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

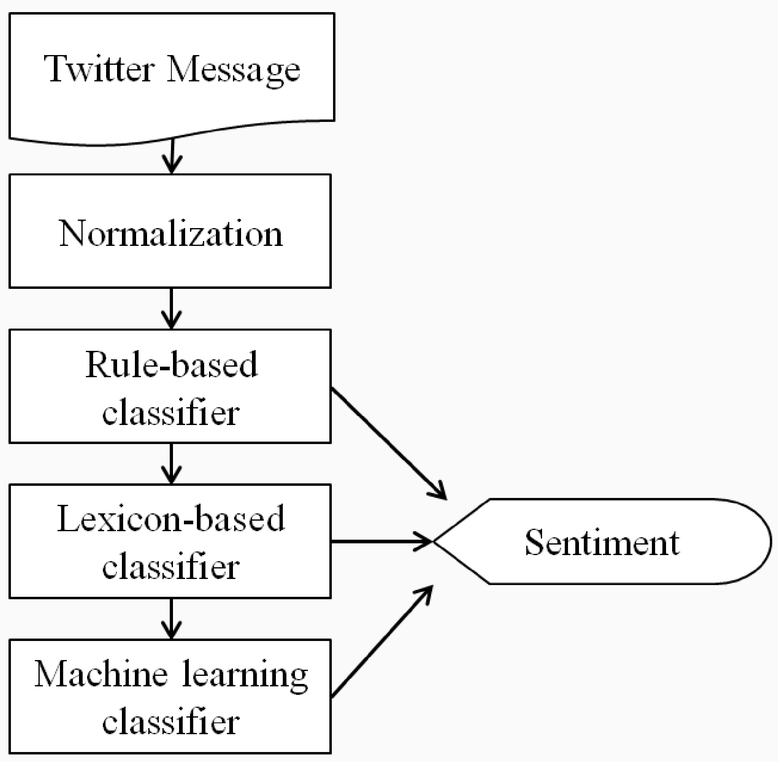
Pedro P. Balage Filho, Lucas Avanço, Thiago A. S. Pardo, Maria G. V. Nunes

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 poster 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.

Architecture



Normalization

- 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

Rule-based Classifier

Rule-based classifier is designed to provide rules that better impact the precision than the recall.
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.

Lexicon-based Classifier

Lexicons:

- Opinion-Lexicon
- NRC Hashtag Sentiment Lexicon
- 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.

Machine Learning Classifier

Features:

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

Comparative

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**

Results

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

Access links

Visit our paper

Check out our source code at github.com

Conclusion

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.