IIT Patna: Supervised Approach for Sentiment Analysis in Twitter

Raja Selvarajan and Asif Ekbal
Department of Computer Science and Engineering
Indian Institute of Technology Patna, India
{raja.cs10,asif}@iitp.ac.in

Abstract

In this paper we report our works for SemEval-2014 Sentiment Analysis in Twitter evaluation challenge. This is the first time we attempt for this task, and our submissions are based on supervised machine learning algorithm. We use Support Vector Machine for both the tasks, *viz. contextual polarity disambiguation* and *message polarity classification*. We identify and implement a small set of features for each the tasks, and did not make use of any external resources and/or tools. The systems are tuned on the development sets and finally blind evaluation is performed on the respective test set, which consists of the datasets of five different domains. Our submission for the first task shows the F-score values of 76.3%, 77.04%, 70.91%, 72.25% and 66.32% for LiveJournal2014, SMS2013, Twitter2013, Twitter2014 and Twitter2014Sarcasm datasets, respectively. The system developed for the second task yields the F-score values of 54.68%, 40.56%, 50.32%, 48.22% and 36.73%, respectively for the five different test datasets.

1 Introduction

During the past few years, the communications in the forms of microblogging and text messaging have emerged and become ubiquitous. Opinions and sentiments about the surrounding worlds are widely expressed through the mediums of Twitter messages (Tweets) and Cell phone messages (SMS). The availability of social content generated on sites such as Twitter creates new opportunities to automatically study public opinion. Dealing with these informal text genres presents new challenges for data mining and language processing techniques beyond those encountered when working with more traditional text genres such as newswire. Tweets and SMS messages are 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 (Nakov et al., 2013). Text also contains lots of misspellings, slang, out-of-vocabulary words, URLs, and genre-specific terminology and abbreviations, e.g., RT for reTweet and #hashtags. The kind of these specific features pose great challenges for building various lexical and syntactic resources and/or tools, which are required for efficient processing of texts. These aspects also introduce complexities to build the state-of-the-art data mining systems. 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., 2011; Pak and Paroubek, 2010; Kouloumpis et al., 2011). Recent studies show the interests 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 (Mandel et al., 2012).

Another aspect of social media data, such as twitter messages, is that they include rich information about the individuals involved in the communication. For e.g., twitter maintains information about who follows whom. ReTweets (reshares of a Tweet) and tags inside of Tweets provide discourse information (Nakov et al., 2013). Efficient modelling of such information is crucial in the sense that it provides a mean to empirically study the social interactions where opinion is conveyed.

Several corpora with detailed opinion and sentiment annotation have been made freely available,

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 324–328,
Dublin, Ireland, August 23-24, 2014.

e.g., the MPQA corpus (Barbosa and Feng, 2005) of newswire text; i-sieve (Kouloumpis et al., 2011) and TASS corpus2 (Villena-Roman et al., 2013) for Twitter sentiment. These resources were either in non-social media or they were small and proprietary. They further focused on message-level sentiment. The SemEval-2013 shared task (Nakov et al., 2013) on sentiment analysis in Twitter releases SemEval Tweet corpus, which contains Tweets and SMS messages with sentiment expressions annotated with contextual phrase-level polarity as well as an overall message-level polarity. Among the 44 submissions, the highest-performing system (Mohammad et al., 2013) made use of Support Vector Machine (SVM) classifier. It obtained the F-scores of 69.02% in the message-level task and 88.93% in the term-level task. Variety of features were implemented based on surface-forms, semantics, and sentiment features. They generated two large wordsentiment association lexicons, one from Tweets with sentiment-word hashtags, and one from Tweets with emoticons. They showed that in message-level task, the lexicon-based features gained 5 F-score points over all the others.

SemEval-14 shared task [1] on sentiment analysis in Twitter is a continuing effort to promote the research in this direction. Similar to the previous year's evaluation campaigns two primary tasks were addressed in this year challenge. The first task (i.e. Subtask A) deals with *contextual polarity disambiguation* and the second task (i.e. Subtask B) was about *message polarity classification*. For Subtask A, for a given message containing a marked instance of a word or phrase, the goal is to determine whether that instance is positive, negative or neutral in that context. In Subtask B, for a given message, the task is to classify whether the message is of 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 submissions as part of our first-time participation in this kind of task (i.e. sentiment classification). We develop the systems based on supervised machine learning algorithm, namely Support Vector Machine (SVM) (Joachims, 1999; Vapnik, 1995). We identify and implement a very small set of features that do not make use of any external resources and/or tools. For each task the system is tuned on the development data, and finally blind evaluation is performed on the test data.

2 Methods

We develop two systems, one for contextual polarity disambiguation and the other for message polarity classification. Each of the systems is based on supervised machine learning algorithm, namely SVM. Support vector machines (Joachims, 1999; Vapnik, 1995) have been shown to be highly effective at traditional text categorization, generally outperforming many other classifiers such as naive Bayes (Joachims, 1999; Vapnik, 1995). They are large-margin, rather than probabilistic, classifiers. For solving the two-class problem, the basic idea behind the training procedure is to find a hyperplane, represented by vector $\vec{w}$, that not only separates the document vectors in one class from those in the other, but for which the separation, or margin, is as large as possible. This search corresponds to a constrained optimization problem; letting c_j in 1,-1 (corresponding to positive and negative classes, respectively) be the correct class of the document d_j, the solution could be written as:
$\vec{w} := \sum_j a_j cj \vec{d_j}, \quad a_j >= 0$
where, the a_j's are obtained by solving a dual optimization problem. Those $\vec{d_j}$ such that a_j is greater than zero are called support vectors, since they are the only document vectors contributing to $\vec{w}$. Classification of test instances consists simply of determining which side of $\vec{w}$'s hyperplane they fall on.

2.1 Preprocessing

We pre-process Tweet to normalize it by replacing all "URLs" to "http://url" and all user-ids to "@usr", and this is performed by the regular expression based simple pattern matching techniques. We remove punctuation markers from the start and end positions of Tweets. For e.g., 'the day is beautiful!' is converted to 'the day is beautiful'. Multiple whitespaces are replaced with single whitespace. Stop-words are removed from each review.

2.2 Features

In this work we use same set of features for both the tasks. Each Tweet is represented as a vector consisting of the following features:

1. **Local contexts**: We extract the unigrams and bigrams from the training and test datasets.

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

A feature is defined that checks the occurrences of these n-grams in a particular Tweet or phrase.

2. **Upper case**: This feature is binary valued with a value set to 1 if all the characters of a phrase or Tweet are capitalized, and 0 otherwise. This indicates that the target message or context contains either positive or negative sentiment.

3. **Elongated words**: The feature checks whether a word contains a character that repeats more than twice. This indicates the presence of a positive sentiment word in the surrounding. This was defined in lines with the one reported in (Mohammad et al., 2013).

4. **Hash tags**: This feature checks the number of hash tags in the Tweet. The value of this feature is set equal to the absolute number of features.

5. **Repeated characters**: This feature checks whether the word(s) have at least three consecutive repeated characters (e.g., happppppppy, hurrrrrey etc.). In such cases, the words are normalized to contain only upto two repeated characters. This helps to capture the words having similar structures.

6. **Negated contexts**: A negated word can affect the polarity of the target word. A negated segment is defined as a sequence of tokens that starts with a negation word (e..g, no, couldn't etc.) and ends with a punctuation marks (e.g.,,,,., :, ;, !, ?). All the words following the negation word are suffixed with NEGATIVE, and the polarity features are also converted with NEGATIVE in line with (Mohammad et al., 2013).

3 Experimental Results and Analysis

The SemEval-2014 shared task datasets are based on SemEval-2013 competition datasets. It covers a range of topics, including a mixture of entities, products and events. Keywords and Twitter hashtags were used to identify messages relevant to the selected topic. The selected test sets were taken from the five different domains. We perform experiment with the python based NLTK toolki[2]. We carried out experiments with the different classifiers. However we report the results of SVM as it produced the highest accuracy with respect to this particular feature set. We use the default parameters of SVM as implemented in this toolkit. We submitted 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.

Class	**precision**	**recall**	**F-score**
Positive	72.02	90.45	80.19
Negative	76.86	53.70	63.23
Neutral	7.69	22.22	3.45
Average	52.19	55.46	53.77

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

Class	**precision**	**recall**	**F-score**
Positive	49.92	63.75	55.99
Negative	42.59	31.94	36.51
Neutral	59.54	53.49	56.35
Average	50.68	49.73	66.39

Table 2: Results on development set for Task-B (in %).

We perform several experiments using the development set. Best results are reported in Table 1 and Table 2 for Task-A and Task-B, respectively. Evaluation shows that for message polarity disambiguation we obtain the average precision, recall and F-score values of 52.19%, 55.46% and 53.77%, respectively. For message polarity classification we obtain the precision, recall and F-Score values of 50.68%, 49.73% and 66.39%, respectively. It is evident from the evaluation that the first task suffers most due to the problems in classifying the tweets having neutral sentiments, whereas the second task faces difficulties in classifying the negative sentiments. We report the confusion matrices in Table 3 and Table 4 for the first

gs\pred	positive	negative	neutral
positive	502	50	3
negative	160	196	9
neutral	35	9	1

Table 3: Confusion matrix for A. Here, gs: Gold standard; pred: Predicted class.

[2] http://www.nltk.org/

gs\pred	positive	negative	neutral
positive	313	43	135
negative	102	92	94
neutral	212	81	337

Table 4: Confusion matrix for B. Here, gs: Gold standard; pred: Predicted class.

and second development sets, respectively. Error analysis suggests that most miss-classifications are because of the less number of neutral instances compared to the positive and negative instances in Task-A. For the Task-B training set, the number of instances of positive and neutral sentiments are very low compared to the negative sentiment.

After tuning the systems on the development sets, we perform blind evaluation on the test datasets. Evaluation results on the test sets are reported in Table 5 for both the tasks. The evaluation is carried out based on the evaluation scripts as provided by the organizers. For message polarity disambiguation we obtain the highest F-score of 77.04% for the SMS data type in Task-A. The system shows the F-scores of 76.03%, 70.91%, 72.25% and 66.35% for LiveJournal2014, Twitter2013, Twitter2014 and Twitter2014sarcasm, respectively. For the second task the system attains the highest F-score value of 54.68% for the LiveJournal2014 dataset. For the other datasets, the system shows the F-scores of 40.56%, 50.32%, 48.22% and 36.73% for the SMS2013, Twitter2013 and Twitter2014Sarcasm, respectively. We followed a simple approach that needs fine-tuning in many places. Currently our systems lack behind the best reported systems by margins of approximately 11-18% F-scores for Task-A, and 19-30% F-scores for 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 developed based on SVM. We use a small set of features, and did not make use of any external resources and/or tools in any of the tasks. Each of the systems is tuned on the development set, and blind evaluation is performed on the test set. Evaluation shows that our system achieves the F-score values in the ranges of 66-76% for Task-A and 36-55% for Task-B.

It is to be noted that this is our first participation, and there are many ways to improve the performance of the models. Firstly we would like to identify more features in order to improve the accuracies. We also plan to come up with proper sets of features for the two task. Efficient feature selection techniques will be implemented to identify the most effective feature set for each of the tasks. We would like to apply evolutionary optimization techniques to optimize the different issues of machine learning algorithm.

Task	Test-set	Average F-score
A	LiveJournal2014	76.03
	SMS2013	77.04
	Twitter2013	70.91
	Twitter2014	72.25
	Twitter2014Sarcasm	66.35
B	LiveJournal2014	54.68
	SMS2013	40.56
	Twitter2013	50.32
	Twitter2014	48.22
	Twitter2014Sarcasm	36.73

Table 5: Results on the test set.

References

Luciano Barbosa and Junlan Feng. 2005. Robust Sentiment Detection on Twitter from Biased and Noisy Data. 39:2-3.

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. 2011. Detecting Sentiment Change in Twitter Streaming Data. *Journal of Machine Learning Research - Proceedings Track*, 17:5–11.

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, June.

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*, Los Angeles,USA.

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.

Julio Villena-Roman, Sara Lana-Serrano, Eugenio Martnez-Camara, Jose Carlos Gonzalez, and Cristobal. 2013. Tass - Workshop on Sentiment Analysis at SEPLN. In *Proceedings of Procesamiento del Lenguaje Natural*, pages 50:37–44.