# Evaluating the Effectiveness of Hashtags as Predictors of the Sentiment of Tweets

Credell Simeon(✉) and Robert Hilderman

University of Regina, Regina, SK S4S 0A2, Canada
{simeon3c,Robert.Hilderman}@uregina.ca

**Abstract.** Recently, there has been growing research interest in the sentiment analysis of tweets. However, there is still a need to examine the contribution of Twitter-specific features to this task. One such feature is hashtags, which are user-defined topics. In our study, we compare the performance of sentiment and non-sentiment hashtags in classifying tweets as positive or negative. By combining subjective words from different lexical resources, we achieve accuracy scores of 83.58 % and 83.83 % in identifying sentiment hashtags and non-sentiment hashtags, respectively. Furthermore, our accuracy scores surpass those scores obtained using models that apply a single lexical resource. We apply derived properties of sentiment and non-sentiment hashtags, including their sentiment polarity to classify tweets. Our best classification models achieve accuracy scores of 81.14 % and 86.07 % using sentiment hashtags and non-sentiment hashtags, respectively. Additionally, our models perform comparably to supervised machine learning algorithms, and outperform a scoring algorithm developed in a previous study.

## 1 Introduction

Since its inception in 2006, Twitter, a microblogging application, has gained increasing popularity with approximately 302 million monthly users and an estimated 500 million new daily posts[1]. Twitter provides a platform whereby registered users can post short text messages called tweets. Tweets are opinionated statements, which convey sentiments about different topical issues. Therefore, we can apply sentiment analysis to determine whether the sentiment contained within the text is either positive or negative [5]. Positive tweets express favorability whereas negative tweets express unfavorability towards a subject. Thus, sentiment analysis is useful in assessing people's attitudes and emotions towards products and services offered by businesses [13], or political candidates in general elections [3].

For sentiment analysis, research studies have applied both machine learning techniques and lexicon-based methods. The lexicon-based approach depends entirely on using opinion lexicons, which are dictionaries of positive and negative words, to detect subjectivity in text [19]. By contrast, supervised machine learning applies learning algorithms to large number of labeled data. Unlike other

[1] https://about.twitter.com/company.

N. Japkowicz and S. Matwin (Eds.): DS 2015, LNAI 9356, pp. 251–265, 2015.
DOI: 10.1007/978-3-319-24282-8_21

text, tweets can contain a significant amount of information which can make the sentiment analysis task challenging [4]. Therefore, it is important to examine the unique nature of tweets, as this plays a significant role in determining their overall sentiment.

Tweets are highly informal text messages, which are restricted to 140 characters. They are conversational in nature and thus, they contain many features including: abbreviations, slangs, acronyms, repetitions of characters in words e.g., “yeaaaah”, punctuation marks, and emoticons. In terms of Twitter-specific features, these are described below.

1. **Retweets** are copies of an original tweet that are posted by other users [7]. 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 consecutive words, e.g., “#believe” and “#wishfulthinking”, respectively. A tweet can contain multiple hashtags, which can be located anywhere in the text.

Of all the features of tweets described previously, hashtags have been selected as the focus of our study. 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 popular content, engage other users, and group related content. Most importantly, hashtags are useful for determining the popularity of topics, and the overall sentiment that is being expressed by groups of users. Consequently, hashtags are being used by many other platforms including photo-sharing applications such as Instagram[2] and social networks like Facebook[3], Tumblr[4] and Google+[5].

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 that were proposed in a previous study [17]. Examples of sentiment and non-sentiment hashtags are “#best” and “#football”, respectively.

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,

[2] http://instagram.com/.

[3] http://facebook.com/.

[4] https://www.tumblr.com/.

[5] https://plus.google.com/.

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 [4,6,19]. Thirdly, by focusing on a single feature, we reduce the effort required in determining the optimal combination of the various features in the tweets. Therefore, our study applies the derived properties of hashtags, including their sentiment polarity, in order to classify tweets as positive and negative. We describe these properties as "derived" because they are not part of the definition of a hashtag, but they are resulting because of it. For instance, a hashtag may contain a subjective word, thus we consider that there are two types of hashtags: sentiment and non-sentiment bearing. Additionally, we consider all hashtags to have a polarity which can be determined by examining the tweets that contain them. Therefore, in this study, we compare the effectiveness of sentiment and non-sentiment hashtags for classifying subjective tweets.

The main contributions of our paper are summarized as follows:

1. It demonstrates the effectiveness of combining different lexical resources to identify sentiment from non-sentiment bearing hashtags.
2. It presents different scoring algorithms for determining the sentiment polarity of hashtags.
3. It demonstrates the effectiveness of using the derived properties of hashtags, including their sentiment polarity, for the sentiment classification of tweets.
4. It shows that non-sentiment hashtags are more effective at classifying tweets as positive and negative, than sentiment hashtags.

The remainder of this paper is organized as follows. Section 2 summarizes previous studies on the sentiment analysis of tweets. Section 3 describes the development of the our approach. Section 4 discusses our experimental results, and compares these results with that of another study. Finally, Sect. 5 presents our conclusions and plans for future work.

## 2 Related Work

Sentiment analysis of tweets has garnered much research interest in recent years. A study by [2] demonstrated that a scoring algorithm can be used to accurately classify positive and negative tweets [2]. They applied the function to two separate datasets, Stanford [7] and Mejaj [5], which used emoticons, and sentiment suggestive words as sentiment labels, respectively. The scoring function calculated an overall score for each tweet by aggregating the difference in the positive and negative probabilities of unigrams, and assigning predefined weights to emoticons and punctuations. After applying stop word removal, stemming, spelling correction and noun identification, the function applied a Popularity Score in order to boost the scores of domain specific words. Tweets were determined to be positive (negative) if the sum of their sentiment scores was greater (less) than zero. Experimental results revealed that the Stanford and Mejaj datasets achieve accuracy scores of 87.2 % and 88.1 %, respectively. Also, these accuracy scores were comparable to that obtained using a SVM classifier.

In terms of the contribution of hashtags to the sentiment analysis of tweets, very few studies have focused on this task. Kouloumpis et al. [9] investigated Twitter hashtags for identifying positive, negative and neutral tweets such that the polarity of the tweet is determined by the hashtag. Using linguistic, lexical and microblogging features extracted from tweets, an AdaBoost.MH classifier achieved accuracy scores of 74 % and 75 % on hashtagged, and emoticon datasets, respectively. However, their study only focused on tweets containing a single hashtag. By contrast, Mohammad [10] analyzed self-labeled hashtagged emotional words in tweets, and concluded that they are good indicators of the sentiment of the entire tweet. In a later study, Mohammad et al. [11] developed a hashtag sentiment lexicon 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 hashtag lexicon to classify sentiment in tweets, the performance of the classifier increased by 3.8 %. Therefore, both studies demonstrate that hashtags can be useful in the sentiment analysis of tweets.

Wang et al. [17] applied a graph-based approach for classifying sentiment in hashtags as either positive or negative by incorporating hashtag co-occurrence information, their literal meaning, and the sentiment polarity distribution of tweets. By doing so, they showed that the polarity distribution of tweets can be combined with hashtag information for sentiment classification.

Rodrigues Barbosa et al. [14] performed a preliminary investigation into determining the effectiveness of hashtags in the sentiment analysis of tweets. The study focused specifically on using hashtags to detect and track online population sentiment. In order to do this, the authors studied hashtag propagation patterns, and the use of hashtags to express sentiment in tweets. Using a dataset of tweets on elections in Brazil, the authors manually identified frequent positive and negative hashtags. After performing analysis on the hashtags, the results revealed that in some cases hashtags were required for defining the sentiment of the tweet. Overall, this study concluded that hashtags may be useful for the sentiment analysis of tweets. By contrast, our study seeks to demonstrate conclusively that hashtags are accurate predictors of the sentiment of tweets.

## 3 Method

In order to investigate the effectiveness of hashtags as predictors of the overall sentiment of tweets, we divide the project into two main phases. In the first phase, we develop a modified lexicon-based approach to automatically classify hashtags as either sentiment or non-sentiment bearing. In the second phase, we apply supervised machine learning to classify tweets containing these hashtags as either positive or negative.

### 3.1 Phase 1: Classification of Hashtags

In the modified lexicon-based approach, the subjective words from different sentiment resources are used to detect subjectivity in the hashtags extracted from the tweets. Hashtags are stripped of their hash symbol, and their stems are found using a Regexp stemmer from Natural Language Processing Toolkit (NLTK)[6].

Sentiment resources refer to both opinion lexicons and word lists of sentiment terms. In our study, we use seven opinion lexicons (listed from smallest to largest): AFINN [12], SentiStrength [16], Bing Liu Lexicon [8], Subjectivity Lexicon [18], General Inquirer [15], NRC Hashtag Sentiment Lexicon [11], and SentiWordNet [1]. For each lexicon, we extract all positive and negative words. However, there are a few lexicons, in which we extract only the strongly subjective words. For SentiStrength Lexicon, we extract positive and negative words with semantic orientations greater than 2.0, and less than $-2.0$, respectively. For NRC Hashtag Sentiment Lexicon, we extract the top 500 adjectives for each sentiment class (positive and negative). For SentiWordNet, 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.

We also use three lists of sentiment words: Steven Hein feeling words[7] which contains 4232 words, The Compass DeRose Guide to Emotion Words[8] which consists of 682 words, and sentiment bearing Twitter slangs and acronyms collected from various online sources[9,10]. Most of these words are not found in the other lexicons. Examples include "fab" for "fabulous", and "HAND" for "Have a Nice Day".

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 selected combinations. These 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 strongly subjective words. 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 list of 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[11], giving a total of 308 words.

[6] www.nltk.org.

[7] http://eqi.org/fw.htm.

[8] http://www.derose.net/steve/resources/emotionwords/ewords.html.

[9] http://www.socialmediatoday.com/content/top-twitter-abbreviations-you-need-know.

[10] http://www.webopedia.com/quick_ref/Twitter_Dictionary_Guide.asp.

[11] http://www.dailywritingtips.com/100-mostly-small-but-expressive-interjections/.

**Model Development.** The classification model uses a binary search algorithm 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, each hashtag is compared with each opinion 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.

The second step focuses on the hashtags that have not been matched after the first step. 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. Both algorithms return a list of substrings sorted in descending order of length. 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, we consider only matches to the most subjective words are considered. Additionally, we only consider substrings of the hashtag that contain 3 or more characters. Our two recursive algorithms are described below.

1. **reduce_hashtag** eliminates the rightmost character from the hashtag after each iteration. The remaining characters form the left substring, whereas the removed character(s) form the right substring. For example, the hashtag, “lovestory” has 10 substrings: “lovestor”, “lovesto”, “lovest”, “estory”, “loves”, “story”, “love”, “tory”, “lov”, and “ory”.
2. **remove_left** removes the leftmost character from the hashtag after each iteration. Using this algorithm, six relevant substrings of the pre-processed hashtag, “lovestory”, are found: “ovestory”, “vestory”, “estory”, “story”, “tory”, and “ory”.

Initially, the reduce_hashtag algorithm is 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. If the search is unsuccessful, then the remove_left algorithm is applied.

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 the presence of extra repeated letters (at least 3), exclamation or question marks.

Table 1 outlines the eight rules for determining whether a hashtag is sentiment bearing. 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.

**Table 1.** 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 |

### 3.2 Phase 2: Classification of Tweets

In this phase, we develop different scoring algorithms that can be used in conjunction with various classifiers, in determining the sentiment polarity of tweets. We only consider tweets with hashtags.

**Model 1.** In this 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. This is the simplest model.

**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 [7]. 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 and stop words are removed from the dataset. Then, each unique word, $word_f$, in the tweet is used as a feature. The frequency of each word in the different sentiment classes is calculated. Then the positive and negative ratios are found using Eqs. 1, and 2. The positive ratio shown in Eq. 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.

$$positive\,ratio(word) = \frac{pos(word) - (neg(word))}{pos(word)} \tag{1}$$

The negative ratio shown in Eq. 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.

$$negative\,ratio(word) = \frac{neg(word) - (pos(word))}{neg(word)} \tag{2}$$

The sentiment polarity of the word, $sp(word)$, is the maximum of the positive and negative ratios. The sentiment weight of each word, $word_{sw}$, is determined as the product of $sp(word)$ and $word_f$.

Additionally, emoticons, punctuation marks, and negating words are also incorporated as features into the model. Positive emoticons and exclamation marks are assigned a polarity of 1, whereas negative emoticons, question marks and negations are assigned a polarity of −1. The feature weight $feature_{fw}$ of each feature is described in Eq. 3

$$feature_{fw} = \frac{count(fw)}{frequency_{fw}} \times sp(feature) \tag{3}$$

where $count(fw)$, is the frequency of the word in the tweet, $frequency_{fw}$, is the total frequency in the dataset and $sp(feature)$, is the sentiment polarity of the word. Then, the sentiment score of each hashtag in the tweet is the weighted average of all the features (including the words) which is determined by Eq. 4 as

$$hastag\_score = \frac{\sum_{i=1}^{n} feature_{fw_i} \times sp(feature_i)}{\sum_{i=1}^{n} feature_{fw_i}} \tag{4}$$

where $n$, is the number of features. The sentiment score for each hashtag in the tweet ranges from −1 to 1. If the score is greater than 0, the hashtag is considered to be positive. Otherwise, the hashtag is considered to be negative.

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 or negative, and assign the sentiment class with the highest frequency.

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

## 4 Experimental Results

### 4.1 Dataset

Our dataset consists of 71,836 unique tweets with hashtags, which are extracted using the Twitter API[12]. The tweets were collected during the FIFA World Cup 2014 using search terms (excluding hashtags) related to the football matches, in order to capture the opinions of the Twitter users during each game. We use Sentiment140 API[13] to automatically assign sentiment labels to the tweets in our dataset. Sentiment140 uses a Maximum Entropy classifier with an accuracy of 83 percent on a combination of unigrams and bigrams [7]. Positive, negative, and neutral tweets are assigned numerical values of 4, 2, and 0, respectively.

[12] http://www.dev.twitter.com/.

[13] http://help.sentiment140.com/api.

### 4.2 Evaluation Measures

Our evaluation measures are accuracy, precision, recall and f-measure. Accuracy measures the number of tweets (hashtags) for each class that are classified correctly. Precision determines the ratio of actual relevant tweets (hashtags) among predicted tweets (hashtags) for the sentiment category. Recall refers to the fraction of relevant tweets (hashtags) actually classified by the model. F-measure is the average of precision and recall.

### 4.3 Classification of Hashtags

Hashtags are extracted from the dataset and manually classified. Hashtags belonging to the same type are grouped. For each hashtag type, we selected all the tweets containing at least one hashtag of the respective type. Then, we divided this group of tweets equally into training and test sets. Table 2 shows the total number of hashtags extracted from the training and test sets.

**Table 2.** Training and test sets for each type of hashtag

| Type | Train | Test | Total |
|---|---|---|---|
| Sentiment | 1,368 | 1,376 | 2,744 |
| Non-sentiment | 3,070 | 3,142 | 6,212 |

In order to evaluate our model, we compare the hashtags extracted in the training and test sets. The hashtags in the test set is compared with the list of determined hashtags in the training set. If the hashtag is found in this list, the same class label is assigned. If it is not found, then similarity testing is performed where we compare their stems and length (threshold of 95 %) of the hashtags to determine a suitable match. Then, we compare the predicted class label assigned by the model to that of actual label of the hashtag assigned during manual annotation in order to evaluate our model.

Table 3 shows examples of the sentiment and non-sentiment hashtags that are identified by our model. Table 4 shows the precision, recall, f-measure, and accuracy metrics (in percent) obtained by our classification model.

**Discussion.** It can be observed from Table 4 that our model achieved higher percentages for accuracy, precision, recall and f-measure in identifying non-sentiment hashtags than 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 model, we created models which used a single lexical resource in order to identify sentiment hashtags from non-sentiment hashtags. Figure 1 shows the accuracy scores for the top five models. It can be observed in Fig. 1 that our model (last column), which used combined

**Table 3.** Examples of sentiment and non-sentiment hashtags classified by our model

| Type | Examples |
|---|---|
| Sentiment | #stressful, #helpmeunderstand, |
| | #shedoesntlookveryhappy, #strong |
| | #celebration #mindblowing |
| Non-sentiment | #budweiser, #dance |
| | #2014fifaworldcup, #children |
| | #teambrazil, #waiting |

**Table 4.** Results for 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** |

resources is the most accurate in identifying sentiment hashtags when compared with models which used a single lexical resource. For the identification of non-sentiment hashtags, our model performs comparably to one of the models which used a single lexical resource. Therefore, the experimental results show that using subjective words from different lexical resources is effective in boosting the identification of sentiment hashtags.

### 4.4 Classification of Tweets

We use the sentiment and non-sentiment hashtags that are classified by our model to select tweets that contain these hashtags. These tweets form our training and test set for each sentiment class. Table 5 show the number of tweets in our training and test sets, for each sentiment class.

**Table 5.** Tweets for sentiment classification

| Dataset | Train | Test | Total |
|---|---|---|---|
| Positive | 2,886 | 2,888 | 5,774 |
| Negative | 16,477 | 16,478 | 32,995 |

In order to classify positive and negative tweets in the test sets, we use the derived properties 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. Then, we assign the tweet the same sentiment polarity as the hashtag. If the tweet contains multiple hashtags from the training set, we apply two derived properties of the hashtags: the type of hashtag and, its frequency in

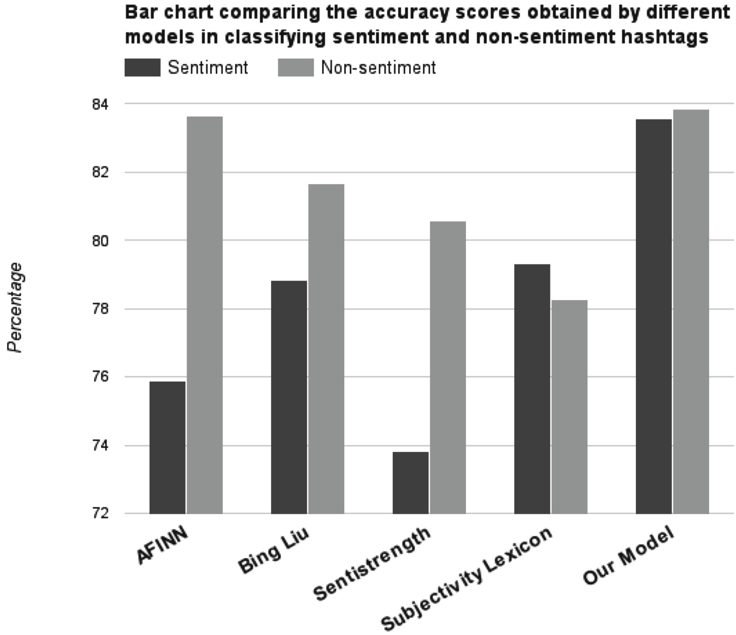

**Fig. 1.** Comparing the accuracy of our model to models using a single resource

the training set. For classifying tweets using sentiment hashtags, 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 polarity as this hashtag. Otherwise, the tweet is assigned the sentiment polarity of the hashtag with the highest frequency. For classifying tweets using non-sentiment hashtags, 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 in NLTK for python. Then the tweet is assigned the same sentiment polarity as this hashtag. Otherwise, the tweet is assigned the sentiment polarity of the hashtag with the highest frequency.

Table 6 shows the precision, recall, f-measure, and accuracy metrics (in percent) for our models on the test set, for each type of hashtag.

**Discussion.** It can be observed from Table 6 that both Model 1 and 2 for non-sentiment hashtags achieve higher accuracy, recall, precision, and f-measure in classifying tweets as positive and negative than Models 1 and 2 for sentiment hashtags. Therefore, non-sentiment hashtags are more effective in classifying tweets as positive and negative than sentiment hashtags.

Furthermore, our experimental results show that Model 1 outperforms Model 2 in classifying tweets using non-sentiment hashtags. For classifying tweets with sentiment hashtags, Model 1 achieves higher precision and f-measure than

**Table 6.** Results for classification of 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 |
| | Bakliwal et al. [2] | 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 |
| | Bakliwal et al. [2] | 71.70 | 81.13 | 71.73 | 75.22 |

Model 2. However, Model 2 achieves slightly higher accuracy and recall than Model 1. Overall, Model 1 is the better classification model.

In order to further evaluate our models, we apply the scoring algorithm created by Bakliwal et al. [2] to our dataset. Experimental results show that all of our models achieve higher accuracy, recall and f-measure scores than the model which applied the scoring algorithm by Bakliwal et al. [2]. However, the models created using the scoring algorithm by Bakliwal et al. [2], achieve the highest precision.

We then compare Model 1 for each hashtag type 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 1 and 0 to indicate the presence and absence of each hashtag in each tweet. Tables 7 and 8 shows the precision, recall, f-measure, and accuracy values (in percent) for the five classifiers for the test set on sentiment and non-sentiment hashtags, respectively.

**Table 7.** Results for classification 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 7 and 8 that our models performed quite comparably to the established classifiers. Additionally, all five models which applied non-sentiment hashtags achieve higher accuracy, precision, recall and f-measure scores than the models which applied sentiment hashtags. Therefore, this suggests that non-sentiment hashtags are more effective than sentiment hashtags in classifying tweets as positive or negative.

**Table 8.** Results for classification 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** |

Overall, all our experimental results show that non-sentiment hashtags are more effective in classifying tweets as positive and negative than sentiment hashtags. Additionally, Model 1 is determined to be the better model. Model 1 significantly outperforms the model created using the scoring algorithm by Bakliwal et al. [2], and performs comparably to that of the established classifiers, which demonstrates that our method is effective.

## 5 Conclusions and Future Work

In this paper, we evaluated the effectiveness of hashtags as accurate predictors of the sentiment of tweets. First, we applied a modified lexicon-based approach, which incorporated subjective words from different lexical resources, in order to accurately distinguish sentiment-bearing hashtags from non-sentiment hashtags. Using this model, we are able to achieve accuracy of 83.58 % and 83.83 % in identifying sentiment hashtags and non-sentiment hashtags, respectively. Furthermore, our accuracy surpassed those scores obtained using models that applied a single lexical resource.

Then, we applied the derived properties of hashtags to classify tweets as positive and negative. We developed and evaluated two separate classification models using training and test datasets of tweets. Our best models achieved accuracy scores of 81.14 % and 86.07 % in classifying tweets using sentiment hashtags and non-sentiment hashtags, respectively. Additionally, the performance of our models outperforms a previously developed algorithm [2] but is comparable to established classifiers. Finally, all our experimental results clearly indicate that non-sentiment hashtags are more effective than sentiment hashtags for the sentiment analysis of tweets.

In terms of future work, we will extend our work to include neutral tweets, and we plan to use hashtags for topic-based sentiment analysis.

## References

1. Baccianella, S., Esuli, A., Sebastiani, F.: 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 2010, Valletta, Malta (2010)
2. Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., Varma, V.: Mining sentiments from tweets. In: Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, WASSA 2012, Portland, Oregon, pp. 11–18 (2012)
3. Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., Hughes, M.: Sentiment analysis of political tweets: towards an accurate classifier. In: Proceedings of the Workshop on Language Analysis in Social Media, Atlanta, Georgia, pp. 49–58 (2013)
4. Bifet, A., Frank, E.: Sentiment knowledge discovery in twitter streaming data. In: Pfahringer, B., Holmes, G., Hoffmann, A. (eds.) DS 2010. LNCS, vol. 6332, pp. 1–15. Springer, Heidelberg (2010)
5. Bora, N.N.: Summarizing public opinions in tweets. Int. J. Comput. Linguist. Appl. **3**(1), 41–55 (2012)
6. Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using twitter hashtags and smileys. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING 2010, Beijing, China, pp. 241–249 (2010)
7. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Technical report, Stanford University (2009)
8. Minqing, H., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2004, Seattle, WA, USA, pp. 168–177 (2004)
9. Kouloumpis, E., Wilson, T., Moore, J.D.: 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)
10. Mohammad, S.: #emotional tweets. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – vol. 1: Proceedings of the main conference andthe shared task, and vol. 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), Montréal, Canada, pp. 246–255, 7–8 June 2012
11. Mohammad, S.M., Kiritchenko, S., Zhu, X.: Nrc-canada: building the state-of-the-art in sentiment analysis of tweets. CoRR, abs/1308.6242 (2013)
12. Nielsen, F.Å.: A new ANEW: evaluation of a word list for sentiment analysis in microblogs. CoRR, abs/1103.2903 (2011)
13. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - vol. 10, EMNLP 2002, Stroudsburg, PA, USA, pp. 79–86 (2002)
14. Rodrigues Barbosa, G.A., Silva, I.S., Zaki, M., Meira Jr., W., Prates, R.O., Veloso, A.: Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment. In: CHI 2012 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2012, Austin, Texas, USA, pp. 2621–2626 (2012). ISBN 978-1-4503-1016-1

15. Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge (1966)
16. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment in short strength detection informal text. J. Am. Soc. Inf. Sci. Technol. **61**(12), 2544–2558 (2010)
17. Wang, X., Wei, F., Liu, X., Zhou, M., Zhang, M.: 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 2011, Glasgow, Scotland, UK, pp. 1031–1040 (2011)
18. Wilson, T., Wiebe, J., Hoffmann, P.: 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 2005, Vancouver, British Columbia, Canada, pp. 347–354 (2005)
19. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B.: Combining lexicon-based and learning-based methods for twitter sentiment analysis. Technical report, HP Laboratories (2011)