EVALUATING THE EFFECTIVENESS OF HASHTAGS AS PREDICTORS OF THE SENTIMENT OF TWEETS

A Project Report

Submitted to the Faculty of Graduate Studies and Research

In Partial Fulfillment of the Requirements

for the Degree of

Masters of Science

in

Computer Science

University of Regina

By

Credell Simeon

Regina, Saskatchewan

May 2015

Abstract

Twitter is a microblogging application, which has gartered much interest in recent years. The main source of attraction is its user-generated content called tweets, that are created daily by users. Tweets are 140-character text messages expressing opinions about different topical issues. They are highly informal, and compact with many different conversational features, some of which are specific to Twitter. One such unique feature is hashtags, which are user-defined topics. Marked by a hash symbol "#" at the beginning, hashtags can range from a single word to a combination of multiple words. The uniqueness of this feature stems from its ability to simultaneously connect related tweets, topics, and communities of Twitter users. Moreover, hashtags can also contain sentiment and topic information. Those that are void of sentiment are referred to as *non-sentiment hashtags* whereas those that convey feelings are referred to as *sentiment hashtags*. Therefore, hashtags are interesting features, and thus in this project, we focus exclusively on evaluating their contribution to the sentiment analysis of tweets.

Sentiment analysis of tweets focuses on determining whether the opinion expressed within the text is either positive or negative. However, the uniqueness of tweets poses a challenge to this task because of the time and effort required for manual annotation and feature selection. Therefore, in order to improve this classification task, we hypothesize that hashtags can be used as accurate predictors of the sentiment of tweets. Furthermore, we need to determine whether tweets that contain sentiment information are better predictors than those with only topic information.

In order to prove our hypothesis, we propose to apply a lexicon-based approach, which incorporates the use of subjective words from different lexical resources, to identify sentiment from non-sentiment hashtags. Then, using a supervised machine learning approach, we intend to apply features of these hashtags to classify tweets according to sentiment. Finally, the results of our experiments are presented.

Acknowledgements

The successful completion of this project is due to the concerted efforts and support that I have received from many individuals.

Firstly, I would like to express my sincere gratitude to Dr. Robert Hilderman for his supervision, knowledge, encouragement, support and guidance throughout this project. It was greatly appreciated.

Special thanks to Dr. Howard J. Hamilton for providing me with the opportunity to work on this project.

I would like to acknowledge Dr. Orland Hoeber for introducing me to the topic of sentiment analysis, and Dr. Lisa Fan on providing greater insight into the topic of classification, which has lead to the development of a significant component of my project.

I would also like to thank Dr. Malek Mouhoub and Dr. Daryl Hepting for participating on my committee, the faculty and staff of the Department of Computer Science for their support, and the Faculty of Graduate Studies and Research for their financial support.

Finally, I want to extend special thanks to my immediate family for their love and support throughout this project.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Since its inception in 2006, Twitter has gained increasing popularity with approximately 241 million monthly users in 2014 [12]. Twitter is a microblogging application, which allows users to communicate with each other by exchanging short text messages called tweets. Tweets can be shared with persons selected by the user to be his/her followers. Therefore, each user is linked to a group of followers, resulting in an interconnection of followers, and the formation of a social network.

Figure 1.1 shows an example of an account of a registered user on Twitter [2]. The main highlight of this account is the user's public timeline, in which the most recent tweets are displayed at the top of the list. These tweets have been posted by persons to whom the user has selected to follow. These persons are referred to as "following". Therefore, Twitter is designed to allow users to quickly discover popular issues that are being discussed among their peers, and by extent, the wider community.

Generally, tweets are opinionated statements, which are posted by users in order to enlist the support of their followers. The content of tweets can vary from personal thoughts to public statements [41]. There is an estimated 500 million tweets that are created by users each day [25]. In recent years, there has been growing interest by the research community in collecting tweets, and analyzing the sentiment contained in them.

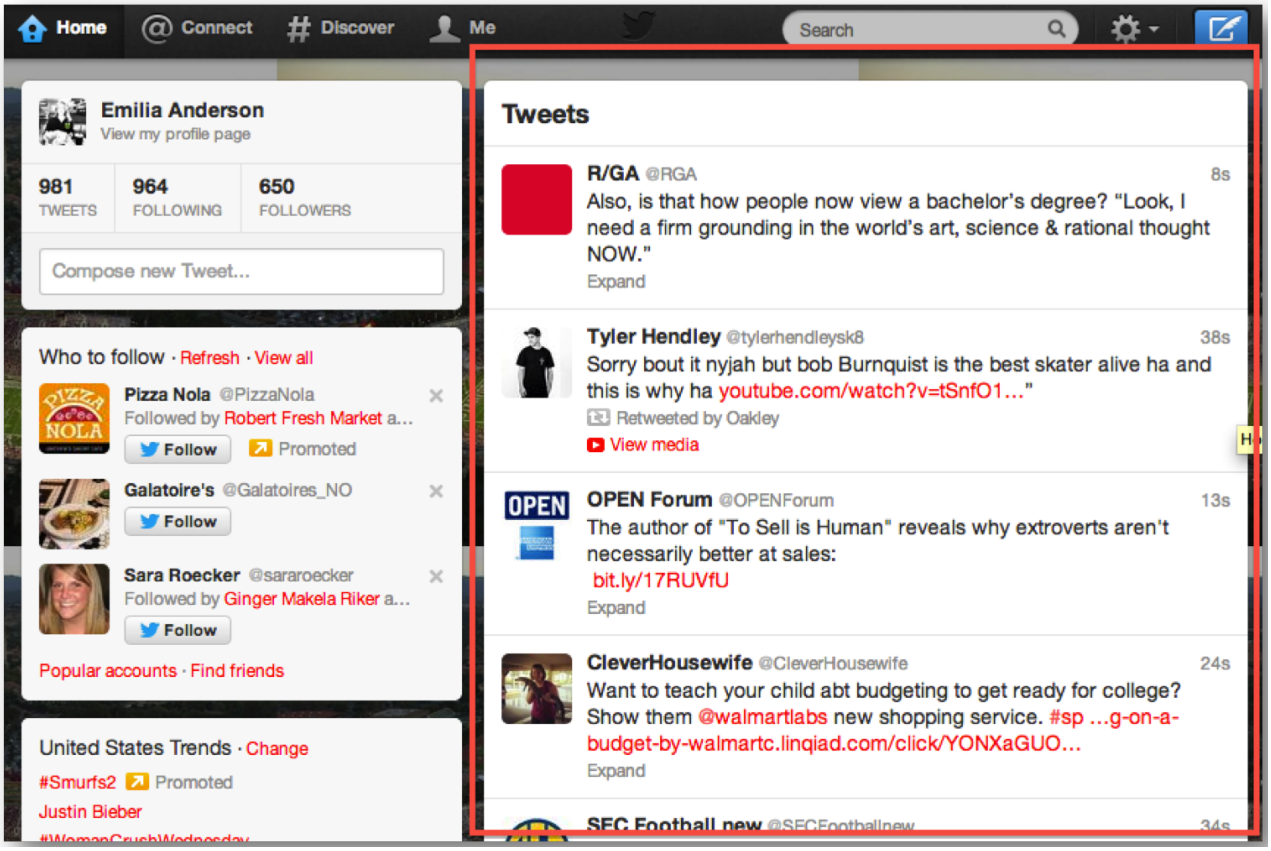

Figure 1.1: Overview of the Twitter application

Here, sentiment refers to a personal feeling that can be expressed either positively or negatively. Although sentiment analysis of text has been studied extensively, tweets are unlike any other text. They are generally structured as a single sentence, compact with many different linguistic and conversational features. As a result, the sentiment analysis of tweets can be a challenging task [8]. Therefore, it is first important to examine the unique nature of tweets, as this plays a significant role in determining their overall sentiment.

Figure 1.2: Example of a tweet with Twitter-specific features

Tweets are highly informal text messages, which are restricted to 140 characters. Figure 1.2 shows an example of a tweet highlighting Twitter-specific features. These are described below.

1. **Retweets** are copies of an original tweet that are posted by other users [19]. They are denoted by the letters, "RT".

2. **Mentions** are used for replying directly to others. They begin with the "@" symbol followed by the name of a Twitter user e.g. "@john".

3. **URL links** are used to direct users to interesting pictures, videos or websites for additional information.

4. **Hashtags** are user-defined topics, keywords or categories denoted by the hash symbol, "#". Hashtags can be a single word or a combination of words connected without whitespaces, e.g. "#believe" and "#wishfulthinking", respectively. A tweet can contain multiple hashtags, which can be located anywhere in the text.

They are also a number of conversational features that are not Twitter-specific, but they are commonly found in tweets. These are described below.

1. **Abbreviations, slangs and acronyms** are useful because of the 140-character limitation of tweets. Figure 1.2 shows examples of these features:

   (a) Slangs e.g."cuz" which means"because".

   (b) Acronyms e.g. "OMG" refers to "Oh My God".

   (c) Abbreviations e.g."dem" meaning"them" ,and "2" meaning "to".

2. **Repeated characters, exclamation or question marks** are used for emphasis [4, 10]. Examples include "soooo" instead of "so", "!!!!" and "??", respectively.

3. **Emoticons** are used for conveying an emotional state [8]. They are facial expressions that are formed by combining standard keyboard characters e.g. ":D" is a positive emoticon.

Of all the features of tweets described previously, hashtags have been selected as the focus of this project. The significance of hashtags lies in their unique ability to simultaneously connect related tweets, topics, and communities of people who share similar interests. Each hashtag is a sharable link, which can be used to promote specific ideas, search for and share popular content, engage other users, and group related content. Moreover, some hashtags have gained popularity by starting new online conventions, such as "#tbt", which means "throwback thursday", where users would post an old photo, video or thought [37]. Most importantly, hashtags are useful for determining the popularity of topics, relationship among topics, and the overall sentiment that is being

expressed by groups of users. Because of their usefulness, hashtags are also used by many other platforms including photo-sharing applications such as Instagram and Flickr, and social networks like Facebook, Tumblr and Google+.

Additionally, hashtags contain sentiment and topic information. Hashtags that contain only topic information are considered to be non-sentiment bearing. However, hashtags that contain sentiment information, such as an emotion expressed by itself or directed towards an entity, are considered to be sentiment bearing. These two types of hashtags are similar to the sentiment, sentiment-topic and topic hashtags proposed by in a previous study [51]. Table 1.1 shows examples of sentiment and non-sentiment bearing hashtags. It can be seen from Table 1.1 that non-sentiment bearing hashtags typically refer to a

Table 1.1: Examples of sentiment and non-sentiment bearing hashtags

Type of hashtag	**Examples**
Sentiment	#excited, #bestplayers, #iloveu, #imGrateful2u4urLove, #gooooalll! #brasilmustwin
Non-sentiment	#canada, #football, #Justinbieber, #FOLLOW, #itsweekend, #watchinDgame

specific entity such as a place, person, object, event or an activity. On the other hand, sentiment bearing hashtags refer to specific feelings or desires that are conveyed to an audience.

In this study, we hypothesize that hashtags can be used as accurate predictors of the overall sentiment of tweets. Based on this assumption, we can identify three major opportunities for improving the sentiment analysis of tweets. Firstly, we might be able to accurately determine the sentiment of a large volume of tweets without having to examine individual tweets. Secondly, we can reduce dependency on manual annotation of tweets, which can be time-consuming and labor-intensive [8, 14, 53]. Thirdly, by focusing on a single feature, we can reduce the effort required in determining the optimal combination of the various features in the tweets. Therefore, in this project, we apply the features of hashtags, including their sentiment polarity, in order to classify tweets as positive,

negative or and neutral. Additionally, we compare the effectiveness of hashtags that contain sentiment information with those that contain only topic information, for this classification task.

The remainder of the chapter is organised as follows. In Section 1.1, a statement of the problem is outlined. In Section 1.2, an overview of sentiment analysis and classification is provided. In Section 1.3, the objectives of the project are outlined. In Section 1.4, the contributions of the project are outlined. In Section 1.5, the organization of this report is provided.

1.1 Statement of Problem

One of the most significant hindrances to the task of sentiment analysis of tweets is the time and effort required for manual annotation of tweets [8, 10, 50, 53]. Other challenges that are associated with sentiment analysis of tweets are related to the unique nature of tweets. Due to the 140-character restriction, tweets are compact with many different features, and can contain highly informal language, such as mispellings, grammatical errors, slangs and abbreviations [8]. In order to overcome these linguistic challenges, many previous studies have selected a combination of features for determining sentiment in tweets [10, 19, 24, 41]. However, this project centers on using a single feature, that of the sentiment contained in hashtags, in order to automatically classify tweets into three classes: positive, negative and neutral.

Table 1.2: Examples of tweets with their respective sentiment labels

Example Tweet	**Sentiment Label**
“ Everybody loves #football and the #worldcup! I #love the #players!! #WorldCupHotties #lovingit”	Positive
“ #hot #sports #news Netherlands Coach Hates The World Cup Third-Place Game”	Negative
“ #News: England out of World Cup 2014 latest reaction!: England are officially out after losing to Cost... http://t.co/SowdL0LLGj #TU”	Neutral

Table 1.2 shows examples of positive, negative and neutral tweets. A tweet is considered to be positive, if it expresses a desirable state. By contrast, a tweet is considered to be negative, if it expresses an undesirable feeling. Neutral tweets are void of any sentiment. It can be observed from Table 1.2 that tweets containing hashtags can have additional sentiment information. Therefore, there exists value in focusing on this feature, in order to increase the accuracy of the sentiment classification of a large volume of tweets. Thus, the main goal of the project is to demonstrate the effectiveness of relying on the features of the hashtag(s) contained in the tweet, in order to assign the tweet to a specific sentiment class. In order to achieve this goal, we first extract the hashtags from a dataset of tweets. Then, we identify sentiment hashtags from non-sentiment bearing hashtags. Afterwards, we group the tweets into two separate datasets based on this classification. Following this, we determine the sentiment polarity of each hashtag i.e. positive, negative or neutral. Finally, we apply the features of the hashtags, including their sentiment polarity, in order to determine the overall sentiment polarity of each tweet.

1.2 Overview of Sentiment Analysis

Sentiment Analysis (SA) or Opinion Mining (OM) studies people's attitudes, emotions and opinions towards an entity [32]. Sentiment Analysis is considered a classification task whereby text can be classified as expressing positive or negative opinion. It can be applied in business intelligence applications and recommender systems for analyzing customer feedback [43] on products and services, in order to improve corporate decisions [10]. It can also be useful to political and social organizations for collecting and monitoring feedback from their supporters. By analyzing the sentiment contained in tweets, entertainers can assess how fans feel about their work [10, 31]. For the research community, the main interest lies in developing methods to determine the sentiment contained in text.

There are two main approaches which can be applied to classify sentiment in text:

lexicon-based approach and supervised machine learning. Here, we provide a brief overview of each approach with a more detailed description in Chapter 2. Firstly, the lexicon-based approach depends entirely on using opinion lexicons. Opinion lexicons are dictionaries of positive and negative words used to identify, and determine the sentiment orientation of text [53]. By contrast, supervised machine learning uses a large number of training data where the sentiment labels are known. Then, it relies on learning algorithms to determine the relationship between the sentiment labels and the linguistic features of the text [48].

For this project, we intend to apply a lexicon-based approach to identify sentiment from non-sentiment hashtags. The assumption is that by using this approach, which depends entirely on opinion lexicons to ascertain subjectivity, we would be able to accurately distinguish sentiment-bearing hashtags from non-sentiment hashtags. In order to test this hypothesis, we develop and evaluate a classification model using words from different opinion lexicons as input.

Additionally, we apply supervised machine learning methods to determine the sentiment polarity of the hashtags classified as sentiment or non-sentiment bearing. Then, we ascertain whether this sentiment polarity can be an accurate predictor of the sentiment of the entire tweet. Here, we assume that the sentiment assigned to the hashtag is reflective of the sentiment of the entire tweet. We also compare the performance of sentiment and non-sentiment hashtags in the sentiment classification of tweets. To test this hypothesis, we develop and evaluate two separate classification models using training and testing datasets of tweets, for each type of hashtag.

1.3 Objective of the Project

The main objective of this project is to determine whether hashtags can be accurate predictors of the sentiment of tweets. In order to achieve this goal, we intend to undertake the following tasks:

1. To analyze a dataset of tweets containing hashtags, which is collected from Twitter.

2. To develop and evaluate a lexicon-based approach to automatically identify sentiment from non-sentiment bearing hashtags extracted from a dataset of tweets.

3. To develop classification models to determine the sentiment polarity of sentiment and non-sentiment bearing hashtags.

4. To apply this sentiment polarity, and other features of hashtags to determine the overall sentiment of tweets.

5. To evaluate our approach to the sentiment analysis of tweets.

1.4 Contributions of the Project

The methods developed in this project make the following contributions to the field of Sentiment Analysis:

1. It demonstrates the effectiveness of combining subjective words from different opinion lexicons to identify sentiment bearing and non-sentiment bearing hashtags.

2. It demonstrates the effectiveness of using the features of hashtags, for the sentiment analysis of tweets.

3. It shows that non-sentiment hashtags are more effective at classifying tweets as positive, negative or neutral, than sentiment hashtags.

4. It shows that non-sentiment hashtags are more effective at classifying subjective tweets, than sentiment hashtags.

1.5 Organization of the Report

The remainder of this report is organized as follows. Chapter 2 presents background information on the sentiment analysis of tweets, Chapter 3 summarizes the previous work conducted on the sentiment analysis of tweets, including their contributions to this field. Chapter 4 describes the development of the our approach including data collection, hashtag extraction, hashtag classification and sentiment classification of tweets. Chapter 5 discusses the experimental results obtained for each classification model, and compares these results with that of another well-known study. Finally, Chapter 6 summarizes the main components and achievements of our approach, and outlines our plans for future work to conclude this report.

Chapter 2

Background

Sentiment Analysis (SA) focuses on determining whether the sentiment contained within a piece of text is positive or negative [10]. There are three different types of sentiment analysis: document-level, sentence-level and aspect-level [32]. Document-level sentiment analysis determines the overall sentiment in an entire document. This type of sentiment analysis has been performed on reviews [16, 30, 43, 52], and news articles [38]. By contrast, sentence-level sentiment analysis identifies each sentence in a document as a separate unit, which contains its own sentiment [23]. Sentences which contain either positive or negative opinion are subjective, whereas those which are void of sentiment are objective. Aspect-level sentiment analysis focuses on classification of sentiment based on the aspects of entities in the text [32].

The sentiment analysis of tweets is considered to be a sentence-level task. Three main reasons contribute to this conclusion. Firstly, tweets are typically structured as a single sentence because of the 140-character restriction. Secondly, tweets typically express opinions about a single entity [11]. Thirdly, each tweet is an independent unit with its own unique sentiment-bearing features. In terms of classification, tweets can be categorized into two classes (positive or negative) [4, 8, 10, 44] or three classes (positive, negative or neutral) [1, 5, 19, 24, 27, 41, 53].

The remainder of this chapter is organized as follows. In Section 2.1, we provide a general overview of the sentiment analysis of tweets. In Section 2.2, we describe the supervised machine learning approach to sentiment analysis of tweets, including pre-processing and feature selection. In Section 2.3, we describe the lexicon-based approach, including different lexical resources.

2.1 Overview of Sentiment Analysis of Tweets

Sentiment analysis uses data mining technologies to detect subjectivity in text. Figure 2.1 shows an overview of the methods used for the sentiment analysis of tweets.

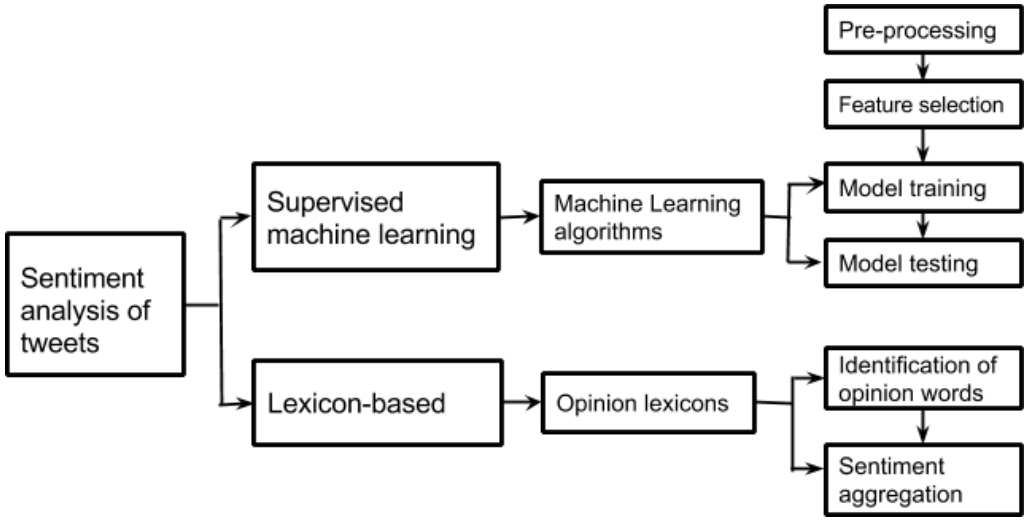

Figure 2.1: Overview of the sentiment analysis of tweets

It can be observed from Figure 2.1 that there are two main approaches: supervised machine learning and lexicon-based methods. Supervised machine learning infers a

learning function from labelled training data. By contrast, the lexicon-based method relies on opinion lexicons to detect subjectivity in the text. Another notable difference between the two approaches is that machine learning requires separate training and testing phases. During training, learning algorithms are applied to features extracted from a set of tweets where the sentiment labels are known, in order to build a classification model. In the testing phase, this model is applied to predict the sentiment of a set of unseen tweets. By contrast, the lexicon-based method does not require a training phase. Another difference is that the lexicon-based method relies on a single feature in the tweet, that of the presence of subjective words. On the other hand, machine learning uses a combination of different features in the tweet, in order to ascertain sentiment. These features are selected after the original data has been pre-processed.

2.2 Supervised Machine Learning

Supervised machine learning relies on a large volume of labelled data [32]. Labelled or training data refers to data, in which the sentiment labels (positive, negative, neutral) are known. Sentiment labels can be obtained manually or automatically. Manual labelling requires the use of human annotators to assign sentiment labels. Due to the large volume of tweets, manual annotation can be time-consuming and labor-intensive [8, 14, 50, 53]. In order to overcome this challenge, noisy labels can be used to automatically assign sentiment labels to tweets.

A noisy label is an indicator in the tweet that can be used to predict the sentiment class to which the tweet belongs [10]. Therefore, the sentiment expressed by the noisy label is considered to be reflective of the sentiment of the entire tweet. Using emoticons or smileys, a tweet is considered to be positive, if it contains a positive emoticon, ":)". Likewise, a tweet is considered to be negative, if it contains a negative emoticon, ":(" [10, 19, 24, 41]. Additionally, sentiment suggestive words can be applied to automatically label tweets,

including sentiment hashtags [14]. For example, if a tweet contains the word "love", it is considered to be positive, whereas the word "awful", implies that the tweet is negative [10]. Tweets that contain ambiguous indicators are removed from the dataset [19].

In the supervised machine learning approach, labelled data is required for training the classifiers. The classifier is a learning algorithm used to determine the association between the features in the tweet and the assigned sentiment labels [48]. Then, the classifier can be applied to predict the sentiment of unlabelled tweets. The unlabelled tweets are referred to as the test dataset. Finally, the predictive performance of the classifier can be evaluated by comparing the predicted sentiment of the tweet to its actual sentiment.

2.2.1 Pre-processing

In the supervised machine learning approach, pre-processing tasks are applied to the original data, which has been extracted from Twitter. This process is important in order to normalize the data, so that the most relevant features can be selected. Common pre-processing tasks include stemming, spelling correction, stop word removal, tokenization and POS tagging.

Stemming reduces inflected words to their base or root form. By applying stemming, the words "stemmer", "stemming", and "stemmed" are reduced to their root "stem" [6].

Spelling correction corrects misspelled words. One way in which words can be misspelt is by the presence of extra repeated letters. Therefore, a spelling correction algorithm can be applied to handle these repeated characters by replacing them with one occurrence, and two occurrences, respectively [4, 6, 10]. For example, the word "happppppy" is replaced with "hapy" and "happy", respectively.

Stop word removal eliminates common words that do not carry any sentiment from the dataset. Examples include definite articles, such as "the" or "an", pronouns and prepositions. Stop word removal has been shown to improve the performance of the classifier [4].

Tokenization is the process of deconstructing tweets into n-grams (tokens). Therefore, it converts the dataset into a bag-of-words, which can include individual words called unigrams, two word combinations (bigrams) or a sequence of three words (trigrams) [34].

Part-of-Speech (POS) tagging is defined as a basic form of syntactic parsing [18]. Words are annotated with the appropriate POS tags based on the context in which they appear in the sentence. An example of applying POS tagger from the Natural Language Processing Toolkit (NLTK) on a sentence is shown below [9].

Sentence: They refuse to permit us to obtain the refuse permit.

Tagged sentence:

[('They', **'PRP'**), ('refuse', **'VBP'**), ('to', **'TO'**), ('permit', **'VB'**), ('us', **'PRP'**),
('to', **'TO'**), ('obtain', **'VB'**), ('the', **'DT'**), ('refuse', **'NN'**), ('permit', **'NN'**)]

Figure 2.2: An example of POS tagging

Figure 2.2 shows both the original sentence, and the resulting tagged sentence. The tags used are "NN" for noun, "PRP" for pronoun, "DT" for definite article, "VB" for verb, "VBP" for present tense verb, and "TO" for "preposition".

2.2.2 Feature Selection

Feature selection focuses on extracting relevant features from tweets. By selecting the most relevant features, the predictive accuracy of the classifier can be improved [34]. Additionally, the running time for the learning algorithms can be reduced [46]. N-grams, especially unigrams and bigrams, are the most common features used in studies on sentiment analysis [5, 6, 19, 24, 41, 43]. They can be represented by binary weights (0 and 1) to denote their presence or absence in the tweet. Also frequency weights can be used to represent the relative importance of the feature in the dataset [32]. Similarly, features

can be represented by their probability distribution in the different sentiment classes [4]. Other useful features for the sentiment analysis of tweets include part-of-speech (POS) tags [1, 41, 44], Twitter-specific attributes [4, 5, 26, 42], and punctuation [4, 14, 28].

2.3 Lexicon-based Approach

This approach relies exclusively on opinion lexicons to identify, and determine the sentiment orientation of tweets i.e. positive, negative or neutral [53]. Opinion lexicons are dictionaries of opinion terms, and their associated semantic orientation. The semantic orientation or polarity refers to the direction in which the word deviates from the other words in its semantic group. Therefore, words expressing a favourable state have a positive orientation, while words expressing an unfavourable state have negative orientation [22]. Semantic orientations are represented as numerical scores, such as "+1" for positive orientation, and "-1" for negative orientation [13].

After the subjective words in the tweet have been identified, their semantic orientations are aggregated to estimate the overall sentiment polarity of the tweet [42]. If the aggregated scores exceeds a certain threshold, the tweet is considered to be positive, otherwise it is classified as negative. Another method is to simply count the number of positive and negative words identified in each tweet. If the count of positive words exceeds the count of negative words, then the tweet is classified as positive. Exceptions can be made for the presence of negating words in the tweet. Negating words such as "not" or "never", can change the semantic orientation of a word. Therefore, the semantic orientation of a positive word can be changed to negative, if a negation is within a certain distance of the word [5, 42].

Although the lexicon-based approach has been applied, there are a number of limitations due to the characteristics of tweets [53]. The main limitation is its inability to handle context dependent words, colloquial expressions, emoticons, and abbreviations found in

tweets [53]. Although these expressions can be subjective, their semantic orientations are not present in opinion lexicons. Consequently, the sentiment of tweets cannot be accurately predicted. Therefore, the lexicon-based approach is unable to adapt to the evolving nature of the language used on Twitter [31]. However, this limitation may be overcome by applying lexicons, which have been specifically designed for Twitter.

Opinion lexicons can be annotated manually or automatically. In a manual process, human subjects are used to determine the sentiment orientation of words, whereas in an automatic process, this orientation is derived from machine learning processes. Consequently, manually created lexicons are smaller in size than those that are automatically created.

2.3.1 Manually Annotated Opinion Lexicons

Examples of manually created lexical resources are described as follows:

1. **SentiStrength** contains over 2500 words extracted from short, social web text. It assigns a single numerical score ranging from 1(no positivity) to 5 (extremely positive) for positivity, and -1(no negativity) to -5 (extremely negative) for negativity [48].

2. **AFINN** is based on Affective Norms for English Words (ANEW) lexicon. It contains 2477 English words, and uses a similar scoring range as SentiStrength. Moreover, AFINN is specifically created for detecting sentiment in microblogs [40].

3. **General Inquirer** is one of the largest manually created lexicons with over 11,000 words. Words are grouped into different sentiment (positive and negative), and mood categories [47].

4. **Bing Liu Lexicon** contains positive and negative words which have been manually extracted from opinion sentences in customer reviews. It contains misspellings, slangs and other social media expressions. Overall, the lexicon contains about 6,800 words, and assigns polarities of "+1" for positivity, and "-1" for negativity [22].

5. **Subjectivity Lexicon** contains about 8,221 words categorized as strong or weak. For each subjectivity clue, a prior polarity (non-numerical score) is assigned, which can be positive, negative or neutral. The lexicon also contains an assigned POS tag for each word [52].

2.3.2 Automatically Annotated Opinion Lexicons

Examples of automatically created lexicons are described as follows:

1. **SentiWordNet 3.0** is an extension of the WordNet Lexical database [33]. It is the largest lexicon with over 115,000 synsets, which have been automatically annotated, using a semi-supervised learning process. A synset is a group of synonymous words with numerical scores for positivity, negativity and objectivity, which sums to a total of one [3].

2. **Sentiment140 Base Lexicon** is formed from unigrams extracted from 1.6 million tweets. These tweets were collected for a study on the sentiment analysis of tweets [19]. Tweets are labelled positive (negative), if they contain a positive (negative) emoticon. There are a total of 62,468 unigrams, 677,698 bigrams and 480,010 non-contiguous pairs in the lexicon [36].

3. **National Research Council (NRC) Hashtag Sentiment Lexicon** is word-sentiment association lexicon. It is created using 78 hashtagged seed words with positive and negative sentiments, and a set of about 775,000 tweets. A tweet is assigned a positive sentiment label, if it contains any of the positive hashtagged seed words, and a negative label, if it contains any of the negative hashtagged seed words. It consists of 54,129 unigrams, 316,531 bigrams, and 308,808 contiguous pairs. This lexicon contains a numerical sentiment score for each word, which is either positive or negative [36].

All of the above lexicons are publicly available resources.

Chapter 3

Related Work

In recent years, various research papers have been published on the sentiment analysis of tweets. For this study, we choose to consider only related work that makes use of the following: supervised machine learning, lexicon-based methods, n-grams, POS tagging, emoticons, hashtags and the microblogging domain. We did not consider any research work which used document-level sentiment classification, feature selection algorithms, and unsupervised machine learning.

In this chapter, we present recent studies on the sentiment of tweets that are relevant to our study. These studies have applied either supervised machine learning or lexicon-based method. The rest of the chapter is organized as follows. Section 3.1 presents recent studies on supervised machine learning. Section 3.2 presents recent studies that have applied lexicon-based methods.

3.1 Recent Studies on Supervised Machine Learning

One of the earliest studies on the sentiment analysis of tweets is performed by Go et al. [19]. Their study evaluates the performance of three machine learning algorithms, Naive Bayes, Support Vector Machine (SVM), and Maximum Entropy on unigrams and bigrams, which are extracted from a dataset of 1.6 million tweets. These tweets have been automatically

labelled using positive and negative emoticons. A tweet is considered to be positive (negative), if it contains a positive (negative) emoticon. This method called distant supervision is proposed as a viable alternative to manual annotation of tweets. The study achieves accuracy scores of 83%, 82.7% and 81.6% using Maximum Entropy, Naive Bayes, and SVM, respectively. Overall, the study demonstrates that emoticons are effective for automatically assigning sentiment labels to tweets, and that machine learning algorithms can be applied successfully to determine the sentiment of tweets.

Davidov et al. [14] employ 50 Twitter hashtags and 15 smileys as sentiment labels for tweets. The study applies supervised machine learning techniques to classify tweets according to different sentiments. Punctuation, words, n-grams and word patterns are extracted from tweets and used with a K-Nearest Neighbor (KNN) classifier. In the multi-classification task, the classifier assigns one of the specific sentiment labels (50 hashtags and 15 smileys) to each tweet in the test set, whereas in the binary classification task, the classifier assigns either a sentiment label or a non-sentiment label. Experimental results obtained from both tasks indicate that smileys are more effective than hashtags as sentiment labels and that word, punctuation and pattern features are useful features. Overall, the study reveals that it may be easier to determine if a tweet expresses a general sentiment than it is to determine its specific sentiment.

Pak and Paroubek [41] also apply a supervised machine learning framework in order to classify 300,000 tweets as positive, negative, or neutral. They adopt the distantly supervised approach used by Go et al. [19], in which positive and negative emoticons are used to automatically label tweets. Neutral tweets are added to the dataset by extracting tweets from the Twitter accounts of popular newspapers. For the classification task, n-grams are used as binary features with POS distribution information to train two multinomial Naive Bayes classifiers, a N-gram based model and a POS-based model. The study also introduces two strategies, entropy and salience, to eliminate common n-grams. Results of a corpus analysis performed on the dataset reveal that some POS

tags (personal pronouns PP, adjectives JJS, and adverbs RB) are strong indicators of subjective text. Experimental results from the classification task indicate that bigrams are the best features for increasing the accuracy of the classifier. In general, the study suggests that n-grams and POS features are effective for the sentiment analysis of tweets.

Kouloumpis et al. [24] investigate Twitter hashtags for identifying positive, negative and neutral tweets. Hashtags are used as sentiment labels such that the polarity of the tweet is determined by the hashtag. A hashtagged dataset is created and used for training an AdaBoost.MH classifier with 500 rounds of boosting. Additionally, the hashtagged dataset is extended by using randomly selected positive and negative tweets labelled with emoticons from the training set used by Go et al. [19]. The classifier is applied to different linguistic features in the dataset including: the top 1000 n-grams, lexical features from Multi-Perspective Question and Answering (MPQA) subjectivity lexicon, POS tags and microblogging features (emoticons, abbreviations and intensifiers). Using this combination of features, the classifier achieved accuracy scores of 74% and 75% on hashtagged, and emoticon datasets, respectively. Furthermore, microblogging features are found to be the most useful in increasing the accuracy of the classifier. Overall, the experimental results obtained from the study demonstrate that the best features for the sentiment classification of tweets are a combination of n-grams, lexical and microblogging features.

Wang et al. [51] propose a graph-based model for hashtag-level sentiment classification. The authors suggest three different types of hashtags: sentiment hashtags which contain only sentiment information, sentiment-topic hashtags which contain both topic and sentiment information, and topic hashtags contain only topic information. In order to classify 2181 hashtags according to sentiment (positive or negative) in a specific time period, the study applies co-occurrence hashtag information, the literal meaning of hashtags, and the sentiment polarity of the tweets containing the hashtags. Three inference algorithms, Loopy Belief Propagation, Relaxation Labeling, and Iterative Classification are used for

the classification task. Experimental results show that the highest accuracy of 77.2% is obtained with Loopy Belief Propagation with enhanced boosting. Overall, the results obtained from the study show that the sentiment polarity distribution of tweets can be combined with the features of hashtags for sentiment classification.

Mohammad [35] showed that self-labelled hashtagged emotional words in tweets corresponded well with annotations provided by trained human judges. His study is significant because it revealed that sentiment conveyed by the hashtagged emotion word is reflective of the sentiment of the entire tweet. In a later study, Mohammad et al. [36] developed a hashtag sentiment lexicon with unigrams and bigrams using a dataset of about 775,000 tweets and 78 hashtagged seed words. A tweet was assigned the same sentiment polarity if it contained any of the positive (negative) hashtagged seed words. By applying the use of the hashtag lexicon for the sentiment classification of tweets, the performance of the classifier increased by 3.8%.

Bakliwal et al. [4] employ a sentiment scoring function to classify tweets in two separate datasets, Stanford [19] and Mejaj [10], which consists of tweets labelled with emoticons, and sentiment suggestive words, respectively. The scoring function applies a Naive Bayes approach to determine the positive and negative probabilities of the unigrams extracted from the datasets. The positive (negative) probability is determined by dividing the positive (negative) frequency of the unigram by its total frequency. Additionally, the scoring algorithm assigns 0.1, -0.1, 1.0 and -1.0 to each occurrence of an exclamation mark, question mark, positive emoticon, and negative emoticon, respectively. The algorithm also applies stop word removal, stemming using Porter's stemming algorithm, noun identification using WordNet [33], and spelling correction to words with extra repeated characters. Finally, a new feature called Popularity Score is introduced to boost the sentiment scores of commonly used words. A popularity factor is determined by taking the difference of the positive and negative frequency, and assigns values of 0.9, 0.8, 0.7,0.5 or 0.1 if the difference is greater than 1000, 500, 250, 100 and less than 50, respectively.

The overall score of each unigram is found by multiplying the difference between the positive and negative probabilities by the popularity factor. Finally, the scores of all the unigrams and features in the tweet are totaled. If the score is greater than 0, the tweet is determined to be positive, otherwise it is determined to be negative. Experimental results using the scoring algorithm reveal that the Stanford and Mejaj datasets achieve accuracy scores of 87.2% and 88.1%, respectively. Also, the scoring function achieved accuracy scores that are comparable to that obtained using a SVM classifier. The study also applies a feature vector approach to ascertain the best features for the sentiment analysis of tweets. Experimental results obtained from this approach reveal that unigrams and bigrams make the most significant contribution to the sentiment analysis of tweets. Overall, the study demonstrates that scoring algorithms can be used successfully for the sentiment analysis of tweets.

Bravo-Marquez et al. [11] combines features from various lexical resources and sentiment analysis methods in a supervised machine learning approach, in order to boost the sentiment classification of tweets. The classification task is subdivided into subjectivity classification where tweets are classified as subjective or objective, and polarity classification where subjective tweets are determined to be either positive or negative. The features extracted from tweets are the strength, emotion and polarity indicators, which refer to the number of positive and negative words identified by different lexicons, the number of emotion words, and the weighted sum of the semantic orientations of the identified words, respectively. The lexical resources used are AFINN [40], SentiWordNet [3], SentiStrength [48], NRC Hashtag Sentiment [36] and Sentiment140 [36]. Feature analysis is performed on two separate datasets and the results show that polarity and strength indicators are the most useful for subjectivity and polarity classification. For both classification tasks, experimental results indicate that combining strength, polarity and emotion indicators is effective in increasing the performance of the classifiers. Overall, the study shows that combining different lexical resources is effective for sentiment

classification.

In general, studies have applied different combinations of features extracted from tweets to machine learning algorithms. Furthermore, the results obtained from these studies show that supervised machine learning is effective for the sentiment analysis of tweets.

3.2 Recent Studies on Lexicon-based methods

Levallois [26] develops a lexicon-based sentiment detection engine for the sentiment analysis of tweets. This engine consists of four main components: detection of semantic features such as emoticons and onomatopes, evaluation of hashtags, tokenization of tweets into n-grams and comparison of n-grams to terms in lexicons, and use of heuristics to assign a final sentiment to the tweet. The lexicons used in the engine have been manually created from tweets, resulting in 1066 terms representing positivity, negativity, strength of sentiment, and negation. The sentiment engine uses a series of heuristics to determine the sentiment of tweets. In a SemEval-2013 task, the engine is applied to classify 3813 tweets as positive, negative and neutral. The experimental results show that the higher f-scores of 69.81% and 64.95% are achieved for the neutral and positive classes, respectively. The study reports a f-score of 48.96% for the negative class and attributes this low value to the rules used to define a negative sentiment.

Likewise, Palanisamy et al. [42] also apply a similar lexicon-based sentiment engine to classify tweets. The sentiment of a tweet (positive, negative or neutral) is found by taking the aggregated sum of all the sentiment-bearing features in the tweet. However, experimental results obtained in this study indicate that precision and recall scores for both the positive and negative classes are above 70%.

Overall, both studies provide evidence that the lexicon-based approach can also be applied to the sentiment analysis of tweets.

Chapter 4

Methodology

In this study, we use both supervised machine learning and lexicon-based approaches, to investigate the effectiveness of hashtags as accurate predictors of the overall sentiment of tweets. Therefore, we divide the project into two main phases. In the first phase, we develop a modified lexicon-based approach to automatically classify hashtags as sentiment or non-sentiment bearing. In the second phase, we apply supervised machine learning to determine the overall sentiment of tweets containing these hashtags.

This chapter is organized as follows. Section 4.1 outlines our assumptions. Section 4.2 provides a general overview of our approach. Section 4.3 details our approach to the classification of hashtags. Section 4.4 describes the sentiment classification of tweets.

4.1 Assumptions

The following assumptions were made as part of our methodology:

1. Each tweet in our dataset contains at least one hashtag.

2. A tweet can contain both sentiment and non-sentiment hashtags.

3. A hashtag is regarded as sentiment-bearing, if it contains an opinion or subjective word.

4. A hashtag is regarded as sentiment-bearing, if it originates from an opinion or subjective word.

5. A hashtag is considered non-sentiment bearing, if it does not contain or originate from an opinion word(s).

6. If a hashtag is not found in any of the lexical resources used in our study, it is regarded as non-sentiment.

7. Each of our aggregated word lists is mutually exclusive (excluding the list of stems).

8. The sentiment label assigned by the automatic classifiers (not developed in this study) is taken as the actual sentiment of the tweet.

4.2 General Overview of our Approach

Figure 4.1 shows that our approach consists of two main phases: Phase 1 and Phase 2. The first phase (Phase 1) focuses on classifying hashtags as either sentiment or non-sentiment bearing. The second phase (Phase 2) focuses on classifying tweets containing the hashtags identified in the first phase, into one of three sentiment classes: positive, negative or neutral.

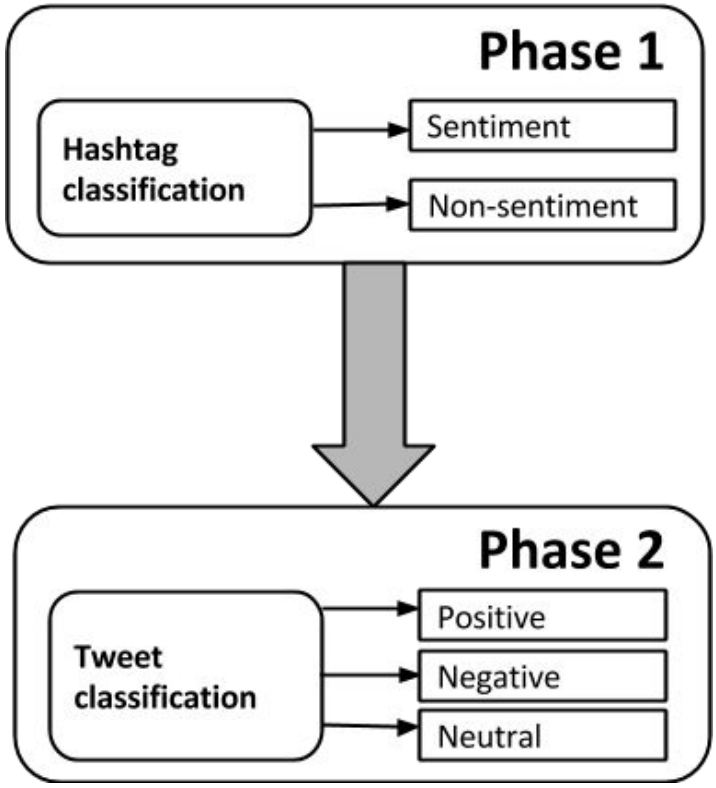

Figure 4.1: General overview of our approach

4.3 Phase 1: Classification of Hashtags

In order to automatically classify hashtags as sentiment or non-sentiment bearing, we develop a classification model using a combination of opinion lexicons and word lists.

4.3.1 Overview of our Approach

For this binary classification task, we develop a lexicon-based approach with some modifications. In our approach, we utilize training and test datasets.

Figure 4.2 shows an overview of our approach. Initially, tweets are downloaded from

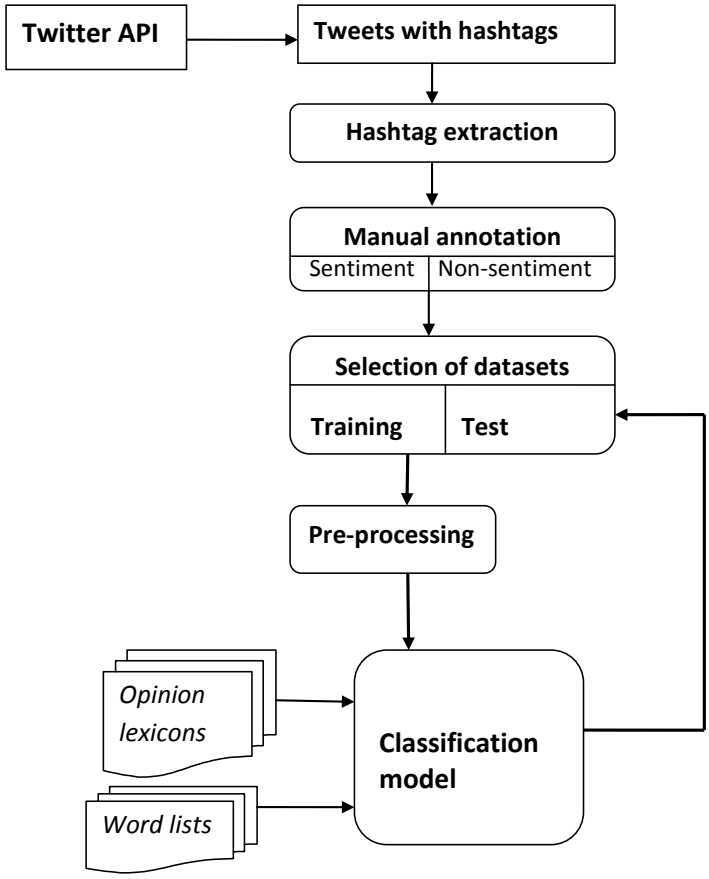

Figure 4.2: Phase 1 - classification of hashtags

Twitter. Then, hashtags are extracted and manually annotated as sentiment and non-sentiment. For each type of hashtag, training and test datasets of tweets are created. Pre-processing tasks are applied to the training sets. Hashtags are stripped of their hash symbol ("#"), and their stems are found and stored in a separate list. Then, a classification model is developed, in which the opinion words from different lexical resources are used to detect subjectivity in the hashtags, in each training dataset. Finally, the model is applied to the test sets to predict the class of the hashtag.

4.3.2 Sentiment Resources

Sentiment resources refer to both opinion lexicons and word lists of sentiment terms. There are a total of 10 different resources that are utilized in the development of our classification model.

4.3.2.1 Opinion Lexicons

In this study, we use seven opinion lexicons: AFINN [40], General Inquirer [47], Subjectivity Lexicon [52], SentiWordNet [3], SentiStrength [48], NRC Hashtag Sentiment Lexicon [36], and Bing Liu Lexicon [22].

Table 4.1: Opinion lexicons

Name of Lexicon	**Total number of words**
AFINN	2477
SentiStrength	2,538
Bing Liu	6,789
Subjectivity Lexicon	8,221
General Inquirer	11,763
NRC Hashtag Sentiment Lexicon	54,129
SentiWordNet	144,510

Table 4.1 lists the opinion lexicons, and number of unique words that are contained in each of them. It can be observed from Table 4.1 that SentiWordNet has the largest number of words, followed by NRC Hashtag Sentiment Lexicon, General Inquirer, and Subjectivity Lexicon, respectively. The smaller lexicons are AFINN, SentiStrength and Bing Liu.

Each lexicon is downloaded separately from the official website that is responsible for its dissemination. Although the structure of each lexicon is different, we concentrate on extracting all positive and negative words. However, there are a few lexicons, which are handled differently in order to extract only the words that are strongly subjective. For the SentiStrength Lexicon [48], we extract positive and negative words with semantic orientations greater than 2.0, and less than -2.0, respectively. For NRC Hashtag Sentiment

Lexicon [36], we extract the top 500 adjectives for each sentiment class (positive and negative). A Part-of-Speech (POS) tagger from the NLTK (Natural Language Processing Toolkit) is used for identifying adjectives. For SentiWordNet [3], we consider only the adjectives (POS tags provided in the lexicon) that have scores for positivity or negativity, which are greater than or equal to 0.5.

4.3.2.2 Word Lists

We also use three lists of sentiment words: Steven Hein feeling words [21] which contains words collected from 1995 to 2013, The Compass DeRose Guide to Emotion Words [15] which consists of words categorized according to different moods, and sentiment bearing Twitter slangs and acronyms collected from various online sources [7, 17]. Most of these words are not found in the other lexicons. Examples include "fab" for "fabulous", and "HAND" for "Have a Nice Day".

Table 4.2: Word lists

Name of list	**Total number of words**
Steven Hein's feeling words	4232
Compass DeRose Guide	682
Twitter slangs and acronyms	222

Table 4.2 shows the total number of words contained in each word list. It can be observed in Table 4.2 that the list of Twitter slangs and acronyms contains the smallest number of words, whereas the list of Steven Hein's feeling words contains the largest number of words.

4.3.2.3 Aggregation of Sentiment Resources

Using these 10 sentiment resources, a total of five aggregated lists of words are created after a series of experiments is performed on the training set to determine the combinations. These resources are described below.

1. **FOW** (Frequently Occurring Words) list contains the most subjective words. These 915 words have occurred in at least five resources. The threshold of five represents half of the total number of resources under consideration.

2. **Stems of FOW** contains the stems of all the opinion words in the FOW list. There are 893 words in this list.

3. **MDW** (More Discriminating Words) list contains all opinion words that are strongly subjective. These 7366 words occur in the smaller opinion lexicons and word lists: AFINN, SentiStrength, Bing Liu and Compass DeRose Guide as well as those which occur in 4 out of the 5 larger lexicons and word lists: NRC Hashtag Sentiment, SentiWordNet, General Inquirer, Subjectivity Lexicon and Steven Hein's feeling words.

4. **LDW** (Less Discriminating Words) list consists of subjective words that occur in at least 2 but not exceeding 3 of the 5 larger lexicons and word lists. These 868 words are considered to be the least subjective.

5. **Twitter slangs and acronyms** which have been manually identified. This list also includes common interjections [39] giving a total of 308 words.

4.3.3 Development of Classification Model

The classification model is developed using the training dataset for each type of hashtag, and the five aggregated lists of words. The model uses a binary search algorithm, in order to determine whether the hashtags in the training datasets meet one of the following criteria:

1. It is an opinion word or originates from an opinion word.

2. It contains an opinion word or feature.

Based on this criteria, the model is divided into two steps. Initially, each of the aggregated word lists are sorted alphabetically. In the first step, we use the binary search algorithm to compare each hashtag with each word in the different word lists. Comparisons are also made between the stem of the hashtag and each opinion word. If a match is found, the search terminates. Otherwise, the search must continue into the second step.

In the second step, the hashtags that have not been matched after the first step are handled. Our aim is to ascertain if the hashtag contains an opinion word (including a word originating from an opinion word) or feature. In order to do this, two recursive algorithms are employed to create substrings of the hashtag. The resulting substrings are compared to the opinion words in the FOW, stems of FOW, and MDW lists because the substrings are smaller representations of the hashtag, and thus, only matches to the most subjective words are considered. Additionally, we only consider substrings of the hashtag that contain 3 or more characters.

In order to create substrings of the hashtag, we utilize two recursive algorithms. The first algorithm called *reduce_hashtag*, employs a recursive process, in which the rightmost character is removed from the hashtag after each iteration. The remaining characters from the hashtag form the left substring, whereas the removed character(s) form the right substring. The final result is a list of substrings of the hashtag sorted in descending order of length. Table 4.3 shows an example of applying the this algorithm to the pre-processed hashtag, "lovestory".

Table 4.3: Example of the reduce_hashtag algorithm

Iteration no.	**left substring**	**right substring**
1	**lovestor**	y
2	**lovesto**	ry
3	**lovest**	**ory**
4	**loves**	**tory**
5	**love**	**story**
6	**lov**	**estory**

It can be observed from Table 4.3 that there are a total of 10 relevant substrings. They include all of the left substrings, and the last four right substrings.

The second algorithm called *remove_left*, utilizes a recursive process, in which the leftmost character is removed from the hashtag after each iteration. Table 4.4 shows an example of applying this algorithm to the pre-processed hashtag, "behappy".

Table 4.4: Example of the remove_left algorithm

Iteration no.	**substring**
1	ehappy
2	happy
3	appy
4	ppy

It can be observed from Table 4.4 that a total of four relevant substrings are found by the algorithm.

Initially, the recursive algorithms are applied to produce a list of substrings. Starting with the largest substring, each one is compared to each opinion word in the FOW, stems of FOW and MDW lists, until a match is found. Comparisons are also made between the stem of the substring and that of each opinion word, and the stem of the substring and each opinion word.

Finally, when all previous attempts to determine if the hashtag contains an opinion word fails, we then ascertain if the hashtag contains an opinion feature. In this study, an opinion feature is any non-word attribute in the hashtag that suggests the expression of a sentiment. Therefore, we consider only extra repeated letters, exclamation or question marks. By using extra repeated characters, the author of a tweet can emphasize certain words, and then use these words to convey a personal feeling. For example the word, "#goallllllll", expresses a feeling of excitement. A previous study [4] suggests that the presence of exclamation marks implies positivity, whereas the presence of question marks implies negativity. Therefore, we evaluate the hashtag for the presence of exclamation or question marks or repeated characters (at least 3).

Table 4.5 outlines the detailed rules for determining whether a hashtag is sentiment bearing.

Table 4.5: Rules for identifying sentiment hashtags

No.	Rules
1	Hashtag = opinion word
2	Hashtag = stem of an opinion word
3	Stem of the hashtag = an opinion word
4	Stem of the hashtag = stem of a FOW
5	Max(substring of the hashtag) = an opinion word
6	Stem of the max(substring of the hashtag) = stem of a FOW
7	Max(substring of the hashtag) = stem of an opinion word
8	Hashtag contains a sentiment feature

It can be observed that from Table 4.5 there are a total of eight rules. If **any** of these rules is found to be true, then the hashtag is determined to be sentiment bearing. Otherwise, the hashtag is non-sentiment bearing.

Finally, the output of the classification model is the predicted class for each hashtag in the training sets. This is the one of the features of the hashtag that would be used in the sentiment classification of tweets.

4.4 Phase 2: Classification of Tweets

The main objective of the second phase is to compare the effectiveness of sentiment and non-sentiment hashtags in classifying tweets into one of three sentiment categories: positive, negative, or neutral. For this classification task, we develop different scoring algorithms that can be used in conjunction with various classifiers.

4.4.1 Overview of our Approach

Figure 4.3 shows an overview of our approach. Initially, we use the sentiment and non-sentiment hashtags that are classified during the hashtag classification task to select tweets, one dataset for each type of hashtag. Then, an automatic classifier is applied to

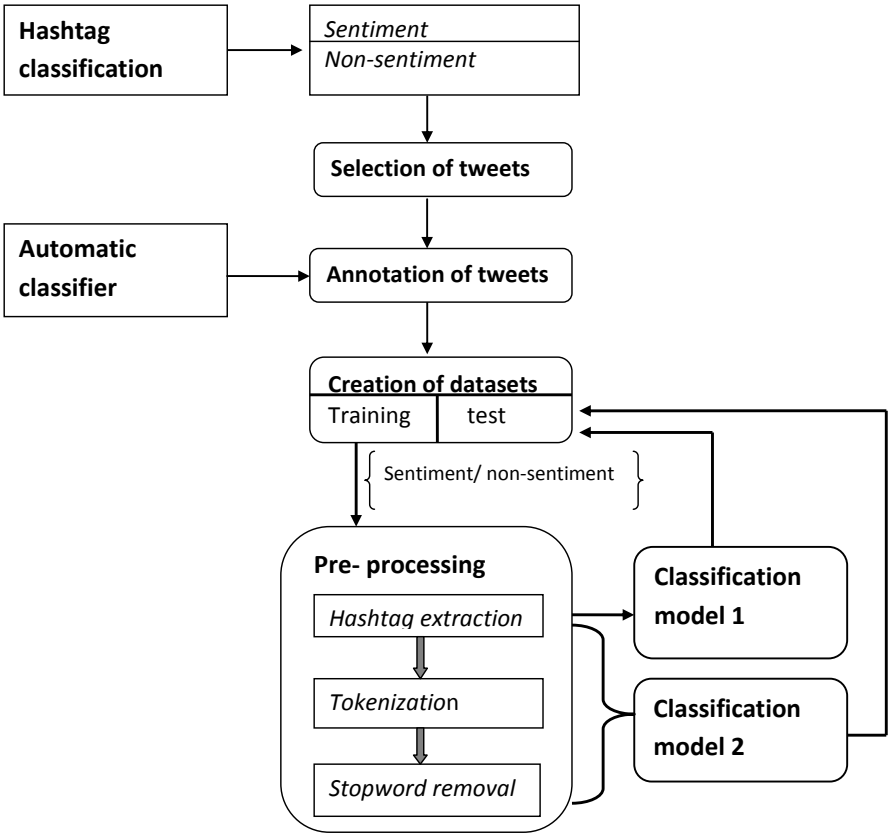

Figure 4.3: Phase 2 - sentiment classification of tweets

the selected tweets, in order to automatically assign sentiment labels (positive, negative or neutral). Afterwards, each dataset is divided into training and test sets. Pre-processing tasks are applied to the training sets. Two different classification models are developed using a supervised machine learning approach. Finally, each model is applied to the test sets, in order to predict the sentiment of unseen tweets.

4.4.2 Development of Model 1

In the first model, the total number of occurrences of each unique hashtag, is determined for each sentiment class. Each unique hashtag is assigned to the sentiment class, for which it has the highest frequency in the training set. This is the simplest model.

Suppose the hashtag, "#love", occurs in a total of 300 tweets. Each tweet has a sentiment label (positive, negative, neutral). Table 4.6 shows the frequency of this hashtag in each sentiment class.

Table 4.6: Example showing frequency of hashtag in each sentiment class

Sentiment class	**Frequency**
Positive	200
Negative	83
Neutral	17

It can be observed from Table 4.6, that the positive sentiment class obtained the highest frequency. Therefore, the hashtag is determined to be positive.

4.4.3 Development of Model 2

This model uses a bag-of-words approach. Tweets in the training set are tokenized into unigrams. Usernames and URL links are replaced with generic tags [19]. Hashtags are extracted, and stored separately. Emoticons are identified, and replaced with tags to indicate their sentiment polarity. Similarly, negating words, repeated questions and exclamation marks are also extracted, and substituted with special tags. All other

punctuation marks as well as stop words are removed from the dataset. Then, each unique word in the tweet is used as a feature.

4.4.3.1 Algorithm development

The frequency of each word in the different sentiment classes is calculated. Then the positive, negative and neutral ratios are found using Equations 4.1, 4.2 and 4.3.

The positive ratio shown in Equation 4.1 is defined as the difference between the number of positive tweets and the number of non-positive tweets that the word occurs in, divided by the number of positive tweets that contains the word.

$$positive\,ratio(word) = \frac{pos(word) - (neg(word) + neu(word))}{pos(word)} \tag{4.1}$$

The negative ratio shown in Equation 4.2 refers to the difference between the number of negative tweets and the number of non-negative tweets that the word occurs in, divided by the number of negative tweets that contains the word.

$$negative\,ratio(word) = \frac{neg(word) - (pos(word) + neu(word))}{neg(word)} \tag{4.2}$$

The neutral ratio shown in Equation 4.3 is defined as the difference between the number of neutral tweets and the number of non-neutral tweets that the word occurs in, divided by the number of neutral tweets that contains the word.

$$neutral\,ratio(word) = \frac{neu(word) - (pos(word) + neg(word))}{neu(word)} \tag{4.3}$$

The sentiment ratio of the word, $sr(word)$, is the maximum of the positive, negative and neutral ratios. The sentiment weight of each word, $word_{sw}$, is determined from

Equation 4.4 as

$$word_{sw} = sr(word) \times word_f \tag{4.4}$$

where $sr(word)$ is the sentiment ratio of the word, and $word_f$ is the frequency of the word in the dataset.

If the maximum ratio of the word is the positive ratio, then the sentiment polarity of the word, $spol(word)$, is positive. If the maximum ratio of the word is the negative ratio, then the sentiment polarity of the word, $spol(word)$, is negative. Otherwise, the sentiment polarity of the word, $spol(word)$, is neutral. We use the values 1, -1 and 0 to represent positive, negative and neutral polarities, respectively.

Then, we determine the total sentiment weight, TSW, of all the words in the tweet by taking the sum of the sentiment weight of all the words as shown in Equation 4.5

$$TSW = \sum_{i=1}^{nw} word_{sw}. \tag{4.5}$$

where nw, is the number of words.

Next, we find the weighted average of each word, $WeightedAvg_{word}$, by dividing the sentiment weight of the word by the total sentiment weight of all the words in the tweet multiplied by the sentiment polarity of the word.

$$WeightedAvg_{word} = \frac{word_{sw}}{TSW} \times spol(word) \tag{4.6}$$

Additionally, emoticons, punctuation marks, and negating words are also incorporated as features into the model. Positive emoticons and exclamation marks are assigned a sentiment polarity of 1, whereas negative emoticons, question marks and negations are assigned a sentiment polarity of -1. In order to find the weighted average of the features in each tweet, $WeightedAvg_{feature}$, its frequency in the tweet, $count(feature)$, is divided

by its total frequency in the dataset, $frequency_{feature}$, multiplied by its given sentiment polarity, $spol(feature)$. This weighted average of the feature is described in Equation 4.7.

$$WeightedAvg_{feature} = \frac{count(feature)}{frequency_{feature}} \times spol(feature) \tag{4.7}$$

Thus, each hashtag in the tweet is scored by taking the sum of the weighted average of each word and the sum of the weighted average of each feature (emoticons, negation, exclamation or question mark) as described in Equation 4.8 as

$$overall\_score = \sum_{i=1}^{nw} WeightedAvg_{word} + \sum_{i=1}^{n} WeightedAvg_{feature} \tag{4.8}$$

where nw is the number of words, and n is the number of features.

The sentiment score for each hashtag in the tweet ranges from -1 to 1. A score of -1 indicates that each feature or word in the tweet has a negative sentiment, and thus, the hashtag(s) are assigned a negative sentiment. A score of 1 indicates that each feature or word in the tweet has a positive sentiment, and thus, the hashtag(s) are assigned a positive sentiment. A score of 0 indicates that each feature or word in the tweet has a neutral sentiment, and thus, the hashtag(s) are assigned a neutral sentiment. For other scores, in order to assign a discrete sentiment of positive, negative, or neutral to the hashtag(s) in each tweet, a threshold is used whereby hashtags with scores above 0.1 are considered positive, those with scores below -0.1 are considered negative, and those with scores between -0.1 and 0.1 are considered neutral.

In order to determine the overall sentiment of the hashtag in the training set, we count the number of times the hashtag is scored as positive, negative, or neutral, and assign the sentiment class with the highest frequency.

At the end of the training phase, we have two additional features for each hashtag: its frequency in the training set, and its sentiment polarity.

4.4.3.2 Example for Model 2

Suppose we take the tweet, "My head hangs in shame #FreeThe7 !! (;/". After tokenization and stopword removal, the relevant words that are used by the model are listed in Table 4.7. For each selected word, its frequency in positive, negative, and neutral tweets are calculated. Next the positive, negative, and neutral ratios are calculated using Equations 4.1, 4.2, and 4.3, respectively. Table 4.8 shows these values.

Table 4.7: Example showing frequency of relevant words in each sentiment class

Word	**Positive**	**Negative**	**Neutral**
head	4	1	5
hangs	15	8	0
shame	0	20	1

Table 4.8: Example of determining the positive, negative and neutral ratios

Word	**Positive Ratio**	**Negative Ratio**	**Neutral Ratio**
head	-0.25	-8	0
hangs	0.467	-0.875	0
shame	0	0.95	-19

The next step would be to identify which of the three ratios is the highest for each word. Table 4.9 shows the sentiment ratio for each word, which is the maximum ratio for each word. Thus, if the neutral ratio is the highest then the word is neutral. After

Table 4.9: Example of selecting of the sentiment ratio

Word	**Sentiment ratio**	**Ratio type**
head	0	neutral
hangs	0.467	positive
shame	0.95	negative

the sentiment ratio for each word is determined, the sentiment weight of each word is calculated using Equation 4.4. Table 4.10 shows the sentiment weight of each word.

The total sentiment weight of all the words in found by using Equation 4.5. The total sentiment weight is 30.691.

Table 4.10: Example showing the sentiment weight of each word

Word	Sentiment weight
head	0
hangs	10.741
shame	19.95

Then, the weighted average of each word is found by using Equation 4.6. Table 4.11 shows the weighted average of each word.

Table 4.11: Example showing the weighted average for each word

Word	Weighted average
head	0
hangs	0.35
shame	-0.65

Additionally, each emoticon, exclamation and question mark in the tweet is also incorporated as features in the model. They are added as weighted features using Equation 4.7.

Table 4.12 shows the total occurrence in the tweet, the total number of occurrences, and their calculated weighted average.

Table 4.12: Example showing the weighted average of features

Feature	Frequency in tweet	Total frequency	Sentiment polarity	Weighted average
exclamation mark	2	10	1	0.2
negative emoticon	1	10	-1	-0.1

Finally, the score of the hashtag, "#Freethe7", is found using Equation 4.8. The calculated score is -0.2, and thus, the hashtag is negative for this specific tweet.

In order to assign a single overall sentiment for this hashtag, the count of the number times the hashtag is scored as positive, negative and neutral is calculated. Suppose the hashtag, "#Freethe7", is scored as positive, negative or neutral in a total of 100, 183, and 17 tweets, respectively. Then, we select the sentiment category which has the largest frequency. Therefore, the overall sentiment of the hashtag is negative.

At the end of the training phase, we have two features of the hashtag,"#Freethe7": its total frequency in the training set which is 300, and its overall sentiment polarity which is negative.

Chapter 5

Experimental Results

A number of experiments are performed to evaluate our classification models on the sentiment analysis of tweets. In this chapter, we present our experimental results.

This chapter is organized as follows. Section 5.1 outlines the hardware and software used. Section 5.2 provides an overview of our dataset. Section 5.3 outlines the interest measures used to evaluate our models. Section 5.4 discusses the results obtained for the hashtag classification task. Section 5.5 presents the results obtained for the sentiment classification of tweets, and compares these results with that of another study.

5.1 Hardware and Software

The experiments were executed on a computer running Windows 8.1 operating system with a 1.70 GHz AMD A8 Quad-Core processor, 8GB of RAM and 684GB hard drive. The software used include Python 2.7 with modules Natural Language Toolkit (NLTK) [29], Sentiment140 Application Programming Interface (API) [45], Umigon online sentiment classifier [26], and Waikato Environment for Knowledge Analysis (WEKA) data mining application (version 3.7) [20]. For a few of our experiments, we use the WEKA implementations of machine learning algorithms Naive Bayes, Maximum Entropy, SVM, and C4.5 on our training and test sets.

We also use the built-in functions in NTLK for POS tagging, tokenization, and stemming. We use a Regrex stemmer, which is available in this module, in both classification tasks. Using this stemmer, we are able to specify the suffices to be removed. The suffixes selected are “ed”, “s”, “er”, “tion”, “ness”, “ing”, and “ment”.

Sentiment140 API uses a Maximum Entropy classifier with an accuracy of 83 percent on a combination of unigrams and bigrams [19, 45]. Positive, negative, and neutral tweets are assigned numerical values of 4, 2, and 0, respectively.

Umigon online sentiment classifier uses a lexicon-based method to assign sentiment labels (positive, negative or neutral) to tweets [26]. This classifier does not have an API so tweets have to be pasted onto the website. Consequently, this limits the number of tweets that can be classified at any one time.

5.2 Dataset

Our dataset consists of tweets extracted using the Twitter API [49]. Twitter provides this API for the public to collect tweets. We collected tweets from June 11^{th}, 2014 to July 2^{nd}, 2014 during the FIFA World Cup 2014. Tweets were scraped from the API using search terms related to the football matches that were being played, in order to capture the opinions of the Twitter users during the game. The search terms used were not hashtags because the intention was to collect a wide variety of hashtags from the dataset of tweets. Of the 1,857,277 tweets that were collected, 635,553 tweets contained at least one hashtag. By eliminating all the re-tweets, our dataset consists of 71,836 tweets with hashtags. We use Sentiment140 API to automatically assign sentiment labels to the tweets in our dataset.

5.3 Interest Measures

The effectiveness of our classification models is evaluated using the standard measures in text classification: accuracy, precision, recall and f-measure defined in equations 5.1, 5.2, 5.3,and 5.4 where TP, FP, FN, and TN represents the number of true positive, false positives, false negatives, and true negatives, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{5.1}$$

Accuracy measures the number of tweets (hashtags) for each sentiment class that are classified correctly. High accuracy indicates that there are significantly more true positives and true negatives than false positives and false negatives. Therefore, the classifier is able to correctly determine whether or not tweets (hashtags) belong to a particular class. By contrast, low accuracy indicates that there are significantly fewer true positives and true negatives than false positives and false negatives.

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

Precision determines the ratio of actual relevant tweets (hashtags) among predicted tweets (hashtags) for the sentiment category. A high precision indicates that there are significantly fewer false positives than true positives. Therefore, the classifier is very exact in its prediction. By contrast, low precision indicates that there are significantly more false positives than true positives, which indicates that when the classifier makes a prediction, it is often incorrect.

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

Recall refers to the fraction of relevant tweets (hashtags) actually classified by the model. High recall indicates that there are significantly fewer false negatives than true negatives. Therefore, the classifier is able to identify a significant number of tweets (hashtags) as belonging to a particular class. On the other hand, low recall means that there are significantly more false negatives than true negatives, which further indicates that the classifier is unable to identify a significant number of relevant tweets (hashtags) that belong to a particular class.

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5.4}$$

F-measure is defined as an average of precision and recall. High f-measure means that values obtained for precision and recall are very close, whereas low f-measure indicates that values for precision and recall are further apart.

5.4 Hashtag Classification

Initially, a total of 9,866 unique hashtags are extracted from the dataset and manually classified, resulting in 2,750 sentiment-bearing hashtags, and 7,116 non-sentiment bearing hashtags. Table 5.1 shows the total number of tweets in the training and test sets. There are 8,225 tweets with sentiment bearing hashtags of which 4,112 and 4113 are in each set, respectively. There are significantly more tweets with non-sentiment hashtags, 24,028 and 24,029 in each set, respectively.

Table 5.1: Distribution of tweets in the training and test sets

Hashtag type	**Training**	**Test**	**Total**
Sentiment	4,112	4,113	8,225
Non-sentiment	24,028	24,029	48,057

5.4.1 Analysis of Sentiment Resources

For this classification task, we combined subjective words from various sentiment resources (opinion lexicons and word lists). Initially, we analyze the different resources to determine the percentage of words that are shared between any two resources. Figure 5.1 shows the table of our results from these comparisons. We compare the words in the opinion lexicons and word lists on the left column of the table with those listed at the top. The highlighted percentage represent the largest percentage of words shared between a specific lexical resource and another resource. For example, AFINN contains 79.18% of the words in Compass DeRose word list.

Sentiment resources	*AFINN*	*Bing Liu*	*SentiStrength*	*Subjectivity Lexicon*	*NRC Hashtag Sentiment*	*General Inquirer*	*SentiWordNet*	*Steven Hein*	*Compass DeRose*
AFINN		19.41	27.54	20.18	**3.36**	12.37	1.22	22.42	**79.18**
Bing Liu	53.23		63.04	**77.92**	6.12	32.08	3.73	40.43	56.74
SentiStrength	**28.25**	23.58		20.3	2.94	21.57	1.48	15.86	25.81
Subjectivity Lexicon	34.92	49.16	34.23		4.3	23.21	2.56	39.11	**57.33**
NRC Hashtag	**73.61**	48.82	62.72	54.47		72.38	11.79	52.89	68.04
General Inquirer	43.13	40.79	**73.33**	46.79	11.54		5.63	34.5	45.31
SentiWordNet	72.23	80.77	85.5	87.74	31.99	**95.82**		77.72	86.95
Steven Hein	38.36	25.21	26.44	**38.66**	4.13	16.92	2.24		**89.44**
Compass DeRose	**21.83**	5.7	6.93	9.13	0.85	3.58	0.4	14.41	

Figure 5.1: Comparisons of words in the different sentiment resources

It can be observed from Figure 5.1 that each sentiment resource contains words that are not found in other resources. Therefore, by combining the resources, we maximise the number of subjective words that are used in our model.

5.4.2 Experimental Setup

In order to evaluate our classification model, we compare the hashtags extracted in the test sets with those obtained in the training set. If the hashtag is found in the list,

the same class label is assigned. Otherwise, similarity testing is performed between the hashtags in the test set and those in the training set.

In similarity testing, we compare the stems of the hashtags in the training and test sets. If a match cannot be determined, then a comparison is applied to ascertain if any of the substrings of the hashtag in the test set is at least 95% of the length of one of the determined hashtags in the training set. If a suitable match is found, the hashtag is assigned the same class label as the hashtag in the training set. Otherwise, the hashtag is predicted to be non-sentiment bearing. Finally, we compare the predicted class label assigned by the model to that of actual label of the hashtag assigned during manual annotation.

5.4.3 Results and Discussion

Table 5.2 shows the accuracy, precision, recall, and f-measure scores (in percent) obtained by our classification model.

Table 5.2: Classification of hashtags

Hashtag type	**Accuracy**	**Precision**	**Recall**	**F-measure**
Sentiment	83.58	86.27	80.96	83.53
Non-sentiment	**83.83**	**94.25**	**84.93**	**89.35**

It can be observed from Table 5.2 that our approach achieved over 80% accuracy, precision, recall and f-measure in classifying both types of hashtags. However, our approach achieved higher percentages for all four measures in identifying non-sentiment hashtags. Therefore, our results suggest that it is easier to identify non-sentiment hashtags than sentiment hashtags.

In order to compare the performance of our approach, we created models which substituted the combined resources for a single resource. Each model is given the name of the sentiment resource (opinion lexicon or word list) used. Figure 5.2 shows the accuracy scores for our approach and these models.

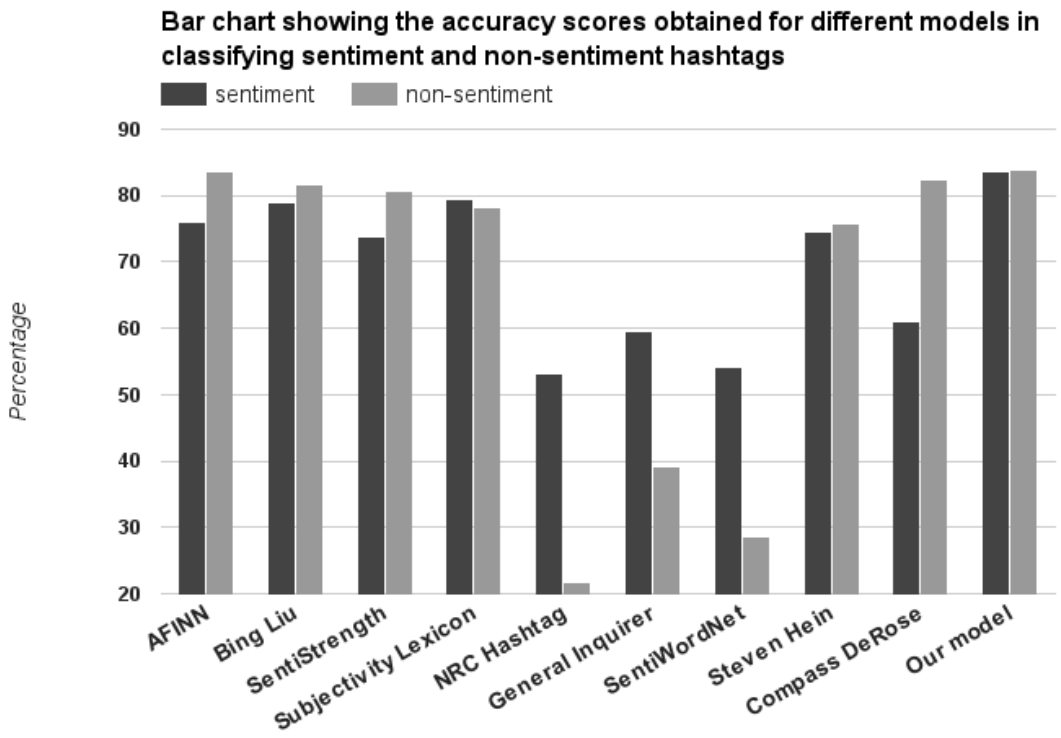

Figure 5.2: Comparison of the accuracy obtained by different models

It can be observed in Fig 5.2 that our model (last column) which used combined sentiment resources is the most accurate in identifying sentiment hashtags, when compared to models which used a single resource. For the identification of non-sentiment hashtags, our approach achieves accuracy scores comparable to each of the models that used a single resource. Therefore, combining subjective words from different sentiment resources increases the accuracy of our approach in identifying sentiment hashtags without compromising its accuracy in identifying non-sentiment hashtags.

It can be observed from Fig 5.3 that our approach (last column), which used the combined resources, attained the highest f-measure in identifying sentiment hashtags when compared to models which used a single resource. For the identification of non-sentiment hashtags, our approach performs comparably to each of the models that used a single resource. Therefore, combining subjective words from different sentiment resources increases the performance of our model in identifying sentiment hashtags without

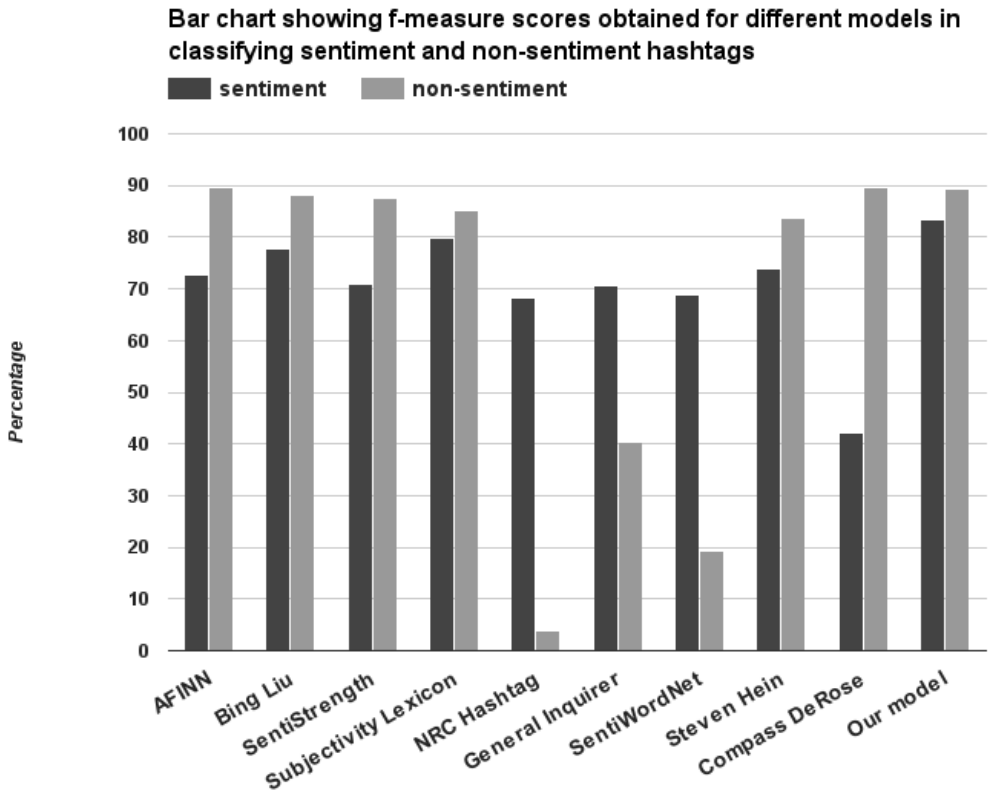

Figure 5.3: Comparing the performance of our model to models using a single resource

compromising the performance of our model in identifying non-sentiment hashtags.

Overall, our approach obtained higher accuracy and f-scores in identifying non-sentiment bearing hashtags than sentiment bearing hashtags. Furthermore, our experimental results show that combining subjective words from different opinion lexicons and word lists is effective in identifying sentiment bearing hashtags from non-sentiment bearing hashtags.

5.5 Sentiment Classification of Tweets

We use the sentiment and non-sentiment hashtags that are classified by our model to select tweets that contain these hashtags. Table 5.3 show the number of tweets in our training and test sets, for each sentiment category.

Table 5.3: Distribution of tweets for each sentiment category

Sentiment	**Training**	**Test**	**Total**
Positive	2,886	2,888	5,774
Negative	16,477	16,478	32,995
Neutral	651	651	1,302

It can be observed from Table 5.3 that the negative class has the highest number of tweets, followed by the positive and neutral classes, respectively.

5.5.1 Experimental Setup

In order to classify tweets in the test sets according to sentiment, we use the features of the hashtags in the training sets. For each tweet in the test set, we determine if it contains at least one hashtag from the corresponding training set. Therefore, we use the presence of the hashtag as a feature to assign the tweet the same sentiment class as the hashtag. If the tweet contains multiple hashtags from the training set, we apply two features from the hashtags: the type of hashtag and, its frequency in the training set. These two features are applied as follows:

1. If the tweet contains multiple hashtags of sentiment type, then we determine the most subjective hashtag in the group by comparing the hashtags to opinion words in the FOW list. If one of the hashtags is found in this list, the tweet is assigned the same sentiment class as this hashtag. Otherwise, the tweet is assigned the sentiment class of the hashtag with the highest frequency.

2. If the tweet contains multiple hashtags of non-sentiment type, we determine the most descriptive hashtag in the group by selecting the hashtag that is not determined to be a noun. We use a POS tagger available in NLTK for python [29]. Then the tweet is assigned the same sentiment class as this hashtag. Otherwise, the tweet is assigned the sentiment class of the hashtag with the highest frequency.

5.5.2 Results and Discussion

Table 5.4 shows the accuracy, precision, recall, and f-measure metrics (in percent) for our models on the test set, for each type of hashtag.

Table 5.4: Sentiment classification of tweets

Hashtag type	**Model**	**Accuracy**	**Precision**	**Recall**	**F-measure**
Sentiment					
	Model 1	**79.27**	**73.14**	**77.42**	**73.88**
	Model 2	78.40	64.58	75.92	67.96
Non-sentiment					
	Model 1	**84.48**	**78.58**	**83.61**	**78.09**
	Model 2	84.17	69.91	83.21	75.96

It can be observed from Table 5.4 that both Models 1 and 2 achieve over 80% accuracy and precision in classifying tweets using non-sentiment hashtags. Both models achieve higher accuracy, recall, and f-measure than the corresponding models which are applied to classify tweets using sentiment hashtags.

In terms of classifying tweets using non-sentiment hashtags, Model 1 outperforms Model 2 in accuracy, precision, recall and f-measure. For sentiment hashtags, Model 1

also achieves higher accuracy, precision, recall and f-measure than Model 2. Based on the results from all four measures, it is evident that Model 1 is the better classification model.

5.5.2.1 Using Tweets with Labelled by Umigon

Additionally, we evaluate and compare the performance of Model 1 on tweets labelled separately by Sentiment140, and Umigon. Sentiment140 was previously used to assign labels to our dataset. Now, we apply a different classifier to automatically label tweets in our dataset. This classifier called Umigon applies a lexicon-based method to determine the overall sentiment of tweets.

Table 5.5 shows the accuracy, precision, recall and f-measure metrics (in percent) obtained by Model 1 on tweets labelled by Sentiment140 and Umigon, respectively.

Table 5.5: Comparing the performance of Model 1 on tweets labelled by Sentiment140 and Umigon

Hashtag type	**Automatic classifier**	**Accuracy**	**Precision**	**Recall**	**F-measure**
Sentiment	Sentiment140	79.27	73.14	**77.42**	73.88
	Umigon	**82.65**	**74.54**	74.60	**74.36**
Non-sentiment	Sentiment140	**84.48**	**78.58**	**83.61**	**78.09**
	Umigon	74.21	68.14	69.41	60.02

It can be observed from Table 5.5 that for sentiment hashtags, Model 1 achieves slightly higher accuracy, precision, and f-measure on tweets labelled by Umigon than on tweets labelled by Sentiment140. For tweets with non-sentiment hashtags, Model 1 achieved higher accuracy, precision, recall, and f-measure on tweets labelled by Sentiment140 than on tweets labelled by Umigon.

By focusing on the performance of Model 1 on tweets labelled by Umigon, we can observe that sentiment hashtags is more effective than non-sentiment hashtags in classifying tweets according to sentiment. This result contradicts with the results obtained from our earlier experiments. However, this is expected as Umigon is a lexicon-based classifier,

which relies on a lexicon to detect subjectivity, and also applies the literal meaning of hashtags in order to determine the overall sentiment of the tweet.

Table 5.6 shows the percentage difference between tweets with sentiment and non-sentiment bearing hashtags that are automatically labelled as positive, negative or neutral. Here, we focus only on the test sets.

Table 5.6: Percentage difference between tweets with sentiment and non-sentiment hashtags

Dataset	**Positive**	**Negative**	**Neutral**
Umigon	**8.83%**	**15.53%**	**24.35%**
Sentiment140	3.44%	5.47%	2.02%

It can be observed from Table 5.6 that the percentage differences between tweets with sentiment and non-sentiment bearing hashtags that are automatically labelled by Umigon for each sentiment category are larger than the percentage differences obtained using Sentiment140. This suggests that of the two classifiers, the Umigon classifier is more dependent on the sentiment and non-sentiment hashtags contained in the tweet in order to determine overall sentiment. Consequently, the highest percentage difference is found in the neutral class. Therefore, this further suggests that the lexicon-based Umigon classifier is less able to detect subjectivity in tweets with non-sentiment hashtags. Thus, this conclusion supports our findings that non-sentiment hashtags are not as effective as sentiment hashtags in classifying tweets labelled by Umigon.

Overall, we can conclude that for the classification of tweets using sentiment hashtags, the performance of our approach is independent of the method used to initially assign labels to tweets. In terms of the classification of tweets using non-sentiment hashtags, the performance of our approach is more dependent on the method used to initially assign labels to tweets.

5.5.2.2 Using Established Classifiers

Additionally, we compare our best model, Model 1, to four established classifiers, Naive Bayes, SVM, Maximum Entropy, and C4.5. We use the WEKA implementation of these classifiers. We modify the training and test sets previously used for Model 1, using binary values to indicate the presence and absence of each hashtag in each tweet. For tweets with sentiment hashtags, we add a feature for the presence or absence of strongly subjective hashtags in the tweet.

Tables 5.7 and 5.8 shows the accuracy, precision, recall, f-measure scores (in percent) for the five classifiers on the test set containing sentiment and non-sentiment hashtags, respectively.

Table 5.7: Classification of tweets using sentiment hashtags

Classifier	**Accuracy**	**Precision**	**Recall**	**F-measure**
Naive Bayes	77.18	74.60	77.20	73.10
SVM	**79.33**	**76.00**	**79.30**	**75.30**
Maximum Entropy	69.47	71.70	69.50	70.30
C4.5	78.99	73.70	79.00	71.60
Model 1	**79.27**	**73.14**	**77.42**	**73.88**

Table 5.8: Classification of tweets using non-sentiment hashtags

Classifier	**Accuracy**	**Precision**	**Recall**	**F-measure**
Naive Bayes	83.49	76.20	83.50	76.40
SVM	83.65	78.40	83.60	78.50
Maximum Entropy	83.17	77.60	83.20	**78.60**
C4.5	**84.04**	**79.10**	**84.00**	78.40
Model 1	**84.48**	**78.58**	**83.61**	**78.09**

It can be observed from Table 5.7 that SVM achieves the highest accuracy, precision, recall and f-measure. It must be noted that our model performs comparably to the established classifiers. We performed a paired 2-tailed T-test at the 95% confidence level on the evaluation measures for our model and SVM. We obtained a p-value of 0.0755 which demonstrate that there is not a statistically significant difference between the results obtained for our model and SVM.

Additionally, it can be observed from Table 5.8 that all five classifiers achieve over 83% accuracy and recall in classifying tweets using non-sentiment hashtags. Our model achieved the highest accuracy. Maximum Entropy achieved the highest f-measure whereas C4.5 achieved the highest precision and recall. We performed a paired 2-tailed T-test at the 95% confidence level on the evaluation measures for our model and C4.5. We obtained a p-value of 0.4333, which demonstrate that there is not a statistically significant difference between the results obtained for our model and C4.5.

Figure 5.4 shows the highest accuracy, precision, recall and f-measure metrics obtained for classifying tweets using sentiment and non-sentiment hashtags, respectively.

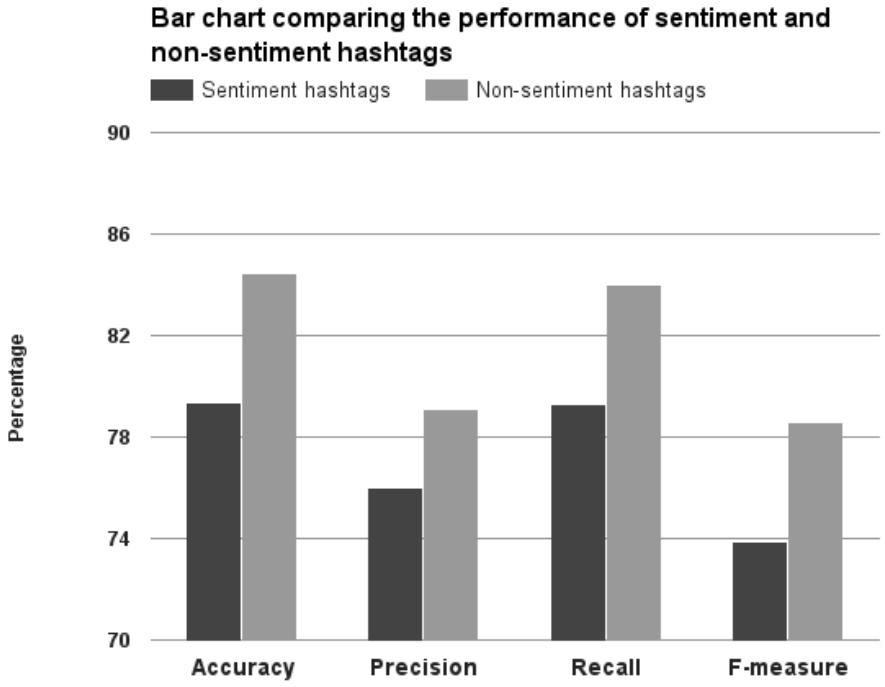

Figure 5.4: Overall performance of sentiment and non-sentiment bearing hashtags

It can be observed from Figure 5.4 that we obtained over 78% in all four evaluation metrics by using the features of non-sentiment bearing hashtags to classify tweets. Additionally, non-sentiment bearing hashtags outperform sentiment bearing hashtags in all four measures. We performed a paired 2-tailed T-test at the 95% confidence level on

the evaluation measures obtained using sentiment hashtags, and non-sentiment bearing hashtags. We obtained a p-value of 0.002, which demonstrate that there is a statistically significant difference between the results obtained for these two types of hashtags.

Overall, our experimental results show that non-sentiment bearing hashtags are more effective than sentiment hashtags for classifying tweets as positive, negative or neutral.

5.5.3 Classification of Subjective Tweets

We also evaluate the performance of our model in classifying only subjective tweets. Therefore, a binary classification task is undertaken to determine the effectiveness of our approach in classifying tweets as positive or negative. In addition to using our models, we developed a classification model (Model 3) using the scoring algorithm developed by Bakliwal et al. [4].

5.5.3.1 Results and Discussion

Table 5.9 shows the precision, recall, f-measure, and accuracy metrics (in percentage) for all the models on the test set, for each type of hashtag.

Table 5.9: Classification of subjective tweets

Hashtag type	**Model**	**Accuracy**	**Precision**	**Recall**	**F-measure**
Sentiment					
	Model 1	81.14	**76.94**	81.23	**77.77**
	Model 2	**81.72**	71.34	**81.84**	73.91
	Model 3 [4]	71.24	76.72	71.22	73.38
Non-sentiment					
	Model 1	**86.07**	**81.96**	**86.03**	**81.50**
	Model 2	85.97	73.89	85.92	79.46
	Model 3 [4]	71.70	81.13	71.73	75.22

It can be observed from Table 5.9 that both Models 1 and 2 achieve higher accuracy, recall, and f-measure scores than the model (Model 3) which applied the scoring algorithm by Bakliwal et al. [4]. Furthermore, our experimental results show that Model 1 outperforms Model 2 in classifying tweets using non-sentiment hashtags. For classifying

tweets using sentiment hashtags, Model 1 achieves higher precision and f-measure than Model 2. However, Model 2 achieves slightly higher accuracy and recall than Model 1. Overall, Model 1 is our best model.

We performed two separate paired 2-tailed T-tests at the 95% confidence level on the accuracy, recall and f-measure values obtained for applying Model 1 and Model 3 to classify tweets using sentiment hashtags, and non-sentiment bearing hashtags. We obtained p-values of 0.0487 and 0.0492, respectively for each test, which demonstrate that there is a statistically significant difference between the results obtained Model 1 and Model 3. Therefore, the tests show that our approach to the sentiment analysis of tweets is more effective.

Additionally, our experimental results show that both our models achieve higher accuracy, recall, precision, and f-measure scores in classifying subjective tweets using non-sentiment bearing hashtags than sentiment bearing hashtags. Therefore, we can conclude that non-sentiment hashtags are more effective than sentiment hashtags in classifying subjective tweets.

5.5.3.2 Using Established Classifiers

We apply Naive Bayes, SVM, Maximum Entropy and C4.5 on our test sets. Tables 5.10 and 5.11 shows the precision, recall, f-measure, and accuracy values (in percent) for the five classifiers (including our model) on the test set for sentiment and non-sentiment hashtags, respectively.

Table 5.10: Classification of subjective tweets using sentiment hashtags

Classifier	**Accuracy**	**Precision**	**Recall**	**F-measure**
Naive Bayes	80.81	77.20	80.80	78.20
SVM	**82.85**	79.70	**82.90**	**79.60**
Maximum Entropy	73.52	75.40	73.50	74.40
C4.5	82.78	**80.10**	82.80	76.90
Model 1	**81.14**	**76.94**	**81.23**	**77.77**

It can be observed from Tables 5.10 that SVM achieves the highest accuracy, recall

Table 5.11: Classification of subjective tweets using non-sentiment hashtags

Classifier	**Accuracy**	**Precision**	**Recall**	**F-measure**
Naive Bayes	85.90	80.50	85.90	79.90
SVM	86.12	82.30	86.10	82.00
Maximum Entropy	85.74	81.70	85.70	**82.10**
C4.5	**86.41**	**83.30**	**86.40**	81.80
Model 1	**86.07**	**81.96**	**86.03**	**81.50**

and f-measure scores whereas C4.5 achieves the highest precision, in classifying subjective tweets using sentiment hashtags. However, when we performed a paired 2-tailed T-test at the 95% confidence level on the evaluation measures for our model and SVM. The p-value of 0.0045 obtained demonstrate that there is a statistically significant difference between our model and SVM. This shows that our model is outperformed by SVM, significantly.

Additionally, it can be observed from Table 5.11 that C4.5 achieved the highest accuracy, precision and recall in classifying subjective tweets using non-sentiment hashtags. We performed a paired 2-tailed T-test at the 95% confidence level on the evaluation measures for our model and C4.5. The p-value of 0.1014 obtained demonstrate that there is not a statistically significant difference between our model and C4.5.

Figure 5.5 shows the highest accuracy, precision, recall and f-measure metrics obtained for classifying tweets using sentiment and non-sentiment hashtags, respectively.

It can observed from Figure 5.5 that, we achieve over 81% for all four evaluation measures by using the features of non-sentiment hashtags to determine the sentiment of tweets. Additionally, non-sentiment bearing hashtags outperform sentiment bearing hashtags in all four measures. We performed a paired 2-tailed T-test at the 95% confidence level on the evaluation measures obtained using sentiment hashtags, and non-sentiment bearing hashtags. We obtained a p-value of 0.0009, which demonstrate that there is a statistically significant difference between these two types of hashtags. Therefore, these results provide empirical evidence that non-sentiment bearing hashtags are more effective than sentiment hashtags for the classification of subjective tweets.

Overall, statistical tests demonstrate that there is a statistically significant difference

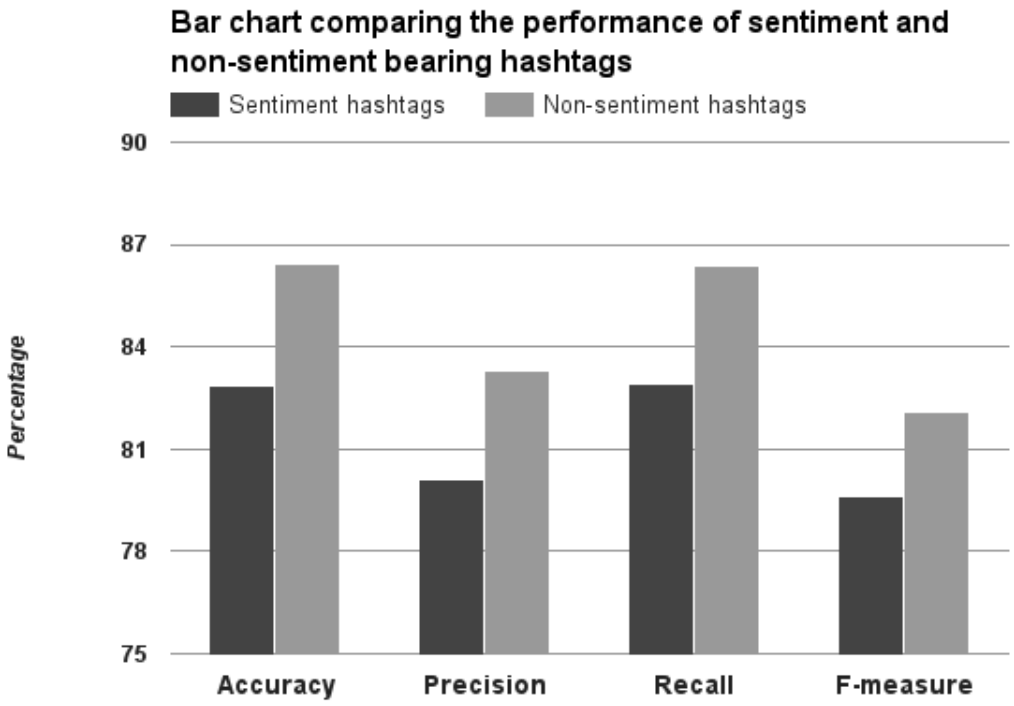

Figure 5.5: Comparing the performance of hashtags in classifying subjective tweets

between the accuracy, recall and f-measure values obtained for our best model and the model created using the scoring algorithm by Bakliwal et al. [4]. Additionally, our results demonstrate that our best model, Model 1, performs comparably to that of the established classifiers Naive Bayes, Maximum Entropy, SVM and C4.5. Therefore, this proves that hashtags can be used as accurate predictors of the sentiment of subjective tweets. Moreover, all our experimental results show that non-sentiment hashtags are better predictors than sentiment hashtags. This is supported by statistical tests, which indicate that there is a statistically significant difference between the results obtained for these two types of hashtags.

Chapter 6

Conclusion and Future Work

With the increasing popularity of Twitter, there is a need to examine some of its most unique features. One such feature is hashtags, which are user defined topics or keywords. In this research, we focused on the problem of determining whether hashtags can be accurate predictors of the sentiment of the entire tweet. We also determined whether hashtags containing sentiment information were a better predictor than those without.

In our approach to solving this problem, we divided the study into two phases. In the first phase, we first extracted hashtags from our dataset of tweets, and manually classified them as sentiment and non-sentiment bearing. Then, we developed a modified lexicon-based approach using subjective words from different lexical resources and a training dataset. Next, we applied our approach to predict the class of the hashtags extracted from the test dataset. In the second phase, we used the predicted hashtags to determine the sentiment (positive, negative or neutral) of tweets. Two classification models were developed using a training set, for each type of hashtag. The models applied scoring algorithms to determine the sentiment polarity of the hashtags. On the test sets, we used a simple classifier that applied this sentiment polarity and other features of the hashtags that were present in the training set, to determine the sentiment of the tweets. Both binary and three-category sentiment classification tasks were performed, and all of

our results were evaluated using recall, precision, accuracy and f-measure.

For the hashtag classification task, our model achieved 94.25% and 86.27% precision in identifying non-sentiment bearing and sentiment bearing hashtags, respectively. Furthermore, we attained the highest accuracy and f-measure scores in identifying sentiment hashtags when compared to models, which applied a single resource. Therefore, our results demonstrate that combining subjective words from different opinion lexicons and word lists is effective for identifying sentiment and non-sentiment bearing hashtags.

For three-category sentiment classification of tweets, Model 1 outperformed Model 2 for each type of hashtag across all evaluation measures. For both sentiment and non-sentiment hashtags, Model 1 was found to be comparable to that of the Naive Bayes, SVM, Maximum Entropy, and C4.5 classifiers. However, the highest accuracy, precision, recall and f-measure scores, were achieved using non-sentiment hashtags.

For binary classification, our models achieved over 80% accuracy, precision and f-measure scores for both types of hashtags. Furthermore, we compared our models to that of a model developed using a scoring algorithm by [4]. Statistical tests demonstrated that there was a statistically significant difference between the accuracy, recall and f-measure values obtained for our best model and the model created using the scoring algorithm by Bakliwal et al. [4]. Therefore, this demonstrates that our method, which applies features of hashtags including their sentiment polarity to determine the overall sentiment of tweets, is more effective.

Analyses of the evaluation measures obtained for both the binary, and the 3-category classification tasks clearly indicate that hashtags are accurate predictors of the sentiment of tweets. Moreover, non-sentiment bearing hashtags are found to be better predictors than sentiment bearing hashtags. Overall, our study provides empirical evidence that hashtags are useful for the sentiment analysis of tweets.

In terms of future work, we intend to investigate hashtags for topic-based sentiment analysis of tweets, and emotional classification of tweets.

Bibliography

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Portland, Oregon, 2011.

[2] Emilia Anderson. "what the tweet?" + twitter (part ii), August 2013. `http://techinreallife.com/2013/08/07/what-the-tweet-twitter-part-ii/`. Retrieved on April 20, 2015.

[3] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC'10, Valletta, Malta, 2010.

[4] Akshat Bakliwal, Piyush Arora, Senthil Madhappan, Nikhil Kapre, Mukesh Singh, and Vasudeva Varma. Mining sentiments from tweets. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 11–18, Portland, Oregon, 2012.

[5] Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 49–58, Atlanta, Georgia, 2013.

[6] Yanwei Bao, Changqin Quan, Lijuan Wang, and Fuji Ren. The role of pre-processing in twitter sentiment analysis. In De-Shuang Huang, Kang-Hyun Jo, and Ling Wang, editors, *Intelligent Computing Methodologies*, volume 8589 of *Lecture Notes in Computer Science*, pages 615–624. Springer International Publishing, 2014.

[7] Vangie Beal. Twitter dictionary: A guide to understanding twitter lingo, August 2014. `http://www.webopedia.com/quick\_ref/Twitter\_Dictionary\_Guide.asp`. Retrieved on September 20, 2014.

[8] Albert Bifet and Eibe Frank. Sentiment knowledge discovery in twitter streaming data. In *Proceedings of the 13th International Conference on Discovery Science*, DS'10, pages 1–15, Canberra, Australia, 2010.

[9] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[10] Nibir Nayan Bora. Summarizing public opinions in tweets. *International Journal of Computational Linguistics and Applications*, 3(1):41–55, 2012.

[11] Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, pages 2:1–2:9, Chicago, Illinois, 2013.

[12] Jeff Bullas. 20 twitter facts and statistics you need to know in 2014, August 2014. `http://www.business2community.com/twitter/20-twitter-facts-statistics-need-know-2014-0966740`. Retrieved on January 25, 2015.

[13] Fermín L. Cruz, José A. Troyano, F. Javier Ortega, and Fernando Enríquez. Automatic expansion of feature-level opinion lexicons. In *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 125–131, Portland, Oregon, 2011.

[14] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Beijing, China, 2010.

[15] Steven J. DeRose. The compass derose guide to emotion words, 2005. `http://www.derose.net/steve/resources/emotionwords/ewords.html`. Retrieved on September 20, 2014.

[16] Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 231–240, Palo Alto, California, USA, 2008.

[17] Tia Fisher. Top twitter abbreviations you need to know, May 2012. `http://www.socialmediatoday.com/content/top-twitter-abbreviations-you-need-know`. Retrieved on September 21, 2014.

[18] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Portland, Oregon, 2011.

[19] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. Technical report, Stanford University, 2009.

[20] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[21] Steven Hein. Feeling words/emotion words, 2013. http://eqi.org/fw.htm. Retrieved on September 20, 2014.

[22] 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*, KDD '04, pages 168–177, Seattle, WA, USA, 2004.

[23] V .S Jagtap and Karishma Pawar. Analysis of different approaches to sentence-level sentiment classification. *International Journal of Scientific Engineering and Technology*, 2(3):164–170, 2013.

[24] Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona*, Catalonia, Spain, 2011.

[25] Raffi Krikorian. New tweets per second record, and how!, August 2013. http://www.internetlivestats.com/twitter-statistics/. Retrieved on January 25, 2015.

[26] Clement Levallois. Umigon: sentiment analysis for tweets based on terms lists and heuristics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 414–417, Atlanta, Georgia, USA, 2013.

[27] Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. Emoticon smoothed language models for twitter sentiment analysis. In Jörg Hoffmann and Bart Selman, editors, *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012*, Toronto, Ontario, Canada, 2012.

[28] Shenghua Liu, Fuxin Li, Fangtao Li, Xueqi Cheng, and Huawei Shen. Adaptive co-training svm for sentiment classification on tweets. In *Proceedings of the 22Nd*

ACM International Conference on Conference on Information & Knowledge Management, CIKM '13, pages 2079–2088, San Francisco, California, USA, 2013.

[29] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Philadelphia, Pennsylvania, 2002.

[30] Justin Martineau and Tim Finin. Delta TFIDF: an improved feature space for sentiment analysis. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009*, San Jose, California, USA, 2009.

[31] Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, and Arturo Montejo-Raéz. Sentiment analysis in twitter. *Natural Language Engineering*, 20(01):1–28, 2014.

[32] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093 – 1113, 2014.

[33] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11): 39–41, November 1995.

[34] Gary D. Miner, Dursun Delen, John Elder, Andrew Fast, Thomas Hill, and Robert A. Nisbet. In Gary MinerDursun DelenJohn ElderAndrew FastThomas HillRobert A. Nisbet, editor, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, pages 929 – 934. Academic Press, Boston, 2012.

[35] Saif M. Mohammad. #emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 246–255, Montreal, Canada, 2012.

[36] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *CoRR*, abs/1308.6242, 2013.

[37] Rebecca Murtagh. The role of #hashtags in social media and search, November 2013. `http://searchenginewatch.com/sew/how-to/2305444/the-role-of-hashtags-in-social-media-and-search`. Retrieved on November 20, 2014.

[38] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2Nd International Conference on Knowledge Capture*, K-CAP '03, pages 70–77, Sanibel Island, FL, USA, 2003.

[39] Mark Nichol. 100 mostly small but expressive interjections, January 2011. `http://www.dailywritingtips.com/100-mostly-small-but-expressive-interjections/`. Retrieved on January 30, 2015.

[40] Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903, 2011.

[41] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta, 2010.

[42] Prabu Palanisamy, Vineet Yadav, and Harsha Elchuri. Serendio: Simple and practical lexicon based approach to sentiment analysis. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 543–548, Atlanta, Georgia, USA, 2013.

[43] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002.

[44] Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I*, ISWC'12, pages 508–524, Boston, MA, 2012.

[45] Sentiment140. `http://help.sentiment140.com/api`.

[46] Anuj Sharma and Shubhamoy Dey. Article: Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis. *IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications*, ACCTHPCA(3):15–20, July 2012.

[47] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA, 1966.

[48] Mike Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength., 2012. `http://sentistrength.wlv.ac.uk/documentation/SentiStrengthChapter.pdf`. Retreived on April 2nd 2014.

[49] Twitter, 2015. `https://dev.twitter.com/`.

[50] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. Harnessing twitter "big data" for automatic emotion identification. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 587–592, Washington, DC, USA, 2012.

[51] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1031–1040, Glasgow, Scotland, UK, 2011.

[52] 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*, HLT '05, pages 347–354, Vancouver, British Columbia, Canada, 2005.

[53] Lei Zhang, Mohamed Ghosh, Riddhiman andDekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. Technical report, HP Laboratories, 2011.