Indian_Institute_of_Technology-Patna: Sentiment Analysis in Twitter

Vikram Singh, Arif Md. Khan and Asif Ekbal
Indian Institute of Technology Patna
Patna, India
(vikram.mtcs13,arif.mtmc13,asif)@iitp.ac.in

Abstract

This paper is an overview of the system submitted to the SemEval-2014 shared task on sentiment analysis in twitter. For the very first time we participated in both the tasks, *viz contextual polarity disambiguation* and *message polarity classification*. Our approach is supervised in nature and we use sequential minimal optimization classifier. We implement the features for sentiment analysis without using deep domain-specific resources and/or tools. Experiments within the benchmark setup of SemEval-14 shows the F-scores of 77.99%, 75.99%, 76.54 %, 76.43% and 71.43% for LiveJournal2014, SMS2013, Twitter2013, Twitter2014 and Twitter2014Sarcasm, respectively for Subtask A. For Subtask B we obtain the F-scores of 60.39%, 51.96%, 52.58%, 57.25%, 41.33% for five different test sets, respectively.

1 Introduction

In current era microblogging is an efficient way of communication where people can communicate without physical presence of receiver(s). Twitter is the medium where people post real time messages to discuss on the different topics, and express their sentiments. The texts used in twitter are generally informal and unstructured in nature. Tweets and SMS messages are very short in length, usually a sentence or a headline rather than a document. These texts are very informal in nature and contains creative spellings and punctuation symbols. Text also contains lots of misspellings, slang, out-of-vocabulary words, URLs, and genre-specific terminology and abbreviations, e.g., RT for re-Tweet and #hashtags. Such kinds of structures introduce difficulties in building various lexical and syntactic resources and/or tools, which are required for efficient processing of texts. Finding relevant information from these posts poses big challenges to the researchers compared to the traditional text genres such as newswire.

In recent times, there has been a huge interest to mine and understand the opinions and sentiments that people are communicating in social media (Barbosa and Feng, 2010; Bifet et al., ; Pak and Paroubek, 2010; Kouloumpis et al., 2011). There is a tremendous interest in sentiment analysis of Tweets across a variety of domains such as commerce (Jansen et al., 2009), health (Chew and Eysenbach, 2010; Salathe and Khandelwal, 2011) and disaster management (Verma et al., 2011; Mandel et al., 2012). Agarwal et al.(Agarwal et al., 2011) used tree kernel decision tree that made use of the features such as Part-of-Speech (PoS) information, lexicon-based features and several other features. They acquired 11,875 manually annotated Twitter data (Tweets) from a commercial source, and reported an accuracy of 75.39%. Semantics has also been used as the feature to improve the performance of sentiment analysis (Saif et al., 2012). For each extracted entity (e.g. iPhone) from Tweets, they added its semantic concept (e.g. Apple product) as an additional feature. Thereafter they devised a method to measure the correlation of the representative concept with negative/positive sentiment, and applied this approach to predict sentiment for three different Twitter datasets. They showed that semantic features produce better recall and F-score when classifying negative sentiment, and better precision with lower recall and F-score in positive sentiment classification. The benchmark corpus were made available with the SemEval-2013 shared task (Nakov et al., 2013) on

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 341–345,
Dublin, Ireland, August 23-24, 2014.

sentiment analysis in twitter. The datasets used are from the domains of Tweets and SMS messages. The datasets were labelled with contextual phrase-level polarity and overall message-level polarity. Among the 44 submissions, the support vector machine based system proposed in (Mohammad et al., 2013) achieved the highest F-scores of 69.02% for Task A, i.e. the message-level polarity and and 88.93% for Task B, i.e. term-level polarity.

The issues addressed in SemEval-13 are further extended in SemEval-14 shared task [1]. The same two tasks, *viz.* Subtask A and Subtask B denoting *contextual polarity disambiguation* and *message polarity classification*. The goal of Subtask A is to determine, for a given message containing a marked instance of a word or phrase, whether that instance is positive, negative or neutral in that context. Given a message, the task is to classify it with its entirety whether it is positive, negative, or neutral sentiment. For messages that convey both positive and negative sentiments, the stronger one should be chosen. In this paper we report on our submitted systems for both the tasks. Our evaluation for the first task shows the F-scores of 77.99%, 75.99%, 76.54%, 76.43% and 71.43% for LiveJournal2014, SMS2013, Twitter2013, Twitter2014 and Twitter2014Sarcasm, respectively for Subtask A. For Subtask B we obtain the F-scores of 60.39%, 51.96%, 52.58%, 57.25%, 41.33% for five different test sets, respectively.

2 Methods

In this section we describe preprocessing steps, features and our methods for sentiment classification

2.1 Preprocessing of Data

The data has to be pre-processed before being used for actual machine learning training. Each Tweet is processed to extract only those relevant parts that are useful for sentiment classification. For example, stop words are removed; symbols and punctuation markers are filtered out; URLs are replaced by the word URL etc. Each Tweet is then passed through the ARK tagger developed by CMU [2] for tokenization and Part-of-Speech (PoS) tagging.

[1] http://alt.qcri.org/semeval2014/task9/

[2] http://www.ark.cs.cmu.edu/TweetNLP/

2.2 Approach

Our approach is based on supervised machine learning. We explored different models such as naive Bayes, decision tree and support vector machine. Based on the results obtained on the development sets we finally select SVM for both the tasks. We also carried out a number of experiments with the various feature combinations. Once the model is fixed with certain feature combinations, these are finally used for blind evaluation on the test sets for both the tasks. We submit two runs, one for each task. Both of our submissions were constrained in nature, i.e. we did not make use of any additional resources and/or tools to build our systems. We adapt a supervised machine learning algorithm, namely Support Vector Machine (Joachims, 1999; Vapnik, 1995). We use its sequential minimal optimization version for faster training[3]. We use the same set of features for both the tasks. Development sets are used to identify the best feature combinations for both the tasks. Default parameters as implemented in Weka are used for the SVM experiments.

2.3 Features

Like any other classification algorithm, features play an important role for sentiment classification. For the very first time we participated in this kind of task, and therefore had to spend quite long time in conceptualization and implementation of the features. We focused on implementing the features without using any domain-dependent resources and/or tools. Brief descriptions of the features that we use are presented below:

- **Bag-of-words**: Bag-of-words in the expression or in the entire Tweet is used as the feature(s).

- **SentiWordNet feature**: This feature is defined based on the scores assigned to each word of a Tweet using the SentiWordNet[4]. A feature vector of length three is defined. The scores of all words of the phrase or Tweet is summed over and normalized in the scale of 3. We define the following three thresholds: if the score is less than 0.5 then it is treated to be a negative polarity; for the score above 0.8, it is assumed to contain positive sentiment;

[3] http://research.microsoft.com/en-us/um/people/jplatt/smo-book.pdf

[4] sentiwordnet.isti.cnr.it/

and the polarity is considered to be neutral for all the other words. Depending upon the score the corresponding bit of the feature vector is set.

- **Stop_word**: If a Tweet/phrase is having more number of stop words then it most likely contains neutral sentiment. We obtain the stop words from the Wikipedia[5]. We assume that a particular Tweet or phrase most likely bears a neutral sentiment if 20% of its words belong to the category of stop words.

- **All_Cap_Words**: This feature is defined to count the number of capitalized words in an entire Tweet/phrase. More the number of capitalized words, more the chances of being positive or negative sentiment bearing units. While counting, the words preceded by # are not considered. We include this with the assumption that the texts written in capitalized letters express the sentiment strongly.

- **Init_Cap**: The words starting with capitalized letter contribute more towards classifying it.

- **Percent_Cap**: This feature is based on the percentage of capitalized characters in a Tweet/phrase. If this is more than 75%, then most likely it is not of neutral type.

- **Psmiley (+ve Smiley)**: Generally people use smileys to represent their emotions. A smiley present in a Tweet/phrase directly represents its sentiment. A feature is defined that takes the value equal to the number of positive smileys. We make use of the list available at this page[6].

- **Nsmiley (-ve Smiley)**: The value of this feature is set to the number of negative smileys present in the Tweet. This list was also obtained from the web[7].

- **NumberPostive words**: This feature takes the value equal to the number of positive words present in the Tweet/phrase. We search the adjective words present in the Tweet in the SentiWordNet to determine whether it bears positive sentiment.

- **NumberNegative words**: This feature takes the value equal to the number of negative words present in the Tweet/phrase. The words are again looked at the SentiWordNet to determine its polarity.

- **NumberNeutral words**: This feature determines the number of neutral words present in the Tweet or phrase. This information is obtained by looking the adjective words in the SentiWordNet.

- **Repeating_char**: It has been seen that people express strong emotion by typing a character many times in a Tweet. For example, happppppppy, hurrrrrey etc. This feature checks whether the word(s) have at least three consecutive repeated characters.

- **LenTweet**: Length of the Tweet is used as the feature. The value of this feature is set equal to the number of words present in the Tweet/phrase.

- **Numhash**: The value of this feature is set equal to the number of hashtags present in the Tweet.

3 Experiments and Analysis

SemEval-2014 shared task is a continuation of the SemEval-2013 shared task. In 2014 shared task, datasets from different domains were incorporated with a wide range of topics, including a mixture of entities, products and events. Messages relevant to the topics are selected based on the keywords and twitter hashtags.

The training set of Task-A has 4,914 positive, 2,592 negative and 384 neutral class instances. The Task-B training set contains 3,057 positive, 1,200 negative and 3,941 neutral sentiments. Developments sets contain 555, 45 and 365 positive, negative and neutral sentiments, respectively for the first task; and 493, 288 and 632 positive, negative and neutral sentiments, respectively for the second task. The selected test sets were taken mainly from the following domains:

LiveJournal2014: 2000 sentences from LiveJournal blogs;

SMS2013: SMS test from last year-used as a progress test for comparison;

Twitter2013: Twitter test data from last year-used as a progress test for comparison;

Twitter2014: A new Twitter test data of 2000

[5]http://en.wikipedia.org/wiki/Stop_words

[6]http://en.wikipedia.org/wiki/List_of_emoticons

[7]http://en.wikipedia.org/wiki/List_of_emoticons

Model	Avg. F-score
Model-1	75.75
Model-2	72.69
Model-3	75.45
Model-4	75.77

Table 1: Results for Task-A on development set(in %).

Tweets;
Twitter2014Sarcasm: 100 Tweets that are known to contain sarcasm.

We build different models by varying the features as follows:

1. **Model-1**: This model is constructed by considering the features, "Repeating_char", "Numhash", "LenTweet", "Percent_Cap", "Init_Cap", "All_Cap", "Bag-of-words", "Nsmiley", "Psmiley", "SentiWordNet" and "Stop_Words".

2. **Model-2**: This model is constructed by the features "Repeating_char", "Percent_Cap", "Numhash", "LenTweet", "Init_Cap", "All_Cap", "Bag-of-words", "SentiWordNet" and "Stop_Words".

3. **Model-3**: This model is built by considering the features "Repeating_char", "Bag-of-words", "SentiWordNet", "Nsmiley" and "Psmiley".

4. **Model-4**: The model incorporates the features "Repeating_char", "Bag-of-words", "SentiWordNet", "Nsmiley", "Psmiley", "Stop_Words", "Numhash", "LenTweet", "Init_Cap" and "All_Cap".

Results on the development set for Task-A are reported in Table 1 that shows the highest performance in Model-4 with the average F-score value of 75.77%. Thereafter we use this particular feature combination for training SVM, and to report the results. Detailed results are reported in Table 2 for both the tasks. It shows 77.99%, 75.99%, 76.54 %, 76.43% and 71.43% F-scores for the LiveJournal2014, SMS2013, Twitter2013, Twitter2014 and Twitter2014Sarcasm, respectively for Subtask A. For Subtask B we obtain the F-scores of 60.39%, 51.96%, 52.58%, 57.25% and 41.33% for the five different test sets, respectively. A closer investigation to the evaluation results reveals that most of the errors are due to the confusions between positive vs. neutral and negative vs. neutral classes.

Comparisons with the best system(s) submitted in this shared task show that we are behind approximately in the range of 6-14% F-score measures for all the domains for Task-A. Results that we obtain in Task-B need more attention as these fall much shorter compared to the best one (in the the range of 14-18%).

Features used	**Classifier**	**Result(Task A)**	**Result(Task B)**
SWN +ve		LiveJournal2014	LiveJournal2014
SWN -ve		77.99	60.39
SWN neutral		SMS2013	SMS2013
#Stop_Words		75.99	51.96
#All_Cap_Words		Twitter2013	Twitter2013
#Numhash		76.54	52.58
Len_Tweet		Twitter2014	Twitter2014
#Init_Cap_Words		76.43	57.25
%_Init_Cap	SVM	T2014S	T2014S
#+ve_Smiley		71.43	41.33
#-ve_Smiley			
#+ve_Words			
#-ve_Words			
#Neutral_Words			
#Bag_of_words			
Rep_character			

Table 2: Result on test sets for Task-A and Task-B.

4 Conclusion

In this paper we report our works as part of our participation to the SemEval-14 shared task on sentiment analysis for twitter data. Our systems were based on supervised classification, where we fixed SVM to report the test results after conducting several experiments with different classifiers on the development data. We implement a set of features that are applied for both the tasks. Our runs are constrained in nature, i.e. we did not make use of any external resources and/or tools. Our results are quite promising that need further investigation. A closer analysis to the results suggest that most of the errors are due to the confusions between positive vs. neutral and negative vs. neutral classes.

This is our first participation, and within the short period of time we developed the systems with reasonable accuracies. There are still many ways to improve the performance. Possible immediate future extension will be to investigate and implement more features, specific to the task.

References

Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment Analysis of Twitter Data. *ACL Workshop on Languages in Social Media LSM-2011*, pages 30–38.

Luciano Barbosa and Junlan Feng. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, China.

Albert Bifet, Geoffrey Holmes, Bernhard Pfahringer, and Ricard Gavald'a. Detecting Sentiment Change in Twitter Streaming Data. *Journal of Machine Learning Research - Proceedings Track*, 17.

Cynthia Chew and Gunther Eysenbach. 2010. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS ONE*, 5(11):e14118+.

Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter Power: Tweets as Electronic Word of Mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

Thorsten Joachims, 1999. *Making Large Scale SVM Learning Practical*, pages 169–184. MIT Press, Cambridge, MA, USA.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG! In *Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM*, pages 538–541, Barcelona, Spain.

Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis, and Jeremy Rodrigue. 2012. A Demographic Analysis of Online Sentiment during Hurricane Irene. In *Proceedings of the Second Workshop on Language in Social Media, LSM 12*, Stroudsburg.

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 Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, Georgia.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 Task 2: Sentiment Analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA.

Alexander Pak and Patrick Paroubek. 2010. Twitter based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval 10*, pages 436–439, Los Angeles,USA.

Hassan Saif, Yulan He, and Harith Alani. 2012. Semantic Sentiment Analysis of Twitter. In *ISWC'12 Proceedings of the 11th International Conference on the Semantic Web - Volume Part I*, pages 508–524.

Marcel Salathe and Shashank Khandelwal. 2011. Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. *PLoS Computational Biology*, 7(10):e14118+.

Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Sudha Verma, Sarah Vieweg, William Corvey, Leysia Palen, James Martin, Martha Palmer, Aaron Schram, and Kenneth Anderson. 2011. Natural Language Processing to the Rescue? Extracting Situational Awareness Tweets during Mass Emergency. In *Proceedings of the AAAI Conference on Weblogs and Social Media*, Velingrad.