the MPQA lexicon tend to have higher Quote Scores as compared to those classified as ”weaksubj”. A word which is classified as ”strongsubj” is a strong subjective clue, meaning that it is subjective in most contexts. On the contrary, a word which is classified as ”weaksubj” is a weak subjective clue, meaning that it is subjective only in certain contexts.

### 4.2.1 The ”MPQA-Wiki Lexicon”

The MPQA lexicon consists of a total of 8222 entries, however there are some words which occur more than once in the lexicon. This is because a word may have different attributes, depending on its usage and context. For example, the word ”diplomatic” appears twice in the MPQA lexicon. In both entries, it is classified as ”weaksubj”, has length of 1, not a stem word and has positive prior polarity; however the

word ”diplomatic” has 2 word types; one is an adjective and the other is a noun. Interestingly, even if a word occurs more than once in the MPQA lexicon due to variations in attributes, its subjectivity classification is found to be always the same for all its entries. Therefore, taking into account only the subjectivity classification attribute in MPQA lexicon, the duplicate entries can be removed and the lexicon is condensed into 5711 unique words.

There is a total of 4639 words which are found to be in common between the lexicon generated from Wikipedia (44613 words) and that of MPQA (5711 words). This group of words is used as a basis of evaluation since each word has been assigned a Quote Score as well as categorised in MPQA, and will be henceforth referred to as **”MPQA-Wiki Lexicon”**. The Venn Diagram in Figure 3 illustrates the overlap between the two lexica. Moreover, the distribution of Quote Scores in ”MPQA-Wiki Lexicon” is visualised in Figure 4. Note that both the x-axis and y-axis in the histogram have linear scales.

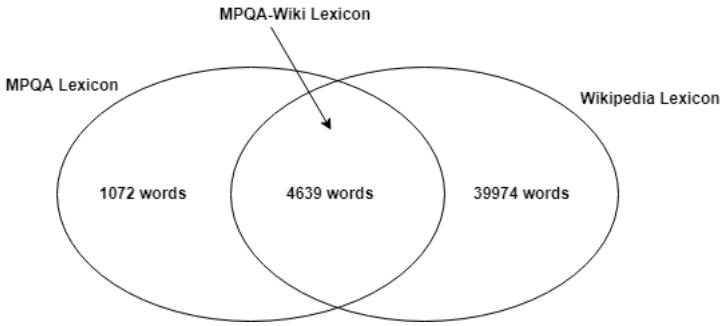


Figure 3: Venn Diagram to show the number of overlapping words between the lexicon generated from Wikipedia and the MPQA lexicon

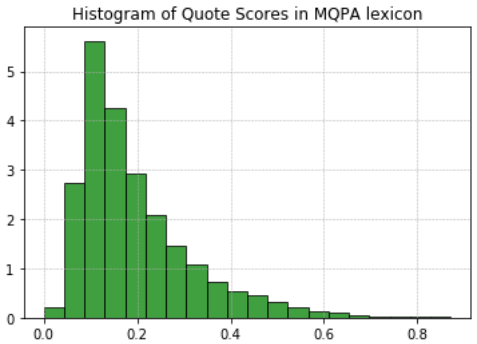


Figure 4: Histogram of Quote Scores in MPQA-Wiki Lexicon

The histogram in Figure 4 in the previous page is also right-skewed, however not as much as compared to the distribution of Quote Scores for the lexicon generated from Wikipedia because it has a lower skewness value of 31.935. The mean, median and standard deviation of the Quote Score distribution of ”MPQA-Wiki Lexicon” are calculated below:

Mean: 0.19538
Median: 0.15824
Standard deviation: 0.12266

Both the mean and median Quote Scores of the ”MPQA-Wiki Lexicon” are observed to be larger than that of the lexicon generated from Wikipedia. This observation suggests that on average, words that are in common between MPQA lexicon and Wikipedia lexicon tend to be more subjective as compared to words in the entire Wikipedia lexicon itself. This initial observation is intuitive as words in the MPQA lexicon are called ”subjectivity clues” so it is not unexpected to obtain higher overall mean and median Quote Scores for those words. This finding is investigated further by the creation and comparison between word groups.

### 4.2.2 Comparison between Word Groups

The words in the ”MPQA-Wiki lexicon” are divided into 3 groups in order to investigate the relationship between Quote Score and the corresponding subjectivity classifications.

Group S: Words in ”MPQA-Wiki Lexicon” which are classified as ”strongsubj” (3069 words)
Group W: Words in ”MPQA-Wiki Lexicon” which are classified as ”weaksubj” (1570 words)
Group N: Words in lexicon generated from Wikipedia which are not in ”MPQA-Wiki Lexicon” (39997 words)

If the difference in mean Quote Scores among the 3 groups is large enough, it can be said that the Quote Scores obtained from Wikipedia fit the subjectivity classification of words in MPQA lexicon. The mean, median and standard deviation for each group are calculated as follows:

| Statistical Measure | Group S | Group W | Group N |
|---|---|---|---|
| Mean | 0.22532 | 0.13686 | 0.11702 |
| Median | 0.19212 | 0.11466 | 0.092409 |
| Standard Deviation | 0.13238 | 0.070929 | 0.095918 |

Table 1: Statistical measures for the word groups

It can be observed in Table 1 in the previous page that median and mean Quote Scores for words which are classified as ”strongsubj” in MPQA (Group S) are larger than those classified as ”weaksubj” (Group W). However, the difference in means and medians between the words classified as ”weaksubj” in MPQA (Group W) and those not in MPQA (Group N) are not as large as in the aforementioned comparison. The standard deviations of all 3 groups are not very different from each other. As such, the following measures are calculated to verify significance of difference between the 3 word groups.

- Cohen's d effect size - quantifiable measure of difference between the two word groups
- Cohen's U3 - proportion of words in a group which have Quote Scores above the mean of the other group
- Overlap coefficient - proportion of overlap between the two groups
- Common language effect size - probability that a randomly-selected word from a group will have higher Quote Score than another randomly-selected word from the other group

Before analysing the results, it is crucial to note the baseline for the comparison, which is when the value of cohen's d effect size is equal to 0. When cohen's d effect size is 0, the overlap coefficient is 1 because both distributions overlap each other completely. In such a case, values of cohen's d U3 and common language effect size are both 0.5. Therefore, the results of pairwise comparison among all 3 groups are summarised in the table below.

| Measure | Group S vs W | Group W vs N | Group S vs N |
|---|---|---|---|
| Cohen's d effect size | 0.83306 | 0.23518 | 0.93694 |
| Cohen's U3 | 0.79759 | 0.59296 | 0.82561 |
| Overlap coefficient | 0.67702 | 0.90639 | 0.63945 |
| Common language effect size | 0.72209 | 0.56604 | 0.74618 |

Table 2: Pairwise comparison of Effect Sizes for the 3 groups of words

Words classified as ”strongsubj” (Group S) vs ”weaksubj” (Group W)

A large Cohen's d effect size indicates that the difference between the two distributions is large. More specifically, Cohen's U3 value of 0.79759 means that 79.759 percent of words classified as ”strongsubj” will have Quote Scores greater than the mean Quote Score of words classified as ”weaksubj”. As compared to the baseline of 50 percent, this is a relatively large proportion of words. There is a moderate overlap of 67.702 percent between the 2 distributions. The common language effect size shows that the probability of a word chosen randomly from the ”strongsubj”

group having a higher Quote Score than another word chosen randomly from the ”weaksubj” group is 0.72209. This probability value is significantly higher than the baseline of 0.5. As such, it can be concluded that there is a large and significant difference in Quote Scores when comparing the group of words classified as ”strongsubj” and those classified as ”weaksubj” in the MPQA lexicon.

Words classified as ”weak” (Group W) vs non-MPQA words (Group N)

A small Cohen's d effect size indicates that the difference between the two distributions is small. More specifically, Cohen's U3 value of 0.59296 means that 59.296 percent of words classified as ”weaksubj” will have Quote Scores greater than the mean Quote Score of words which are not part of the MPQA lexicon, which is only slightly more than the baseline of 50 percent. There is a significantly large overlap of 90.639 percent between the 2 distributions. The common language effect size shows that the probability of a word chosen randomly from the ”weaksubj” group having a higher Quote Score than another word chosen randomly from group of words that are not part of the MPQA lexicon is 0.56604, which is only slightly more than the baseline value of 0.5. As such, it can be concluded that there is only a small difference in Quote Scores when comparing the group of words classified as ”weaksubj” and words which are not part of the MPQA lexicon.

Words classified as ”strong” (Group S) vs non-MPQA words (Group N)

The Cohen's d effect size for this comparison is the largest as compared to the previous 2 comparisons. Given that there is already a large difference between the distributions of words classified as ”strongsubj” and ”weaksubj”, as well as a small difference between distribution of words classified as ”weaksubj” and those not in the MPQA lexicon, this observation is intuitive. The values of Cohen's U3 and common language effect size are also the largest, and degree of overlap is the lowest.

Based on the results obtained above, the following conclusions can be made:

- Distribution of words in the MPQA lexicon has a higher mean and median Quote Scores as compared to distribution of words which are not part of the lexicon. The higher mean and median Quote Scores for distribution of words in MPQA validate the fact that they are subjective clues.
- Within the MPQA lexicon, words which are classified as ”strongsubj” tend to have higher Quote Scores as compared to those which are classified as ”weaksubj”. This is proven by large and significant difference in means of distributions of Quote Scores for these 2 categories.
- Therefore, for words which are common between the MPQA lexicon and the lexicon generated from Wikipedia, the subjectivity categorisation of words corroborates the corresponding Quote Scores.

## 4.3 Evaluation with SentiWordNet Lexicon

In addition to evaluation with subjectivity categorisation in MPQA lexicon, a lexicon with numerical subjectivity scoring system for words is required to validate the assignment of Quote Score values to words. SentiWordNet fulfils this requirement as it is a lexical resource for open mining in which words are assigned a Positive Score and a Negative Score. The sum of Positive Score and Negative Score lies between 0 and 1 and represents the strength of sentiment in a word. The Objective Score is defined in the SentiWordNet documentation as the complement of this sum, as shown in the equation below:

$$ObjectiveScore = 1 - PositiveScore + NegativeScore$$

This equation could be interpreted in another way; the degree of subjectivity for words in SentiWordNet could be deduced to be simply the sum of Positive Score and Negative Score, henceforth referred to as **"SentiWordScore"**. "SentiWordScore" from the SentiWordNet lexicon is then compared with the Quote Score of words in the lexicon generated from Wikipedia, namely the Quote Score. The equation below summarises the relationship between the aforementioned terminologies:

$$QuoteScore \equiv SentiWordScore = PositiveScore + NegativeScore$$

### 4.3.1 The "SentiWordNet-Wiki Lexicon"

The SentiWordNet lexicon contains 206941 entries; however there are words which appear in multiple entries because a word may convey varying degree of positive and negative sentiments, resulting in several SentiWordScore values. In those cases, the mean SentiWordScore of the word is calculated and the duplicate entries are removed, resulting in 147292 unique words. The lexicon of words is manually inspected and it is found that a large number of words are not regular English words as they contain numbers and punctuation such as underscore and fullstop. As the focal point of this Master Thesis is on English words in Wikipedia, these non-English words need to be excluded so as to fairly compare words in the SentiWordNet lexicon and the lexicon generated from Wikipedia.

The lexicon of words is compared against the NLTK (Natural Language Toolkit) word corpus to check whether the words belong to the English dictionary. As a result, the SentiWordNet lexicon is further reduced to 51871 words. These words are then compared with the lexicon generated from Wikipedia and 34511 unique words are found to be in common with the lexicon generated from Wikipedia. The Venn Diagram in Figure 5 on the following page illustrates the overlap between the two lexica.

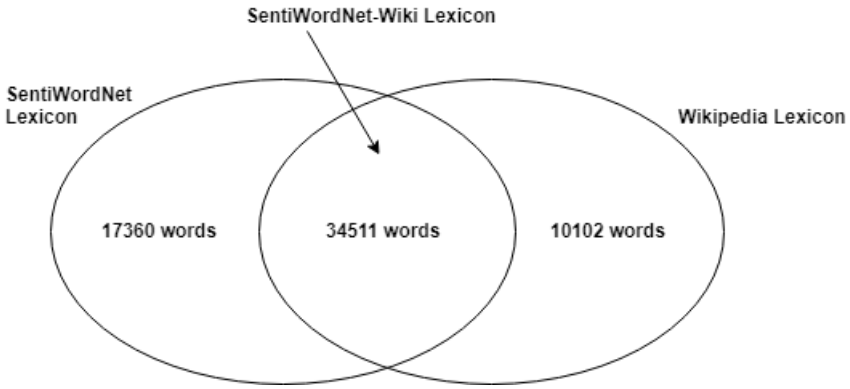


Figure 5: Venn Diagram to show the number of overlapping words between the lexicon generated from Wikipedia and the SentiWordNet lexicon

Based on Figure 5 above, 77.36 percent of words in the lexicon generated from Wikipedia could be matched with words in the SentiWordNet lexicon. The number of matching words with SentiWordNet lexicon is much higher than that of MPQA lexicon since the SentiWordNet lexicon contains significantly more words. This group of words in common will be referred to as **"SentiWordNet-Wiki Lexicon"**. The distributions of Quote Score as well as SentiWordScore for words within the "SentiWordNet-Wiki Lexicon" are visualised in Figure 6 below. Note that the y-axis uses a logarithmic scale, whereas the x-axis uses a linear scale in the histogram.

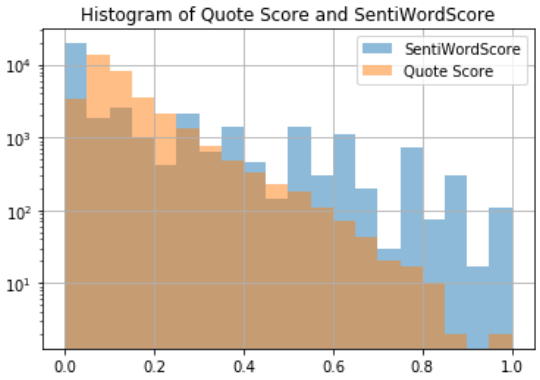


Figure 6: Histogram of Quote Score and SentiWordScore

Based on Figure 6 in above, it can be observed that both distributions are right-skewed, with a large majority of words having low values of Quote Score and SentiWordScore. The mean, median and standard deviation for both distributions are calculated, as shown in Table 3 on the following page.

| Statistical Measure | Quote Score | SentiWordScore |
|---|---|---|
| Mean | 0.13098 | 0.14571 |
| Median | 0.10077 | 0 |
| Standard Deviation | 0.10031 | 0.22195 |

Table 3: Statistical measures for Quote Score and SentiWordScore

It is observed in Table 3 that the means of distribution of Quote Score and SentiWordScore are quite similar to each other. However, the distribution of SentiWordScore has a significantly larger standard deviation as compared to that of Quote Score.

### 4.3.2 Relationship between Quote Score and SentiWordScore

The Kendall's Tau B as well as Spearman's rank correlation coefficient are calculated as a means of measuring the degree of agreement between the two subjectivity scoring systems.

- Kendall's Tau B = 0.24374 (p-value = 0)
- Spearman's rank correlation = 0.32541 (p-value = 0)

Since both Kendall's Tau B and Spearman's rank correlation coefficient have values above zero, but are relatively small, it can be deduced that there is mild agreement between the two subjectivity scoring systems. The p-values for both tests are zero, which means that the results of the tests are significant. Furthermore, since both subjectivity scoring systems are numeric-based and have the same range of 0 to 1, a plot is generated to visualise the correlation between Quote Score and SentiWordScore for words in the "SentiWordNet-Wiki Lexicon".

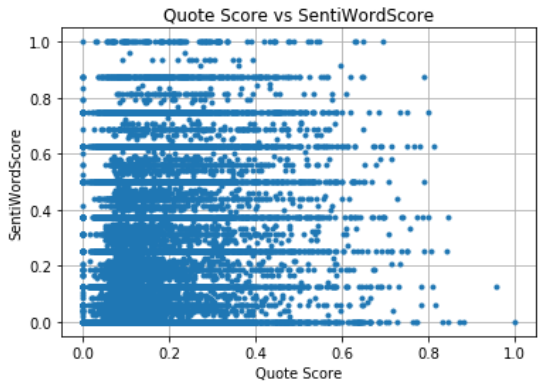


Figure 7: Scatterplot of Quote Score and SentiWordScore

The scatterplot in Figure 7 in the previous page shows that the points are mostly concentrated at the lower left corner of the plot, which verifies the observation that both distributions are right skewed. Another key observation from this plot is that there are many points which form multiple horizontal ”lines”, which indicates the coarse granularity of SentiWordScore assigned to words. Many words have the exact same SentiWordScore, which would result in large number of tied ranks if Mann-Whitney U test or Wilcoxon signed rank test were to be used. As such, effect size is chosen as the method of evaluation of whether the Quote Score generated from Wikipedia is corroborated by SentiWordScore. The results are as shown below:

- Cohen's d effect size = 0.085472
- Cohen's U3 = 0.53406
- Overlap coefficient = 0.96591
- Common language effect size = 0.52410

A small Cohen's d effect size indicates that the difference between the two distributions is small. The overlap coefficient is very high; there is a 96.591 percent overlap between the 2 distributions. If a word is randomly selected from each distribution, probability of the word from the ”SentiWordScore” distribution having a higher Quote Score than another word chosen randomly from the Quote Score distribution is 0.52410, which is only slightly higher than the baseline of 0.5.

Based on the results obtained above, the following conclusions can be made:

- The distributions of Quote Score and SentiWordScore have a very high degree of overlap with each other.
- Positive values of rank correlation coefficients between Quote Score and SentiWordScore of words show that they are mildly in agreement.
- Therefore, for words which are common between the SentiWordNet lexicon and the lexicon generated from Wikipedia, the SentiWordScore corroborates the corresponding Quote Score of words.

## 4.4 Validation with Human Evaluators

In addition to evaluation against lexica from previous researches, the lexicon generated from the English Wikipedia is also validated using human evaluators. This is particularly important as there are words in the Wikipedia lexicon which are neither in the MPQA lexicon nor in the SentiWordNet lexicon. The Venn diagram in Figure 8 below shows that there is a remainder of 10032 words out of 44613 words (22.487 percent) in the Wikipedia lexicon which are pending review by human evaluators. Subsequently, a questionnaire is created to determine whether the Quote Score of words is corroborated by assessment of human evaluators.

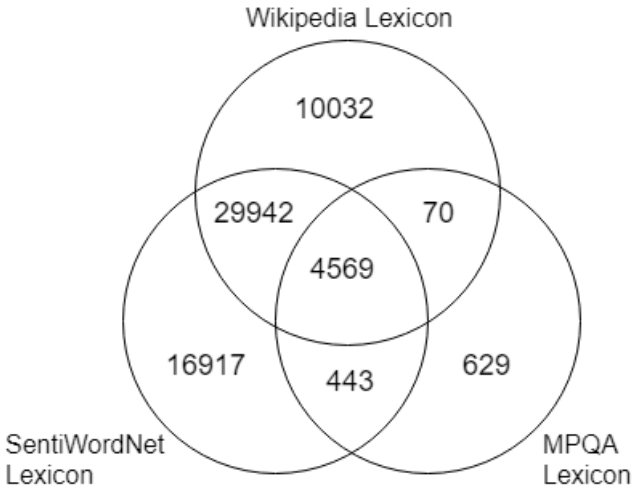


Figure 8: Venn Diagram to illustrate the number of common words among the lexica

### 4.4.1 Composition of Questionnaire

A questionnaire consisting of 100 words is created in Google Forms and distributed via social media. Participants are tasked with selecting the most suitable category for each word, namely; ”Strongly Subjective”, ”Mildly Subjective”, ”Objective” or in the worst case, ”I don’t know”. A total of 13 responses are obtained over a period of 1 week. Unknown to the participants, the 100 words in the questionnaire are actually divided into 2 groups:

- Group 1: Quality control group (20 words). The subjectivity categorisation and scores of these words are already established, but are included in the survey so as to gauge the reliability of the human evaluator in rating assignment. These 20 words are randomly selected from a small pool of words which adhere to the following criteria:

  - Condition 1: Difference between SentiWordScore and Quote Score of the word is less than 0.05
  - Condition 2: If the word is classified as ”strongsubj”, both SentiWordScore and Quote Score of the word are above their respective means. If the word is classified as ”weaksubj”, both SentiWordScore and Quote Score of the word are below their respective means.

- Group 2: Words pending evaluation (80 words). These 80 words are randomly sampled from the 10032 words which have been neither evaluated by MPQA nor SentiWordNet lexicon.
  - ”High” bucket: Top 500 words with the highest Quote Scores
  - ”Low” bucket: Words which have Quote Score below the mean of Quote Score (6236 words)
  - ”Medium” bucket: The remainder words that are not in the other 2 buckets (3296 words)

The overview of the buckets and their respective number of words and mean Quote Scores is summarised in Table 4 below. The mean Quote Score increases accordingly from Low to Medium to High Bucket. As such, it is forecasted that High Bucket would have the highest proportion of words rated as Highly Subjective by human evaluators whereas Low Bucket would have the highest proportion of words rated as Objective.

| Bucket Type | Number of words | Mean Quote Score |
|---|---|---|
| Low | 20 | 0.048509 |
| Medium | 30 | 0.16710 |
| High | 30 | 0.43270 |

Table 4: Number of words in each bucket and the corresponding mean Quote Score

### 4.4.2 Aggregation of Results

Prior to validation and interpretation of results, the responses in the questionnaire are first aggregated. The degree of subjectivity for a particular word may be perceived differently by questionnaire participants and yet a decision must be made to assign the word to one of the three categories; Objective, Mildly Subjective or Strongly Subjective. A **simple majority voting model** is used in which the selection of category by one participant constitutes one vote, and the category with the highest overall number of votes will be assigned to the word. In a case where the majority of votes is for ”I don’t know”, the category with the second highest number of votes will be assigned to the word instead. This is because the ”I don’t know” category does not provide any meaningful information in the context of evaluating

the words. If there would be a word where 2 or more categories receive equal and the highest number of votes, the word will be excluded from evaluation to eliminate ambiguity in the categorisation process. The categorisation process will be demonstrated using an excerpt of words from the questionnaire and the corresponding responses of participants, as shown in Table 5 below.

| Words | folky | lune | toleration |
|---|---|---|---|
| Objective | 1 | 3 | 4 |
| Mildly Subjective | 6 | 0 | 4 |
| Strongly Subjective | 2 | 0 | 2 |
| I don't know | 1 | 7 | 0 |
| Categorization Outcome | "Mildly Subjective" | "Objective" | ambiguous |

Table 5: Responses for sample of words in questionnaire

With reference to Table 5, the word "folky" has 1 vote for "Objective", 6 votes for Mildly Objective, 2 votes for "Strongly Subjective" and 1 vote for "I don't know". Using the simple majority voting rule, the category with the most votes is "Mildly Subjective", hence the word "folky" is classified as "Mildly Subjective" by human evaluators. As for the word "lune", the majority of questionnaire participants voted "I don't know", and the category with second highest number of votes is "Objective". In this case, votes for "I don't know" are not taken into consideration and hence the word "lune" is assigned the category of "Objective". Categorisation of the word "toleration" is rather ambiguous, since "Objective" and "Mildly Subjective" receive equal and majority of votes. Words which do not have a category with distinct majority votes are excluded from the evaluation to eliminate ambiguity.

### 4.4.3 Reliability of Questionnaire Participants

20 words in the questionnaire belong to the "quality control group", which is used as a sanity check for responses to the questionnaire. A minimum "agreement threshold" that must be fulfilled for each individual response with respect to the quality control group is set at 50 percent. In other words, a participant must have selected the "correct" category for at least 10 out of 20 words in the quality control group. The definition of what constitutes a "correct" answer is detailed below:

- If the word is classified as "strongsubj" in MPQA lexicon, a response which classifies the word as "Mildly Subjective" or "Strongly Subjective" is deemed to be correct.
- If the word is classified as "weaksubj" in MPQA lexicon, a response which classifies the word as "Objective" or "Mildly Subjective" is deemed to be correct.

The aforementioned criteria as well as the minimum agreement threshold are defined in such a way that most responses would easily fulfil them. However, responses which deviate significantly from the existing classification of words in quality control group are excluded. Furthermore, questionnaire participants who are unable to categorise many words in the quality control group (multiple "I don't know" responses) would obtain a lower score and not be taken into consideration when tallying results. As such, responses from 3 participants did not meet the 50 percent minimum threshold and are excluded, as a result, only 10 responses are taken into account.

### 4.4.4 Interpreting Results of Questionnaire

After the validation and aggregation process, the main focus is on the actual group of words to be evaluated in which only the Quote Score is known. There were initially 80 words which are to be evaluated, however, after excluding words with ambiguous classification by human evaluators during the aggregation process, there are 70 words remaining. The analysis of the results takes into account only responses from 10 participants, as 3 failed to meet the minimum agreement threshold of the quality control group. As stated in section 4.4.1, words pending evaluation are composed of word sampled from 3 different buckets, as summarised below.

- "High" bucket: Top 500 words with the highest Quote Scores
- "Low" bucket: Words which have Quote Score below the mean of Quote Score
- "Medium" bucket: The remainder words that are not in the other 2 buckets

The main objective is to determine whether there is indeed a significant difference among the 3 buckets with respect to proportion of Objective, Mildly Subjective and Strongly Subjective categorisation by human evaluators. Table 6 below shows the result tally from the questionnaire with subjectivity categories and bucket types as dimensions. Note that the number of words in all buckets has decreased slightly due to removal of words with ambiguous classification.

| Bucket Type | Strongly Subjective | Mildly Subjective | Objective | SUM |
|---|---|---|---|---|
| Low | 1 | 2 | 15 | 18 |
| Medium | 0 | 4 | 21 | 25 |
| High | 5 | 5 | 17 | 27 |

Table 6: Number of words in each subjectivity category segregated by bucket type

As the number of words in the buckets are not equal, the results are also represented in proportions so as to have a fair comparison, as depicted in Table 7 on the following page.

| Bucket Type | Strongly Subjective | Mildly Subjective | Objective |
|---|---|---|---|
| Low | 0.055556 | 0.11111 | 0.83333 |
| Medium | 0 | 0.16000 | 0.84000 |
| High | 0.18519 | 0.18519 | 0,62963 |

Table 7: Proportion of words in each subjectivity category segregated by bucket type

Based on Table 7 above, it can be observed that High Bucket has the highest proportion of Strongly Subjective and Mildly Subjective words while having the lowest proportion of Objective words as compared to the other 2 buckets. This observation is intuitive as the mean Quote Score of words in the High Bucket is larger than that of the other 2 buckets, hence it is expected that there is a higher tendency for questionnaire participants to categorise words that belong to the High Bucket as Strongly Subjective or at least Mildly Subjective. Nevertheless, when comparing Low Bucket and Medium Bucket, only a minor difference in the proportion of words classified into the 3 subjectivity categories is observed. Moreover, a common feature among all 3 buckets is that the majority of words have been classified by human evaluators as Objective. While this observation corroborates mean Quote Score for each bucket type, this also highlights the fact even human evaluators tend to perceive that the majority of words in Wikipedia are Objective.

#### 4.4.5 Areas of Improvement

During the course of data collection, several voluntary feedback are received from questionnaire participants. Furthermore, some areas of improvement have been identified which could mitigate the challenges faced during aggregating of results as well as improve participation rate, as detailed below:

- **Difficulty level of words.** The difficulty level of words is underestimated during the preparation of the questionnaire, as even native English speakers did not know some of the words. Nevertheless, since words which constitute the questionnaire are randomly sampled, the degree of difficulty cannot be controlled directly. A possible solution to this issue is the provide an appendix with the definitions of all words in the questionnaire. However, care must be taken to provide all possible meanings of a word so that the questionnaire participant can make an informed decision.
- **Questionnaire size versus audience size.** The questionnaire consists of 100 words to be evaluated, which is a relatively large task. On the other hand,

the target group for survey participants are those who with high proficiency in the English language, which is a relatively small group. In hindsight, the reverse may have been more effective, meaning that the questionnaire could have been shortened significantly while reaching out to a larger audience in order to gain more samples.

- **Participation Incentives.** The participation rate could have been increased if a certain tangible reward is offered to lucky participants, such as cash vouchers. However, caution must be exercised not to encourage ”junk responses”, such as multiple submissions from the same person or random selection of answers in order to complete the questionnaire as quickly as possible.

# 5 Conclusion and Future Work

The subjectivity lexicon is generated from the English Wikipedia with moderate success. The Quote Score is calculated based on simple ratio of word frequency in quotations and outside quotations. Quotation detection in Wikipedia turned out to be a non-trivial task, as quotation marks are occasionally missing in the text corpus. A conservative approach is taken when dealing with incomplete quotation marks, taking into consideration only quotes with complete pairs of opening and closing quotation marks. This could be improved in future researches where a taxonomy study should be conducted to gain better understanding of the context of quotations to estimate the missing starting or end point of quotes for a particular language. Moreover, further research could be done to explore the possible motivations behind quotations in Wikipedia, apart from only functioning as a means of direct citation.

The Quote Score of words in the subjectivity lexicon is compared against categorical subjectivity classification in MPQA lexicon as well as numerical-based subjectivity score in SentiWordNet lexicon. With the same techniques applied in this Master Thesis, the subjectivity lexicon could be reproduced based on Wikipedia of another language, and the Quote Score could act as a baseline reference point for further subjectivity research for that specific language. Moreover, another possibility is to incorporate aspects of grammar and sentence structure into the lexicon generated from Wikipedia as a form of smoothing to the subjectivity scores of words for a particular language.

A questionnaire has been commissioned to have human evaluators determine the degrees of subjectivity for a small sample of words in the Wikipedia corpus that are not part of the MQPA and SentiWordNet lexica, which are then compared against the corresponding Quote Scores. While these "new" words present an opportunity to extend the vocabulary list of existing subjectivity lexica based on the small sample, a dedicated team of language experts would be required to evaluate all these "new" words in order to verify their suitability to be added to existing subjectivity lexica.

In addition to that, the proposed use cases of this subjectivity lexicon, such as analysis of online product reviews as well as detection of fake news, could be an interesting follow-up theme for this research. The methods used in generating the subjectivity lexicon in this research is intentionally kept simple and not language-dependent so that it could appeal to the commercial world in terms of ease of implementation. It is hoped that the subjectivity lexicon, which is generated from an openly-available resource, will in turn benefit users and organisations in the online world alike.

# 6 References


Anindya Ghose, P. G. I. (2009). *Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics.* Ieee Transactions on Knowledge and Data Engineering, 1–15.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). *SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*, 0, 1–12.

Chenlo, J. M., & Losada, D. E. (2013). *A machine learning approach for subjectivity classification based on positional and discourse features.* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8201 LNCS, 17–28.

Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., & Lee, L. (2009). *How Opinions are Received by Online Communities: A Case Study on Amazon.com Helpfulness Votes.*

Deng, L. (2015). *MPQA 3.0: An Entity/Event-Level Sentiment Corpus.* NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1323–1328. Retrieved from https://people.cs.pitt.edu/ wiebe/pubs/papers/naacl2015.pdf

Esuli, A., & Sebastiani, F. (2006). *SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining.* Proceedings of the 5th Conference on Language Resources and Evaluation, 417–422.

Fellegi, I. (2010). *Survey Methods and Practices.* Statistics Canada.

Jamsen, J., & Corley, K. (2007). *E-Survey Methodology. Handbook of Research on Electronic Surveys and Measurements*, 1–8.

Khanna, S., & Shiwani, S. (2013). *Subjectivity detection and Semantic orientation based Methods for Sentiment Analysis*, 4(9), 868–873.

Lin, C., He, Y., & Everson, R. (2011). *Sentence Subjectivity Detection with Weakly-Supervised Learning*, 1153–1161.

Mudambi, S. M., & Schuff, D. (2010). *What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com.* MIS Quarterly, 34(1), 185–200.

Riloff, E., & Wiebe, J. (2003). *Learning Extraction Patterns for Subjective Expressions.* Proceedings of the 2003 Conference on Empirical Methods in Natural Language Pro-

cessing -, 10, 105–112.

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). *Deception Detection for News: Three Types of Fake News.* Proceedings of the Association for Information Science and Technology, 52(1), 1–4.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). *Fake News Detection on Social Media: A Data Mining Perspective*, (i).

Simm, W., Ferrario, M. A., Piao, S., Whittle, J., & Rayson, P. (2010). *Classification of Short Text Comments by Sentiment and Actionability for VoiceYourView.* Proceedings - SocialCom 2010: 2nd IEEE International Conference on Social Computing, PASSAT 2010: 2nd IEEE International Conference on Privacy, Security, Risk and Trust, 552–557.

Tschiatschek, S., Singla, A., Rodriguez, M. G., Merchant, A., & Krause, A. (2018). *Fake News Detection in Social Networks via Crowd Signals.*

Wiebe, J., Wilson, T., & Cardie, C. (2005). *Annotating Expressions of Opinions and Emotions in Language.* Language Resources and Evaluation, 39(2–3), 165–210.

Wilson, T. A. (2001). *Finegrained Subjectivity and Sentiment Analysis: Recognising the Intensity, Polarity, and Attitudes of Private States.*

Zheng, C., He, G., & Peng, Z. (2015). A Study of Web Information Extraction Technology Based on Beautiful Soup. Journal of Computers, 10(6), 381–387.

Farkiya, A., Saini, P., & Sinha, S. (2015). Natural Language Processing using NLTK and WordNet, 6(6), 5465–5469.

Flöck, F. (n.d.). *f-squared*. Retrieved December 2, 2017, from https://f-squared.org/

*Multi-Perspective Question Answering*. (n.d.). Retrieved December 2, 2017, from http://mpqa.cs.pitt.edu/

*SentiWordNet.* (n.d.). Retrieved December 2, 2017, from http://sentiwordnet.isti.cnr.it/

*Natural Language Toolkit.* (n.d.). Retrieved December 5, 2017, from http://www.nltk.org/

# 7 Appendix

The English Wikipedia dataset used to generate the subjectivity lexicon can be downloaded directly via the following link.
`http://f-squared.org/wiki_html_082016/wiki_html_082016.tar.bz2`

The project source code can be found via the GitHub link below. The SentiWordNet and MPQA lexica are also included.
`https://github.com/waynekoblenz/masterthesiscode`