Using Combined Lexical Resources to Identify Hashtag Types

Credell Simeon
University of Regina
3737 Wascana Parkway
Regina, Sk S4S 0A2
simeon3c@uregina.ca

Robert Hilderman
University of Regina
3737 Wascana Parkway
Regina, Sk S4S 0A2
Robert.Hilderman@uregina.ca

Abstract

This paper seeks to identify sentiment and non-sentiment bearing hashtags by combining existing lexical resources. By using a lexicon-based approach, we achieve 86.3% and 94.5% precision in identifying sentiment and non-sentiment hashtags, respectively. Moreover, results obtained from both of our classification models demonstrate that using combined lexical, emotion and word resources is more effective than using a single resource in identifying the two types of hashtags.

1 Introduction

In recent years, there has been increasing use of microblogs like Twitter where users post short text messages called tweets. One of the most unique and distinctive features found in tweets are hashtags. They are user-defined topics or keywords that are denoted by the hash symbol "#", followed immediately by a single word or multi-word phase joined without spaces (Qadir and Riloff, 2013). A valid hashtag is a community-driven convention that connects related tweets, topics and communties of users. Therefore, they are ideal for promoting specific ideas, searching for and organizing content, tracking customers feedback, and building social conversations. By using hashtags, Twitter users can significantly increase the engagement of their audience (Khan, 2015).

Moreover, hashtags may contain sentiment information. Examples include "#goodluck", "#enjoy", "#wellplayed", and "#worldcupfever". These hashtags can be useful in determining the overall opinion of tweets. Qadir and Riloff (2014) suggest that such hashtags reflect the emotional state of the author, while others (Davidov et al., 2010; Mohammad, 2012) concur that these emotions are not conveyed by the other words in the tweet. By contrast, some hashtags do not contain any sentiment information. Examples include "#soccer", "#USA", "#worldcup", and "#imwatching", respectively. They can be useful in event detection and topic classification of tweets. In our study, hashtags with sentiment information and those without are referred to as sentiment and non-sentiment bearing, respectively.

Because of the heightened interest in the sentiment analysis of tweets, it is important that we are able to identify sentiment and non-sentiment bearing hashtags, accurately. Therefore, in this paper, we propose using existing lexical and word resources to automatically classify these two types of hashtags. We apply a lexicon-based approach to develop two classification models, which use subjective words from different lexical, emotion and word resources. By employing this approach, we intend to demonstrate that using combined resources is more effective than using a single resource for identifying sentiment from non-sentiment bearing hashtags.

Paper organization The rest of the paper is organized as follows: Section 2 outlines related work, Section 3 details the opinion lexicons used, Section 4 describes our proposed methodology, Section 4 discusses our experimental results, and Section 6 presents our conclusion.

2 Related Work

Very few research studies have focused on analyzing hashtags. Wang et al. (2011) proposed that there were three types of hashtags: topic, sentiment-topic and sentiment. Each type refers to the kind of information that is contained within the hashtag such that sentiment-topic hashtags contain both topic and sentiment information. Therefore, there are two types of hashtags with sentiment information, and one type that refer only to topic information. They also classified positive and negative hashtags by using a graph-based approach

Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pages 169–174,
Lisboa, Portugal, 17 September, 2015.

that incorporated their co-occurrence information and literal meaning, and the sentiment polarity of tweets. Experimental results showed that the highest accuracy of 77.2% was obtained with Loopy Belief Propagation with enhanced boosting.

In terms of the most relevant work, Simeon and Hilderman (2015) showed that sentiment and non-sentiment hashtags are accurate predictors of the overall sentiment of tweets. The authors applied a lexicon-based approach to identify the two hashtag types, and then employed supervised machine learning to classify positive and negative tweets containing these hashtags. The experimental results obtained indicated that non-sentiment hashtags are better predictors than sentiment hashtags.

By contrast, Qadir and Riloff (2013) applied a bootstrapping approach in order to automatically learn hashtagged emotion words from unlabeled data. Hashtags were categorized as belonging to one of five sentiment categories: affection, anger/rage, fear/anxiety, joy and sadness/disappointment. Using five hashtags as seed words for each emotion class and a logistic regression classifier, additional hashtags were learned from unlabeled tweets. The learned hashtags were then used to classify emotion in tweets. Experimental results for emotional classification showed that their method achieved higher precision than recall. In a later study, Qadir and Riloff (2014) extended their work to include hashtag patterns and phrases associated with these five sentiments.

In this study, we focus on classifying hashtags into two types: sentiment and non-sentiment bearing. Our main goal is to demonstrate that combining lexical, emotion and word resources is more effective for this classification task than using a single lexical resource. Furthermore, by using this approach, we can reduce dependency on manual annotation, and increase the use of hashtags in the sentiment analysis of tweets.

3 Opinion lexicons

Opinion lexicons are dictionaries of positive and negative terms. For our approach, we employ a number of publicly available lexical resources. They include the manually annotated opinion lexicons of SentiStrength (Thelwall, 2012), AFINN (Nielsen, 2011), Bing Liu (Hu and Liu, 2004), General Inquirer (Stone et al., 1966) and Subjectivity lexicon (Wilson et al., 2005), and the automatically annotated lexicons of SentiWordNet (Baccianella et al., 2010) and NRC Hashtag Sentiment lexicon (Mohammad et al., 2013). They are described below.

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

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, it is specifically created for detecting sentiment in microblogs.

3. **General Inquirer** contains over 11,000 words grouped into different sentiment (positive and negative), and mood categories.

4. **Bing Liu Lexicon** contains about 6800 positive and negative words extracted from opinion sentences in customer reviews. It contains misspellings, slangs and other social media expressions.

5. **Subjectivity Lexicon** contains about 8,221 words categorized as strong or weak. For each word, a prior polarity (non-numerical score) is assigned, which can be positive, negative or neutral.

6. **SentiWordNet 3.0** is the largest lexicon containing over 115,000 synsets. A synset is a group of synonymous words with numerical scores for positivity, negativity and objectivity, which sums to a total of one.

7. **NRC Hashtag Sentiment Lexicon** consists of 54,129 unigrams. It is word-sentiment association lexicon that was created using 78 positive and negative hashtagged seed words, and a set of about 775,000 tweets.

4 Proposed Methodology

For this binary classification task, we develop lexicon-based approaches with some modifications. We utilize training and test datasets.

4.1 Overview of the Approach

Initially, tweets are downloaded using the Twitter API. Hashtags are extracted and manually annotated. Tweets containing at least one hashtag of a

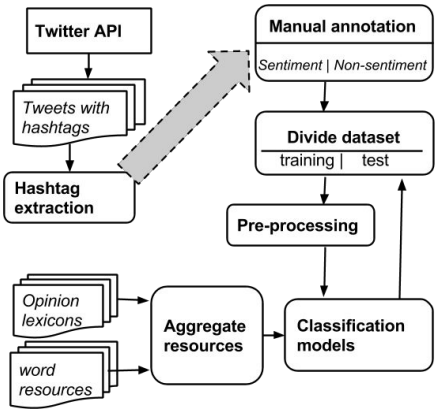

Figure 1: Overview of our approach

particular type are grouped. Then each group is divided into training and test sets. Pre-processing tasks are applied to the training hashtags. Then, classification models are developed and applied to the training hashtags. These models use aggregated lists of opinion words obtained from different lexical and word resources. Finally, each model is applied to the test set.

4.2 Pre-processing

Training hashtags are stripped of their hash symbol, "#". Stemming is applied to the extracted hashtags using a Regrexp stemmer from the Natural Language Processing Toolkit (NLTK) (Loper and Bird, 2002). Using this stemmer, we remove the following suffices:"ed", "ition", "er", "ation", "es", "ness", "ing" and "ment".

For each lexicon, we extract all positive and negative words. However, for a few lexicons, 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 the larger resources we focus only on the adjectives because they are sentiment-bearing (Khuc et al., 2012). As a result, for NRC Hashtag Sentiment Lexicon, we use a POS tagger from NLTK to extract the top 500 adjectives for each sentiment class whereas for SentiWordNet, we consider only the adjectives (as indicated in the lexicon) that have scores for positivity or negativity, which are greater than or equal to 0.5.

4.3 Aggregation of subjective Words

Additionally, we include emotional words from three online resources: Steven Hein feeling words (Hein, 2013) which has 4232 words, The Compass DeRose Guide to Emotion Words (DeRose, 2005) which has 682 words, and SentiSense affective lexicon in which we selected all the adjectives and adverbs in the gloss of the synsets that are categorized as adjectives (de Albornoz et al., 2012). We also include a group of manually identified sentiment-bearing Twitter slangs/acronyms (Fisher, 2012; Nichol, 2014), and some common interjections (Beal, 2014). These words are not typically found in the opinion lexicons. Examples include "fab" for "fabulous", and "OMG" for "Oh my God".

Overall, we use a total of 11 resources. We then combine all the unique words from each of the resources. All duplicates are removed. Then, 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. Each aggregated list of words is mutually exclusive. These lists are described below.

1. **(FOW)** (Frequently Occurring Words) list contains the most subjective words. These **542** words have occurred in at least six resources. The threshold of six represents over half of the total number of resources used.

2. **Stems of FOW** contains the stems of all the opinion words in the FOW list. This list contains **522** words.

3. **LDW** (Less Discriminating Words) list consists of opinion words that occur in at least 2 but not exceeding 3 of the 5 larger resources: NRC Hashtag Sentiment, SentiWordNet, General Inquirer, Subjectivity Lexicon and Steven Hein's feeling words. These **1031** words are considered to be the least subjective.

4. **MDW** (More Discriminating Words) list contains words that are strongly subjective. These remaining **7763** words are not FOW or LDW.

5. **Twitter slangs and acronyms** and common interjections, giving a total of **308** words.

4.4 Model Development

We develop two classification models, which use our aggregated lists of subjective words as input.

4.4.1 Model 1

This model uses a binary search algorithm to compare each hashtag with each subjective word. Comparisons are also made between the stem of the hashtag and each subjective word. If a match is found, the search terminates. Otherwise, the search must continue into the second step where substrings of the hashtag are created using two recursive algorithms. The list of substrings contain at least 3 characters and are sorted in descending order of length.

The first algorithm, called *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. The second algorithm, called *remove_left*, removes the leftmost character from the hashtag after each iteration. After employing both algorithms, the pre-processed hashtag "behappy" has 6 unique substrings: "behapp", "behap", "beha", "beh", "ehappy", and "happy". The resulting substrings of the hashtag are compared to the opinion words in FOW, stems of FOW, and MDW lists because these substrings are smaller representations of the hashtag, and thus, we consider only matches to the most subjective words.

If this search is unsuccessful, we then ascertain if the hashtag contains any non-word attribute in the hashtag that suggests the expression of a sentiment. We consider only the presence of exclamation or question marks (Bakliwal et al., 2012) and repeated characters (at least 3).

Table 1 outlines the eight rules for identifying sentiment hashtags. If none of these rules is found to be true, then the hashtag is determined to be sentiment bearing. Otherwise, the hashtag is non-sentiment bearing.

Rules
Hashtag = opinion word
Hashtag = stem (opinion word)
Stem of the hashtag = an opinion word
Stem of the hashtag = stem of FOW
Max(hashtag substring) = an opinion word
Stem (max(hashtag substring)) = stem of FOW
Max(hashtag substring) = stem (opinion word)
Hashtag contains a sentiment feature

Table 1: Rules for identifying sentiment hashtags

4.4.2 Model 2

In this model, we apply a bootstrapping technique. First, we obtain seed words by using our aggregated lists to find hashtags that are subjective words (including those hashtags that have substrings that are at least 95% in length to a subjective word in our aggregated lists). We then use these seed hashtagged words in order to learn additional hashtags. We employ these four rules: the seed word must be a substring of the hashtag (minimum threshold of 35%) or the stem of the hashtag, and the stem of the seed word must be a substring of the hashtag (minimum threshold of 35%) or the stem of the hashtag. If any of these rules apply, then the hashtag is considered be sentiment bearing. Otherwise, the hashtag is considered to be non-sentiment bearing.

5 Experiment and Results

In this section, we present our experiments that are carried out to evaluate our approach.

5.1 Dataset

Tweets were collected from June 11 to July 2, 2014 during the FIFA World Cup 2014. Tweets were scraped from Twitter using search terms related to the football matches that were being played, in order to capture the opinions of fans. The search terms used were not hashtags as our intention was to acquire a wide variety of hashtags that were created by users. We collected a total of 635,553 tweets containing at least one hashtag. After removing all retweets, hashtags were extracted from the dataset and manually classified. For each hashtag type, we selected the tweets containing at least one hashtag of the respective type. Then, we divided this dataset of tweets equally into training and test sets. Table 2 shows the total number of hashtags in the training and test sets, for each type of hashtag.

Hashtag type	**Training**	**Test**	**Total**
Sentiment	1,368	1,376	2,744
Non-Sentiment	3,070	3,142	6,212

Table 2: Training and test set for each hashtag type

5.2 Experimental setup

In our experiment, we compare the hashtags extracted in the test sets with those from the training set. If the test hashtag is found in the list of

training hashtags, the same class label is assigned. Otherwise, we perform similarity testing.

In similarity testing, we compare the stems of the hashtags in the training and test sets. If a match cannot be determined, we ascertain if the test hashtag contains a substring that is at least 95% of the length of one of the training hashtags. If a suitable match is found, the same class label is assigned to the test hashtag. Finally, we compare the predicted class label assigned by the model to that of actual label of the hashtag assigned during manual annotation.

5.3 Results and Discussion

Tables 3 and 4 shows the accuracy (A), precision (P), recall (R), and f-measure (F), metrics (in percent) for Model 1 and 2, respectively. It can be

Hashtag type	**A**	**P**	**R**	**F**
Sentiment	83.7	86.3	81.2	83.7
Non-sentiment	**84.1**	**94.5**	**85.0**	**89.5**

Table 3: Classification results for Model 1

Hashtag type	**A**	**P**	**R**	**F**
Sentiment	78.7	84.2	72.1	77.7
Non-sentiment	**82.6**	**91.9**	**85.8**	**88.8**

Table 4: Classification results for Model 2

observed from both tables 3 and 4 that our models achieved higher percentages for all four evaluation measures in identifying non-sentiment hashtags than sentiment hashtags. Therefore, we can conclude that it is easier to identify non-sentiment hashtags than sentiment hashtags by combining existing lexical resources. This may be due to the fact that sentiment hashtags contain subjective expressions that are not found in lexical resources. Examples of misclassified sentiment hashtags include "#rootingforyou", "#bringbackourplayers", "needasoccerplayer", and "#historyinthemaking".

In order to determine the effectiveness of using combined resources, for each model, we substituted the combined resources for a single resource. Figure 2 shows the average accuracy and f-measure scores for using single and combined resources for Model 1 and 2, respectively.

It can be observed in Figures 2 and 3 that by using combined lexical, emotion and word resources, Model 1 and 2 achieve the highest average accuracy and f-measure in identifying sentiment and non-sentiment hashtags when compared to using a single resource. Furthermore, this is more acute for Model 1 than Model 2.

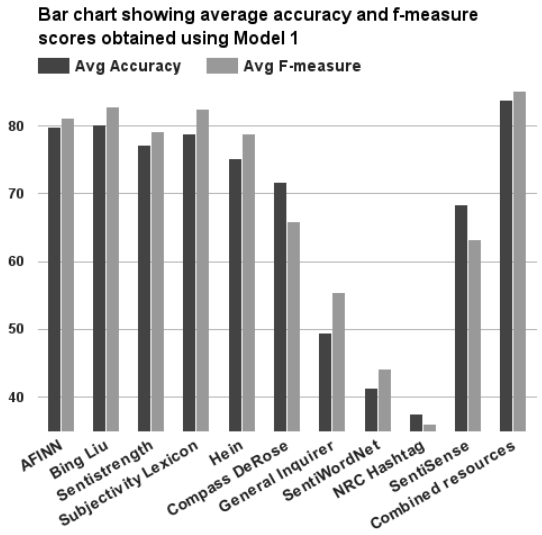

Figure 2: Performance of Model 1

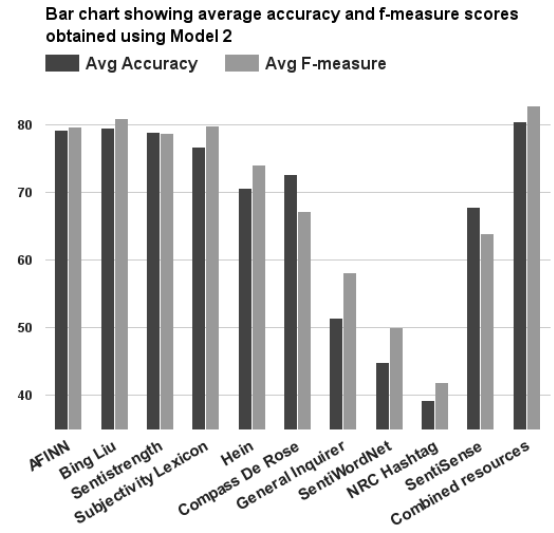

Figure 3: Performance of Model 2

6 Conclusion

In this paper, we applied a lexicon-based approach to identify hashtag types. Our experimental results show that by using combined lexical, emotion and word resources, we can identify non-sentiment hashtags more accurately and precisely than sentiment hashtags. Furthermore, using these combined resources is more effective than using a single resource in identifying hashtag types. In the future, we plan to develop hashtag segmentation algorithms to improve this classification task.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. 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.

Akshat Bakliwal, Piyush Arora, Senthil Madhappan, Nikhil Kapre, Mukesh Singh, and Vasudeva Varma. 2012. 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.

Vangie Beal. 2014. Twitter dictionary: A guide to understanding twitter lingo, August. Retrieved on September 20, 2014.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. 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.

Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervs. 2012. Sentisense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Steven J. DeRose. 2005. The compass derose guide to emotion words. Retrieved on September 20, 2014.

Tia Fisher. 2012. Top twitter abbreviations you need to know. Retrieved on September 21, 2014.

Steven Hein. 2013. Feeling words/emotion words. Retrieved on September 20, 2014.

Minqing Hu and Bing Liu. 2004. 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.

Jahwan Khan. 2015. How to use hashtags for optimal social media engagement. Online, March. Retrieved on June 26, 2015.

Vinh Ngoc Khuc, Chaitanya Shivade, Rajiv Ramnath, and Jay Ramanathan. 2012. Towards building large-scale distributed systems for twitter sentiment analysis. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 459–464, Trento, Italy.

Edward Loper and Steven Bird. 2002. 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.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task*, pages 246–255, Montréal, Canada.

Mark Nichol. 2014. 100 mostly small but expressive interjections. Retrieved on January 30, 2015.

Finn Årup Nielsen. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, pages 93–98, Heraklion, Crete, Greece.

Ashequl Qadir and Ellen Riloff. 2013. Bootstrapped learning of emotion hashtags #hashtags4you. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–11, Atlanta, Georgia.

Ashequl Qadir and Ellen Riloff. 2014. Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1203–1209, Doha, Qatar.

Credell Simeon and Robert Hilderman. 2015. Evaluating the effectiveness of hashtags as predictors of the sentiment of tweets. In *Proceedings of the 18th International Conference on Discovery Science, (DS-2015)*, Lecture Notes in Computer Science, Banff, Alberta. To appear.

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

Mike Thelwall. 2012. Heart and soul: Sentiment strength detection in the social web with sentistrength. Retreived on April 2nd 2014.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. 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.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. 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.