NILC_USP: An Improved Hybrid System for Sentiment Analysis in Twitter Messages

Pedro P. Balage Filho, Lucas Avanço, Thiago A. S. Pardo, Maria G. V. Nunes
Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
São Carlos - SP, Brazil
{balage, taspardo, gracan}@icmc.usp.br avanco@usp.br

Abstract

This paper describes the NILC_USP system that participated in *SemEval-2014 Task 9: Sentiment Analysis in Twitter*, a re-run of the SemEval 2013 task under the same name. Our system is an improved version of the system that participated in the 2013 task. This system adopts a hybrid classification process that uses three classification approaches: rule-based, lexicon-based and machine learning. We suggest a pipeline architecture that extracts the best characteristics from each classifier. In this work, we want to verify how this hybrid approach would improve with better classifiers. The improved system achieved an F-score of 65.39% in the Twitter message-level subtask for 2013 dataset (+ 9.08% of improvement) and 63.94% for 2014 dataset.

1 Introduction

Twitter is an important platform of social communication. The analysis of the Twitter messages (tweets) offers a new possibility to understand social behavior. Understanding the sentiment contained in such messages showed to be very important to understand user behavior and also to assist market analysis (Java et al., 2007; Kwak et al., 2010).

Sentiment analysis, the area in charge of studying how sentiments and opinions are expressed in texts, is usually associated with text classification tasks. Sentiment classifiers are commonly categorized in two basic approaches: lexicon-based and machine learning approaches (Taboada et al., 2011). A lexicon-based classifier uses a lexicon to provide the polarity, or semantic orientation, of each word or phrase in the text. A machine learning classifier uses features (usually the vocabulary in the texts) obtained from labeled examples to classify the texts according to their polarity.

In this paper, we present a hybrid system for sentiment classification in Twitter messages. Our system combines the lexicon-based and machine learning approaches, as well as uses simple rules to aid in the process. Our system participated in *SemEval-2014 Task 9: Sentiment Analysis in Twitter* (Rosenthal et al., 2014), a re-run for the SemEval 2013 task under the same name (Nakov et al., 2013). The task goal was to determine the sentiment contained in tweets. The task included two sub-tasks: a expression-level classification (Task A) and a message-level classification (Task B). Our system participated only in Task B, where, for a given message, it should classify it as positive, negative, or neutral.

The system presented is an improved version of the system submitted for Semeval 2013. Our previous system had demonstrated that a hybrid approach could achieve good results (F-measure of 56.31%), even if we did not use the state-of-the-art algorithms for each approach (Balage Filho and Pardo, 2013). In this way, this work aims to verify how much this hybrid system could improve in relation to the previous one by including modifications on both lexicon-based and machine learning approaches.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 428–432,
Dublin, Ireland, August 23-24, 2014.

2 Related work

The analysis of Tweets has gained lots of interest recently. One evidence is the expressive number of participants in the *SemEval-2013 Task 2: Sentiment Analysis in Twitter* (Nakov et al., 2013). There were a total of 149 submissions from 44 teams. The best performing system on twitter dataset for task B was reported by Mohammad et al. (2013) with an F-mesaure of 69.02%. Their system used a machine learning approach and a very rich feature set. They showed that the best results were achieved using a built-in positive and negative lexicon and a bag-of-words as features.

Other important system in Semeval 2013 was reported by Malandrakis et al. (2013). The authors presented a hybrid system for twitter sentiment analysis combining two approaches: a hierarchical model based on an affective lexicon and a language modeling approach. The system achieved an F-mesaure of 60.14%.

Most work in sentiment analysis uses either machine learning or lexicon-based techniques. However, few studies have shown promising results with the hybrid approach. König and Brill (2006) proposed a hybrid classifier that uses human reasoning over automatically discovered text patterns to complement machine learning. Prabowo and Thelwall (2009) evaluated the effectiveness of different classifiers. Their study showed that the use of multiple classifiers in a hybrid manner could improve the effectiveness of sentiment analysis.

3 System Architecture

Our system is described as a pipeline solution of four main processes: normalization, rule-based classification, lexicon-based classification and machine learning classification. This is the same architecture presented by our system in 2013.

This pipeline architecture works as a back-off model. In this model, each classifier tries to classify the tweets by using the underlying approach. If a certain degree of confidence is achieved, the classifier will provide the final sentiment class for the message. Otherwise, the next classifier will continue the classification task. The last possibility is the machine learning classifier, responsible to deliver the class when the previous two could not achieve the confidence level. We decided to use this back-off model instead of a voting system, for example, due to the high precision achieved for the rule-based and the lexicon-based classifiers.

The aim of this pipeline architecture is to improve the classification process. In Balage Filho and Pardo (2013), we have shown that this hybrid classification approach may outperform the individual approaches.

In the following subsections, we detail the components of our system. In the next section, we explain how the confidence level was determined.

3.1 Normalization and Rule-based Classifier

The normalization module is responsible for normalizing and tagging the texts. This module performs the following operations:

- Hashtags, urls and mentions are transformed into codes;
- Emoticons are grouped into representative categories (such as 'happy', 'sad', 'laugh') and are converted to particular codes;
- Part-of-speech tagging is performed by using the Ark-twitter NLP (Owoputi et al., 2013)

The rule-based classifier is designed to provide rules that better impact the precision than the recall. In our 2014 system, we decided to use the same rule-based classifier from the 2013 system. The rules in this classifier only verify the presence of emoticons in the text. Empirically, we evidenced that the use of emoticons indicates the actual polarity of the message. In this module, we consider the number of positive and negative emoticons found in the text to determine its classification.

3.2 Lexicon-based Classifier

The lexicon-based classifier is based on the idea that the polarity of a text can be given by the sum of the individual polarity values of each word or phrase present in the text. For this, a sentiment lexicon identifies polarity words and assigns polarity values to them (known as semantic orientations).

In the 2013 system, we had used SentiStrength lexicon (Thelwall et al., 2010). In 2014, we improved our lexicon-based classifier by using a larger sentiment lexicon. We used the sentiment lexicon provided by Opinion-Lexicon (Hu and Liu, 2004) and a list of sentiment hashtags provided by the NRC Hashtag Sentiment Lexicon (Mohammad et al., 2013). For dealing with negation, we used a handcrafted list of negative words.

In our algorithm, the semantic orientations of each individual word in the text are added up. In this approach, the algorithm searches for each word in the lexicon and only the words that were found are returned. We associate the value +1 to the positive words, and -1 to the negative words. If a polarity word is negated, its value is inverted. This lexicon-based classifier assumes the signal of the final score as the sentiment class (positive or negative) and the score zero as neutral.

3.3 Machine Learning Classifier

The machine learning classifier uses labeled examples to learn how to classify new instances. The features used for this 2014 system were completely changed from 2013 system. We inspired our machine learning module in the work reported by Mohammad et al. (2013). The features used by the classifier are:

1. unigrams, bigrams and trigrams
2. the presence of negation
3. the presence of three or more characters in the words
4. the sequence of three or more punctuation marks
5. the number of words with all letters in uppercase
6. the total number of each tag present in the text
7. the number of positive words computed by the lexicon-based method
8. the number of negative words computed by the lexicon-based method

We use a Linear Kernel SVM classifier provided by the python sckit-learn library with C=0.005[1].

4 Hybrid Approach and Tuning

The organization from *SemEval-2014 Task 9: Sentiment Analysis in Twitter* provided four datasets for the task: a training dataset (TrainSet) with 9675 messages directly retrieved from Twitter; a development dataset (DevSet), with 1654 messages; the testing dataset from 2013 run, which was not used; and the testing dataset for 2014 with 8987 messages. The 2014 testing dataset was composed of 5 different sources:

- Twitter2013: Twitter test data from 2013 run
- SMS2013: SMS test data from 2013 run
- Twitter2014: 2000 tweets
- LiveJournal2014: 2000 sentences from LiveJournal blogs
- Twitter2014Sarcasm: 100 tweets that contain sarcasm

As we said in the previous section, our system is a pipeline of classifiers where each classifier may assign a sentiment class if it achieves a particular confidence threshold score. This confidence score is a fixed value set for each system in order to have a decision boundary. This decision was made by inspecting the results obtained for the development set. Tables 1 and 2 shows how the rule-based and lexicon-based classifiers perform for the development dataset in terms of score. The score obtained by the rule-based classifier consists of the difference between the number of positive emoticons and the number of negative emoticons found in the messages. The score obtained by the lexicon-based classifier represents the total semantic orientation obtained by the algorithm by adding up the semantic orientation for their lexicon.

Inspecting Table 1, for the best threshold, we adjusted the rule-based classifier boundary to decide when the score is different from zero. For values greater than zero, the classifier will assign the positive class and, for values below zero, the classifier will assign the negative class. For values equal to zero, the classifier will call the lexicon-based classifier.

Table 1: Correlation between the rule-based classifier scores and the gold standard classes in the DevSet

Rule-based classifier score	Gold Standard Class		
	Negative	Neutral	Positive
-1	22	3	3
0	311	709	495
1	7	26	73
2	0	0	2
3 to 6	0	1	2

Inspecting Table 2, for the best threshold, we adjusted the lexicon-based classifier to assign the

[1] Available at http://scikit-learn.org/

positive class when the total score is greater than 1 and negative class when the total score is below -2. For any other values, the classifier will call the machine learning classifier.

Table 2: Correlation between the lexicon-based classifier score and the gold standard classes in the devset

Lexicon-based classifier scores	Gold Standard Class		
	Negative	Neutral	Positive
-7 to -4	2	0	0
-3	10	4	0
-2	48	18	7
-1	111	99	35
0	108	432	178
1	48	143	210
2	11	39	104
3 to 5	3	4	47

As the machine learning classifier is responsible for the final stage, we did not have to decide any threshold for this classifier. However, we empirically identified a bias toward the positive class (the negative class was barely chosen). In order to correct this problem, we setup the machine learning classifier to decide for the negative class whenever the SVM score for this class is bigger than -0.4. Next section shows the results achieved for the Semeval test dataset.

5 Results

Table 3 shows the results obtained by each individual classifier and by the hybrid classifier for the Twitter2014 messages in the testset. In the task, the systems were evaluated with the average F-score obtained for positive and negative classes.

Table 3: Average F-score (positive and negative) obtained by each classifier and the hybrid approach for the Twitter2014 testset

Classifier	Twitter2014 Testset
Rule-based	14.03
Lexicon-Based	47.55
Machine Learning	63.36
Hybrid Approach	**63.94**

Table 4 shows the improvement of the system over the 2013 run. Unlike last year, we notice that the performance of this hybrid system is very close to the performance of the machine-learning.

Table 4: Comparison of the average F-score (positive and negative) obtained by each classifier and the hybrid approach for the Twitter2013 testset for 2013 and 2014 versions

Classifier	2013 system	2014 system
Rule-based	14.37	13.31
Lexicon-Based	44.87	46.80
Machine Learning	49.99	63.75
Hybrid Approach	**56.31**	**65.39**

Table 5 shows the scores for each source in the testset. Last column shows our system rank among the 50 systems that participated in the competition. For the entire testing dataset, our algorithm had 503 (5%) examples classified by the rule-based classifier, 3204 (36%) by the lexicon-based classifier and 5280 (59%) by the machine learning classifier.

6 Conclusion

We described our improved hybrid classification system used for *Semeval-2014 Task 9: Sentiment Analysis in Twitter*. This work showed that this hybrid classifier can be improved as its modules are too. However, we noticed that, improving the lexicon and machine learning modules, the overall score tends towards the machine learning score.

The source code produced for the experiment is available at https://github.com/pedrobalage.

Acknowledgments

We would like to thank the organizers for their work in constructing the dataset and in the overseeing of the task. We also would like to thank FAPESP and SAMSUNG for supporting this work.

References

Pedro Balage Filho and Thiago Pardo. 2013. NILC_USP: A Hybrid System for Sentiment Analysis in Twitter Messages. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 568–572, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

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, New York, NY, USA. ACM.

Table 5: Results for Twitter TestSet

TestSet Source	Majority Baseline	Our Score	Best Result	Our Rank
Twitter2013	29.2	**65.39**	72.12	15th
SMS2013	19.0	**61.35**	70.28	16th
Twitter2014	34.6	**63.94**	70.96	19th
LiveJournal2014	27.2	**69.02**	74.84	18th
Twitter2014Sarcasm	27.7	**42.06**	58.16	34th

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA. ACM.

Arnd Christian König and Eric Brill. 2006. Reducing the human overhead in text categorization. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 598–603, New York, NY, USA. ACM.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

Nikolaos Malandrakis, Abe Kazemzadeh, Alexandros Potamianos, and Shrikanth Narayanan. 2013. SAIL: A hybrid approach to sentiment analysis. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 438–442, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia, June. Association for Computational Linguistics.

Rudy Prabowo and Mike Thelwall. 2009. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment Analysis in Twitter. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, Dublin, Ireland.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307, June.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, December.