Dependency-based Topic-Oriented Sentiment Analysis in Microposts

Prasadith Buddhitha and Diana Inkpen

University of Ottawa
School of Information Technology and Engineering
Ottawa, ON, K1N6N5, Canada
pkiri056@uottawa.ca and Diana.Inkpen@uottawa.ca

Abstract

In this paper, we present a method that exploits syntactic dependencies for topic-oriented sentiment analysis in microposts. The proposed solution is based on supervised text classification (decision trees in particular) and freely-available polarity lexicons in order to identify the relevant dependencies in each sentence by detecting the correct attachment points for the polarity words. Our experiments are based on the data from the Semantic Evaluation Exercise 2015 (SemEval-2015), task 10, subtask C. The dependency parser that we used is adapted to this kind of text. Our classifier that combines both topic- and sentence-level features obtained very good results (comparable to the best official SemEval-2015 results).

1 Introduction

Identifying opinionated factual information has become widely popular during the current years. The growth of social media has enhanced the amount of information being shared among groups of people as and when it is being generated due to various activities. With the availability of various less-complex and economical telecommunication media, human expression has become frequently embedded within the information being transmitted. Such freely available information has attracted many stakeholders with a wide range of interdisciplinary interests. Microblogs have become one of the most popular and widely-used sources of information, where users share real-time information on many topics.

Twitter has become one of the most popular microblogging platforms in recent years. According to Twitter (2015), 500 million tweets are being posted per day with 302 million monthly active users. As more and more interest has emerged in identifying the key information contained within the messages, greater difficulties are being introduced due to the informal nature of the message representation. With the limitation of 140 characters, the informal nature of the messages has introduced slang, new words and terminology, URLs, creative spelling, misspelling, punctuations and abbreviations such as #hashtag and "re-tweet" (RT).

With the representation of valuable information about one or more interests enriched with user perception and the sheer amount of volume has challenged the researchers in Natural Language Processing and Machine Learning to generate mechanisms to extract the valuable information, which could be beneficial for the interested parties from different domains, such as marketing, finance, and politics. Identification of the perception, which could also be termed as opinion mining or sentiment analysis has resulted in many researches based on supervised and unsupervised learning methods.

The widely-spread enthusiasm in the Twitter sentiment analysis is supported by various research-based events such as the Semantic Evaluation Exercise (SemEval). The research we present in this paper is based on the SemEval 2015 task 10, dedicated to Sentiment Analysis in Twitter. The task is subdivided into four subtasks emphasizing different levels such as expression, message, topic and trend (SemEval-2015).

We focus on "topic-based message polarity classification"; that is, given a message and a topic, we classify whether the message is of positive, negative, or neutral sentiment towards the given topic (SemEval-2015). The task will be approached through the use of sentiment lexicons at both topic and sentence level. Our solution

will use several freely available general-purpose sentiment lexicons and tweet-specific sentiment lexicons, the latter provided by the National Research Council (NRC) of Canada.

The following paragraphs will briefly define the different terminologies being used in the rest of this paper.

Tokenization: Text normalization is a key part in Natural Language Processing (NLP), which is commonly being used in many NLP tasks including the proposed solution. Tokenization can be considered as one of the initial and key functions in text normalization where a given text is divided in to a sequence of characters, which can be treated as a group. The character sequence can be treated as a word or any other token such as a punctuation mark or a number (Jurafsky & Martin, 2008).

Sentiment analysis: As described in Scherer's typology of affective states, sentiment analysis can be defined as detecting attitudes (Scherer, 1984). The polarity can be identified as a type of attitude, which could be categorized into one of the states such as positive, negative or neutral, as well as being assigned with a weighted value indicating the strength within the assigned category (Manning & Jurafsky, 2015).

N-grams: N-grams can be broadly defined as a contiguous sequence of words (Jurafsky & Martin, 2008). The N-grams can be represented as N-tokens, where the tokens could be words, letters, etc. Depending on the number of tokens, N-gram models can be termed as unigrams (N-gram with the size of one), 2-gram (bigram), 3-gram (trigram), four-gram or five-gram, which can be considered as the most commonly-used in statistical language models. The number of items within the language model can be based on the processing task. Our proposed solution mainly considers unigrams and bigrams.

Decision Trees: Decision trees can be explained in the most abstract form as if-then-else statements arranged in a tree. The most informative features extracted from the training data are according to their information gain (Quinlan, 1986). They have the advantage that the model leant is interpretable; the user can inspect the tree in order to understand why a certain decision was made. Decision trees do not always get the best results compared to other machine learning algorithms (but they happened to work very well for our particular task). The key step in making decision trees effective will be the selection of suitable features for our task. In our solution, the selected features are based on the polarity words from the sentence that are in dependency relations to the targeted topic.

2 Related Work

There has been a large volume of research on sentiment analysis. It started with identifying subjective and objective statements. Subjective statement can further be classified into positive, negative, or neutral, possibility with intensity level for the first two. Many researches have been done on opinion mining and sentiment analysis for customer reviews (Pang & Lee, 2008) and, more recently, on Twitter messages (Jansen, Zhang, Sobel, & Chowdury, 2009; Kouloumpis, Wilson, & Moore, 2011; Pak & Paroubek, 2010; Bifet, Holmes, Pfahringer, & Gavaldà, 2011).

Over the years many techniques have been adopted by researchers on Twitter sentiment analysis, such as lexical approaches and machine learning approaches (Fernandez, Gutierrez, G'omez, & Martinez-Barco, 2014) (Bin Wasi, Neyaz, Bouamor, & Mohit, 2014). Lexicon-based systems focused on creating repositories of words labeled with polarity values, possibly calculated based on the association of these words and with other words with known polarities (Fernandez et al., 2014). In addition, well-performing hybrid systems have also been proposed for Twitter sentiment analysis by combining hierarchical models based on lexicons and language modeling approaches (Balage Filho, Avanco, Pardo, & Volpe Nunes, 2014).

The large impact of using polarity lexicons in supervised learning can also be seen in the top seven-ranked participants in the SemEval-2015, task 10, subtask C. According to Boag, Potash, & Rumshisky (2015); Plotnikova et al. (2015); Townsend et al. (2015); Zhang, Wu, & Lan (2015) put emphasis on publicly available lexicons such as the NRC Hashtag Sentiment Lexicon, the Sentiment 140 Lexicon, the NRC Emotion Lexicon and SentiWordNet for feature engineering. In addition to lexicon features, many of the top scored systems used linguistic and Twitter-specific features. These systems have mainly used supervised machine learning implemented through classifiers such as Support Vector Machine (SVM) and logistic regression to obtain the best results. It is interesting to note that Townsend et al. (2015), ranked sixth for subtask C, have used the Stanford parser configured for short documents with the use of a

caseless parsing model. The authors have argued that TweeboParser (Kong et al., 2014), which is explicitly created for parsing Twitter messages, lacks in dependency type information due to the use of a simpler annotation scheme rather than using an annotation scheme like Penn Treebank. Kong et al. (2014) have argued that Penn Treebank annotations produce low accuracies specifically with informal text such as tweets and it is more suited for structured data, and due to this reason they have used a syntactically-annotated corpus of tweets (TWEEBANK). Despite these claims, the TweeboParser has achieved an accuracy of over 80% on unlabelled attachments. The parser has contributed nearly 10% accuracy increase in our proposed solution through topic-level feature extraction, which has accumulated towards a comparable Macro F1-measure of 0.5310 in contrast to a lower Macro F1 measure of 0.2279 obtained by Townsend et al. (2015) using the reconfigured Stanford parser.

As many effective sentiment analysis solutions are based on machine learning and lexicon-based techniques (Balage Filho et al., 2014), our proposed solution will also be focused on supervised machine learning that use features computed by using freely available lexicons, while focusing on general and Twitter-specific language constructs.

Many of the proposed solutions in sentiment analysis have used key natural language processing techniques such as tokenizing, part-of-speech tagging, and bag-of-words representation for preliminary preparation tasks (Bin Wasi et al., 2014; Mohammad & Turney, 2013; Kiritchenko, Zhu, & Mohammad, 2014). Due to the informal nature of the Twitter messages, text-preprocessing techniques have to be given special consideration. Bin Wasi et al. (2014), Mohammad & Turney (2013) and Kiritchenko et al. (2014) used the Carnegie Mellon University (CMU) ARK tools for tasks such as tokenizing and part-of-speech tagging, which handles text with Twitter-specific characteristics such as identifying special characters and tokens according to Twitter-specific requirements (Owoputi, O'Connor, Dyer, Gimpel, Schneider, & Smith, 2013). In addition to the CMU ARK tokenizer, our proposed solution uses the TweeboParser for Twitter text dependency parsing, which allows us to identify the syntactic structure of the Twitter messages.

It could be argued that supervised or semi-supervised machine learning techniques provide higher accuracy levels compared to unsupervised machine learning techniques and also the consideration must be given to the specific domain which the task is implemented on (Villena-Roman, Garcia-Morera, & Gonzalez-Cristobal, 2014). This is why we build a supervised classifier based on the SemEval training data, and we are planning to extend it in future work with a large amount of unlabeled Twitter data.

Many systems in the past gave little consideration to hashtags, but this has changed recently, as their impact on the sentiment value of a message was demonstrated. Research has been conducted by using hashtags as seeds representing positive and negative sentiment (Kouloumpis et al., 2011) and also by creating hashtag sentiment lexicons from a hashtag-based emotion corpus (Mohammad & Kiritchenko, 2015). The same lexicon created by Mohammad & Kiritchenko (2015) is being used by our proposed classifier to identify hashtags associated to opinions and emotions; we add a stronger emphasis on the hashtag representation.

According to Zhao, Lan, & Zhu (2014), emoticons are also considered to be providing considerable sentiment value towards the overall sentiment of a sentence. Emoticons were identified using different mechanisms such as through the use of Christopher Potts' tokenizing script (Mohammad, Kiritchenko, & Zhu, 2013). Our proposed solution has adopted the MaxDiff Twitter sentiment lexicon to identify both the emoticons and their associated sentiment values (Kiritchenko et al., 2014), as it will be described in section 4.2.

Many proposed solutions normalize the informal messages in order to assist in sentiment identification (Zhao et al., 2014; Bin Wasi et al., 2014). We do not need to do this, because the lexicons we used contain many such Twitter-specific terms and their associated sentiment values (Mohammad et al., 2013; Kiritchenko at al., 2014).

3 Data

The dataset is obtained from the SemEval-2015 Task 10 for subtask C[1]. The dataset constitute of trial and training data. The training data includes the Twitter ID, the target topic and the polarity towards the topic (SemEval-2015). Due to privacy reasons, the relevant Twitter messages were

[1] We did not participate in the task, we downloaded the data after the evaluation campaigned

not included and a separate script has being provided in order to download the messages. After executing the script, the message "Not Available" is being printed if the relevant tweet is no longer available.

Our final dataset contains 391 Twitter messages, out from 489 given Twitter IDs for the task, where 96 IDs were removed due to unavailability of the Twitter messages, one record due to a mismatched ID and one record because it was a duplicate ID. The original dataset included around 44 topics and approximately ten tweets per topic (SemEval-2015). From the extracted 391 tweets, 110 tweets were labeled with positive topic polarity, 44 as negative, 235 as neutral and 2 were off-topic. According to Rosenthal et al. (2015), having access to less training tweets does not have a substantial impact on the results being generated, because the task participants that used less training data have produced higher results.

In order to make the dataset more relevant and accurate, both URLs and usernames were normalized, where the URLs are renamed as http://someurl and the usernames as @someuser. The tweets were also tokenized using the tokenizing tool provided by Carnegie Mellon University (CMU), known as Tweet NLP.

The Twitter messages in our dataset were composed of only one sentence (and one target topic in the message), this is why in this paper, the terms "sentence-level" and "message-level" are used interchangeably. This is due to the short nature of the tweets (also they are rarely fully-grammatical sentences due to the informal communication style). In rare case, when a tweet might contain more than one sentence, for future test data, our method will use only the sentence(s) that contain the topic(s) of interest.

4 Experiments

Our experiments had the goal of building a supervised classifier that can decide whether the message is positive, negative or neutral towards the given topic.

The experiments were conducted in two parts where features were extracted at sentence level and topic level, using different lexicons. The following sections will describe our features and the tools that we used to extract them, mainly the dependency parser and the lexicons.

4.1 Dependency Parsing

The dependency parser for English tweets, TweeboParser from CMU, was used to generate the syntactic structure of each tweet. Given an input, a single tweet per line, an output of the tokenized tweet is generated with their associated part-of-speech tags and syntactic dependencies. The generated prediction file is structured according to the "CoNLL" format representing different columns such as, token position (ID), word form (FORM), coarse grained part-of-speech tag (CPOSTAG), fine grained part-of-speech tag (POSTAG), most importantly the head of the current token (HEAD) indicating the dependencies and the type of dependency relation (DEPREL) (Buchholz, 2006). The generated syntactic structure for the following tweet:
"They say you are what you eat, but it's Friday and I don't care! #TGIF (@ Ogalo Crows Nest) http://t.co/l3uLuKGk"
is presented in Table 1. For this example, there are several conjunctions (CONJ), and one multi-word expression (MWE) is identified. Some other dependency relations were missed in this case, due to the imperfect training of the parser on small amounts of Twitter data.

This example tweet is from our dataset, and according to the annotations provided by the SemeEval task, the target topic is "Crows Nest", the general message polarity is "positive", and the polarity towards the given topic is "neutral".

4.2 Feature Extraction

Feature extraction was conducted at sentence level and at topic level. Feature extraction was mainly based on sentiment lexicons. NRC Canada has provided several tweet-specific sentiment lexicons, which were used in capturing many Twitter specific content displaying different levels of polarity such as positive, negative and neutral, and also accompanied with a finite set of values representing evaluative intensity towards specific polarity categories (Kiritchenko et al., 2014). Mentioned below are the different lexicons being used at both sentence and topic levels.

ID	FORM	CPOS TAG	POS TAG	HEAD	DEPREL
1	They	O	O	2	_
2	Say	V	V	9	CONJ
3	You	O	O	4	_
4	Are	V	V	2	_
5	What	O	O	7	_
6	You	O	O	7	_
7	Eat	V	V	4	_
8	,	,	,	-1	_
9	But	&	&	0	_
10	it's	L	L	9	CONJ
11	Friday	^	^	10	_
12	And	&	&	0	_
13	I	O	O	14	_
14	don't	V	V	12	CONJ
15	Care	V	V	14	_
16	!	,	,	-1	_
17	#TGIF	#	#	-1	_
18	(@	P	P	0	_
19	Ogalo	^	^	21	MWE
20	Crows	^	^	21	MWE
21	Nest	^	^	18	_
22	)	,	,	-1	_
23	http://t.co/l3uLuKGk	U	U	-1	_

Table 1. TweeboParser output for a tweet.

The *NRC hashtag emotion lexicon* consists in a list of words and their association with eight emotions: anger, fear, anticipation, trust, surprise, sadness, joy and disgust. The association between the tweets and the emotions were calculated through the identification of emotion-word hashtags in tweets (Mohammad et al., 2013). The file is formatted according to category (e.g. anger, fear, etc.), the target word, and the associated score. The relevant score indicates the strength of the association between the category and the target word (Mohammad et al., 2013). Higher scores indicate stronger associations between the category and the target word (Mohammad et al., 2015).

The *NRC word-emotion association lexicon* contains a list of words and their association with eight emotions, anger, fear, anticipation, trust, surprise, sadness, joy and disgust, and also the polarity towards the relevant word represented either as positive or negative (Mohammad et al., 2013). The lexicon is structured according to the target word, the emotion category and the association value indicating to which category the word belongs. The value 1 indicates that it belongs to the relevant category; the value is 0 if it does not (Mohammad et al., 2013).

The *MaxDiff Twitter sentiment lexicon* represents unigrams with associative strength towards positive sentiment. The data was obtained by manual annotation through Amazon Mechanical Turk (Kiritchenko et al., 2014). Each entry of the lexicon consists of the term and its relevant associative values ranging from -1 indicating the most negative score and +1 indicating the most positive score (Mohammad et al., 2013).

Sentiment140 lexicon is a collection of words with the associated positive and negative sentiment (Mohammad et al., 2013). The lexicon is divided into unigrams and bigrams, where each entry contains the term, the sentiment score and the number of times the term appeared with a positive emoticon and the number of times the term appeared with a negative emoticon. The sentiment score is calculated using the pointwise mutual information (PMI), by subtracting the associated score of the term with negative emoticons from the associated score with positive emoticons (Mohammad et al., 2013).

SentiWordNet 3.0 was designed for assisting in opinion mining and sentiment classification in general (not for Twitter messages). SentiWordNet is a result of annotating WordNet synonym entries according to their polarity weighting (Sebastiani & Esuli, 2010). The scores given for positive, negative and neutral classes range between zero and one, and the summation of all three scores is 1. SentiWordNet 3.0 is based on WordNet version 3.0 and the entries include POS and ID columns identifying the associated WordNet synonym set.

4.3 Sentence level feature extraction

Sentence-level feature extraction is conducted mainly using the above-mentioned lexicons. Hashtags in a tokenized Twitter message were looked up in the hashtag emotion lexicon, and the scores were aggregated according to the associated values for each category of emotion. If the given hashtags are not being associated with any of the emotion classes, a value of zero is being returned for the sentence for the specific emotion class.

As an additional attribute, the aggregated emotion values were compared to the maximum value, which is being assigned as the representative nominal class for the given sentence.

In order to compute the features based on the word emotion lexicon, the tokenized Twitter message was matched against the lexicon and the associated values were aggregated according to each individual emotion class in order to represent the sentence

The MaxDiff Twitter sentiment lexicon is used to identify the aggregated scores for a sentence with the associated values given for unigrams. As the values represent positive sentiment towards a given word calculated using the MaxDiff method of annotation, positive and negative value aggregation has resulted in a representation of a sentiment value for the given tweet.

Also, SentiWordNet is used to obtain an aggregated value for the sentence by matching words between the tokenized tweet and the SentiWordNet synonym sets. In addition to the sentence level SentiWordNet score, the given topic in a message is also being evaluated against the synonym set to identify if it carries a sentiment value.

Tokenized Twitter bigrams are also being used to identify related bigram lexical entries against the "Sentiment140" lexicon. In total, at message level, the classifier was trained on nine features using the hashtag emotion lexicon, ten features using the word-emotion association lexicon, and one feature each using the MaxDiff Twitter sentiment lexicon and SentiWordNet. Also the Sentiment140 lexicon for unigrams and bigrams was used in identifying one feature each at message level.

4.4 Topic-level feature extraction

Topic-level feature extraction is implemented similar to sentence-level feature extraction using the above-mentioned lexicons. The key motivation behind the identification of the dependent words is the nature of the task, where it is required to identify the sentence polarity towards a given topic. It is noted that the sentence level polarity and the sentence polarity towards a given topic can be different, as the topic might or might not contribute towards the overall polarity of the sentence. Dependency parsing is being used mainly to identify the sentiment contribution made by the dependent tokens towards the topic, as all the tokens within the sentence might not contribute equally towards the sentiment of a sentence. In contrast to the feature extraction based on the associated tokens towards the left and right of the specific topic (Kiritchenko et al., 2014), the dependency token identification can be intuitively considered as an effective methodology due to the following reasons: the neighbouring tokens might emphasize less the sentiment value; and, most importantly, the token selection can be limited based on their dependency relation to the topic.

The output obtained from the TweeboParser is analyzed to identify both tokens being dependent on the topic and the relevant dependencies that the topic has with the rest of the tokens within a given sentence. The multiword topics are considered as units and the dependencies towards and from them are identified. Extracted topic dependencies are evaluated using the given lexicons to identify different attributes, as mentioned above under different lexicon features. The features are identified against both unigrams and bigrams according to the given lexicon.

In total, at topic level the classifier was trained on nine features using the hashtag emotion lexicon, ten features using the word-emotion association lexicon, two features using the SentiWordNet and one feature using the MaxDiff Twitter sentiment lexicon. In addition, the Sentiment140 lexicon with unigrams and bigrams was used to identify one feature each at topic level.

In summary, a total of 47 features were used to train the classifier (23 at message level and 24 at topic level) and considerable improvement was obtained by using both sentence- and topic-level features, as it will be described in the next section.

5 Results

The evaluation measure that we report is the one used in the SemEval task: the macro average F1 measure for the positive class and for the negative class (excluding the neutral class). The key reason that could be identified as the motivation behind the use of this macro F1-measure is the unequal distribution of the classes (the neutral class being dominant in the dataset).

As the first step in evaluation the most efficient and effective machine-learning algorithm to be used as the main classifier was identified as decision trees, compared with the results[2] obtained for different classifiers such as Support Vector Machines (SVM) and Naïve Bayes. Decision trees resulted[3] with the highest macro average F1 measure for both positive and the negative classes, given all the feature vectors.

[2] Comparing the weighted average F1 measure, the results obtained using a t-test with both sentence- and topic- level features for decision trees (0.64) was noticeably higher than SVM- (0.60) and statistically significant than Naïve Bayes-algorithm (0.44).

[3] Decision trees macro average F1 measure (0.48) was substantially higher than both SVM (0.39) and Naïve Bayes (0.35) macro average F1 measure.

To understand the impact of different features identified through lexicons and the impact sentence- and topic-level features has on the overall classifier performance, we separately ran the decision trees algorithm on sentence- and topic-level features. Table 2 illustrates the impact each lexicon has on the classification results, which could be identified by removing individual lexicons (one or more features) at sentence-level, and then at topic-level and by comparing with the results when using all the features.

Features	Macro F1-measure	
	Sentence level	Topic level
All	0.4435	0.3500
Remove Hashtag emotion lexicon	0.4925	0.3000
Remove MaxDiff Twitter sentiment lexicon	0.4435	0.3615
Remove Word-emotion association lexicon	0.4435	0.3500
Remove Sentiment140 lexicon (unigrams)	0.4270	0.0870
Remove Sentiment140 lexicon (bigrams)	0.4035	0.1770
Remove SentiWordNet	0.2115	0.2685

Table 2. Classification results based on different lexicons illustrated separately on sentence- and topic-level, by removing one lexicon at a time.

By analyzing the Table 2 results (compared with the baseline accuracies) as well as through attribute subset evaluation and also by calculating the information gain[4] with respect to the class on separate features at sentence- and topic-level, we could identify that SentiWordNet and Sentiment140 lexicon features have more influence on the classifier performance followed by Word-emotion, MaxDiff and Hashtag emotion lexicons.

By implementing different combinations of features, both at sentence- and topic-level, we showed that the most influential features were extracted using the following lexicons: SentiWordNet, Sentiment140 lexicon and NRC emotion lexicon.

Table 3 summarizes the results obtained for different combinations of features, at both sentence and topic level. The first line includes all features at both levels. The second line removes all the sentence-level features and keeps only topic-level features in the first column of results and removes all the topic-level features but keep the sentence level features in the second columns of results (the same as the first line of results in Table 2). Then the next lines remove one or more lexicons at a time from each level, and in the last three lines from both levels.

Features (Lexicons)	Macro F1-measure	
	Sentence level	Topic level
Include all features	0.4845	
Remove all features at one level but keep them for the other level	0.3500	0.4435
Sentiment140 lexicon (bigrams)	0.4680	0.4805
Sentiment140 lexicon (unigrams)	0.4730	0.4945
SentiWordNet	0.4745	0.4825
MaxDiff Twitter sentiment lexicon	0.4845	0.4995
Word-emotion association lexicon	0.4845	0.4845
Hashtag emotion	0.5140	0.4745
Hashtag + Word-emotion	0.5140	0.4745
Hashtag + Word-emotion + MaxDiff	0.5165	-
Hashtag + Word-emotion + MaxDiff + Sentiment140 lexicon (unigrams) (topic)	0.5230	
Hashtag + Word-emotion + MaxDiff (sentence) + MaxDiff (topic)	0.5275	
Hashtag + MaxDiff (sentence) + MaxDiff (topic)	0.5310	

Table 3. Comparison of the classification results generated using sentence- and topic-level features together, while removing subsets of features at sentence-level, at topic-level or at both levels.

[4] Information gain and attribute subset evaluation were not solely considered due to macro average F1 measure where it only considers the positive and negative polarities.

	Team	Twitter 2015
1	TwitterHawk	0.5051
2	KLUEless	0.4548
3	Whu_Nlp	0.4070
4	whu-iss	0.2562
5	ECNU	0.2538
6	WarwickDCS	0.2279
7	UMDuluth-CS8761	0.1899

Table 4. Official SemEval-2015 task 10 subtask C results.

6 Discussion

The results obtained were compared against the official results of the SemEval 2015 task 10 subtasks C. The top seven results from SemEval are presented in Table 4.

Comparatively, the proposed classifier using sentence and topic level features based on lexicons has obtained a macro-F1 score higher than the best result from Table 4. The good results that we obtained were mainly due to the use of the publicly available lexicons and the rich set of lexicons provided by NRC Canada through extensive research on sentiment analysis of short informal text. Both sentence and topic level features have contributed to the higher accuracy level while sentence level features can be identified as the main contributor. Use of topic level features identified through topic dependencies has provided a substantial improvement to the overall results by increasing the macro-F1 measure from 0.4435 (using only sentence level features) to 0.5310 (using both sentence- and topic-level features, but with less sentence level features compared to topic level features). We also showed that use of all the emotion features as a single feature with separate nominal classes achieved better results compared to having separate nominal classes for each emotion (sadness, fear, anger, etc.).

The best results of 0.5310 macro F1-score were obtained with the use of a combination of topic-level and sentence-level features. Although the topic-level features' contribution on top of the sentence-level features was small, the macro-F1 score for topic-level features only was 0.35, a good score in itself for this difficult task.

We also note that the impact on the F1-measure from the emotion and NRC MaxDiff lexicons at both sentence and topic level was at a lower range, while the majority of the impact was contributed by the SentiWordNet and Sentiment140 lexicons. It could be identified that the use of lexicon-based features within a classification task resulted in generating an accurate classifier as long as features at both sentence- and topic- level were considered.

7 Conclusions and Future Work

The identification of both sentence level features and topic level dependencies with the use of lexicons designed especially for short informal texts, such as tweets, have made our proposed solution to achieve very good results. It was also identified that introducing more features based on lexicons at both sentence- and topic- level could further increase the accuracy of the classifier.

In future work, in addition to lexicon-based features, factors that have high impact on sentiment such as identification of negation, part-of-speech tagging and tag frequencies could also be considered in order to improve the accuracy of the classifier. Further identification of dependency relations by training the dependency parser with additional dependency relation labels, could also improve the accuracy level of the classifier. We also plan to do more extensive testing, on large amounts of tweets that arrive in real time for various target topics.

References

Balage Filho, P., Avanco, L., Pardo, T., & Volpe Nunes, M. d. (2014). NILC_USP: an improved hybrid system for sentiment analysis in Twitter messages. ACL Special Interest Group on the Lexicon - SIGLEX (p. 0). Dublin: Dublin City University - DCU.

Bifet, A., Holmes, G., Pfahringer, B., & Gavaldà, R. (2011). Detecting Sentiment Change in Twitter Streaming Data. Journal of Machine Learning Research, 5-11.

Bin Wasi, S., Neyaz, R., Bouamor, H., & Mohit, B. (2014). CMUQ@Qatar:Using Rich Lexical Features for Sentiment Analysis on Twitter. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 186-191). Dublin: Association for Computational Linguistics.

Boag, W., Potash, P., & Rumshisky, A. (2015). TwitterHawk : A Feature Bucket Approach to Sentiment Analysis. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), (SemEval), 640–646. Retrieved from http://www.aclweb.org/anthology/S15-2107

Buchholz, S. (2006, 06 14). CoNLL-X Shared Task: Multi-lingual Dependency Parsing. Retrieved 04 18, 2015, from CoNLL-X Shared Task: Multi-lingual Dependency Parsing: http://ilk.uvt.nl/conll/#dataformat

Fernandez, J., Gutierrez, Y., G'omez, M. J., & Martinez-Barco, P. (2014). GPLSI: Supervised Sentiment Analysis in Twitter using Skipgrams. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 294-299). Dublin: Association for Computational Linguistics.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, 60, 2169-2188.

Jurafsky, D., & Martin, J. H. (2008). Speech and Language Processing (2nd Edition). California: Pearson.

Kiritchenko, S., Zhu, X., & Mohammad, S. (2014). Sentiment Analysis of Short Informal Texts. Journal of Artificial Intelligence Research (JAIR), 723-762.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., & Smith, A. N. (2014). A Dependency Parser for Tweets. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1001-1012). Doha: Association for Computational Linguistics.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011 (pp. 538-541). Catalonia: AAAI Press.

Manning, C., & Jurafsky, D. (2015, 03 28). Sentiment Analysis. Retrieved 03 28, 2015, from Natural Language Processing: http://spark-public.s3.amazonaws.com/nlp/slides/sentiment.pdf

Mohammad, S. M., & Kiritchenko, S. (2015). Using Hashtags to Capture Fine Emotion Categories from Tweets. Computational Intelligence, 31, 301-326.

Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013). Atlanta: SemEval-2013.

Mohammad, S., & Turney, P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence, 436-465.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, A. N. (2013). Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 380-390). Atlanta: Association for Computational Linguistics.

Pak, A., & Paroubek, P. (2010). Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. Proceedings of the 5th International Workshop on Semantic Evaluation (pp. 436-439). Stroudsburg: Association for Computational Linguistics.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2, 1-135.

Plotnikova, N., Kohl, M., Volkert, K., Lerner, A., Dykes, N., Ermer, H., & Evert, S. (2015). KLUEless : Polarity Classification and Association. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 1(SemEval), 619–625. Retrieved from http://www.aclweb.org/anthology/S15-2103

Quinlan, J.R. (1986). Induction of Decision Trees. Machine Learning 1(1). Kluwer Academic Publishers.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., & Stoyanov, V. (2015). SemEval-2015 Task 10 : Sentiment Analysis in Twitter. Proceedings of the 9th International Workshop on Semantic Evaluation, (SemEval), 451–463. Retrieved from http://www.aclweb.org/anthology/S15-2078

Sebastiani, F. & Esuli, A. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) (pp. 19-21). Valletta: European Language Resources Association (ELRA).

SemEval 2015. (2015, 01 01). SemEval-2015. Retrieved 03 02, 2015, from Data and Tools: http://alt.qcri.org/semeval2015/task10/index.php?id=data-and-tools
http://alt.qcri.org/semeval2015/task10/

Scherer, K. R. (1984). Emotion as a multicomponent process: A model and some cross-cultural data. Personality & Social Psychology 5, 37-63.

Townsend, R., Tsakalidis, A., Wang, B., Liakata, M., Cristea, A., & Procter, R. (2015). WarwickDCS : From Phrase-Based to Target-Specific Sentiment

Recognition. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), (SemEval), 657–663. Retrieved from http://www.aclweb.org/anthology/S15-2110

Twitter. (2015, 01 01). About. Retrieved 05 17, 2015, from Twitter: https://about.twitter.com/company

Villena-Roman, J., Garcia-Morera, J., & Gonzalez-Cristobal, C. J. (2014). DAEDALUS at SemEval-2014 Task 9: Comparing Approaches for Sentiment Analysis in Twitter. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 218-222). Dublin: Association for Computational Linguistics.

Zhang, Z., Wu, G., & Lan, M. (2015). ECNU : Multi-level Sentiment Analysis on Twitter Using Traditional Linguistic Features and Word Embedding Features. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), (SemEval), 561–567. Retrieved from http://www.aclweb.org/anthology/S15-2094

Zhao, J., Lan, M., & Zhu, T. (2014). ECNU: Expression- and Message-level Sentiment Orientation Classification in Twitter Using Multiple Effective Features. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 259-264). Dublin: Association for Computational Linguistics and Dublin City University.