AUEB: Two Stage Sentiment Analysis of Social Network Messages

Rafael Michael Karampatsis, John Pavlopoulos and **Prodromos Malakasiotis**
mpatsis13@gmail.com, annis@aueb.gr, rulller@aueb.gr
Department of Informatics
Athens University of Economics and Business
Patission 76, GR-104 34 Athens, Greece

Abstract

This paper describes the system submitted for the Sentiment Analysis in Twitter Task of SEMEVAL 2014 and specifically the Message Polarity Classification subtask. We used a 2–stage pipeline approach employing a linear SVM classifier at each stage and several features including morphological features, POS tags based features and lexicon based features.

1 Introduction

Recently, Twitter has gained significant popularity among the social network services. Lots of users often use Twitter to express feelings or opinions about a variety of subjects. Analysing this kind of content can lead to useful information for fields, such as personalized marketing or social profiling. However such a task is not trivial, because the language used in Twitter is often informal presenting new challenges to text analysis.

In this paper we focus on sentiment analysis, the field of study that analyzes people's sentiment and opinions from written language (Liu, 2012). Given some text (e.g., tweet), sentiment analysis systems return a sentiment label, which most often is positive, negative, or neutral. This classification can be performed directly or in two stages; in the first stage the system examines whether the text carries sentiment and in the second stage, the system decides for the sentiment's polarity (i.e., positive or negative).[1] This decomposition is based on the assumption that subjectivity detection and sentiment polarity detection are different problems. We choose to follow the 2–stage approach, because it allows us to focus on each of the two problems separately (e.g., features, tuning, etc.). In the following we will describe the system with which we participated in the Message Polarity Classification subtask of Sentiment Analysis in Twitter (Task 9) of SEMEVAL 2014 (Rosenthal et al., 2014). Specifically Section 2 describes the data provided by the organizers of the task. Sections 3 and 4 present our system and its performance respectively. Finally, Section 5 concludes and provides hints for future work.

[1] For instance a 2–stage approach is better suited to systems that focus on subjectivity detection; e.g., aspect based sentiment analysis systems which extract aspect terms only from evaluative texts.

2 Data

At first, we describe the data used for this year's task. For system tuning the organizers released the training and development data of SEMEVAL 2013 Task 2 (Wilson et al., 2013). Both these sets are allowed to be used for training. The organizers also provided the test data of the same Task to be used for development only. As argued in (Malakasiotis et al., 2013) these data suffer from class imbalance. Concerning the test data, they contained 8987 messages broken down in the following 5 datasets:

- LJ_{14}: 2000 sentences from LIVEJOURNAL.
- SMS_{13}: SMS test data from last year.
- TW_{13}: Twitter test data from last year.
- TW_{14}: 2000 new tweets.
- TWSARC_{14}: 100 tweets containing sarcasm.

The details of the test data were made available to the participants only after the end of the Task. Recall that SMS_{13} and TW_{13} were also provided as development data. In this way the organizers were able to check, i) the progress of the systems since last year's task, and ii) the generalization capability of the participating systems.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 114–118,
Dublin, Ireland, August 23-24, 2014.

3 System Overview

The main objective of our system is to detect whether a message M expresses positive, negative or no sentiment. To achieve that we follow a 2–stage approach. During the first stage we detect whether M expresses sentiment ("subjective") or not; this process is called subjectivity detection. In the second stage we classify the "subjective" messages of the first stage as "positive" or "negative". Both stages utilize a Support Vector Machine (SVM (Vapnik, 1998)) classifier with linear kernel.[2] Similar approaches have also been proposed in (Pang and Lee, 2004; Wilson et al., 2005; Barbosa and Feng, 2010; Malakasiotis et al., 2013). Finally, we note that the 2–stage approach, in datasets such the one here (Malakasiotis et al., 2013), alleviates the class imbalance problem.

3.1 Data preprocessing

A very essential part of our system is data preprocessing. At first, each message M is passed through a twitter specific tokenizer and part-of-speech (POS) tagger (Owoputi et al., 2013) to obtain the tokens and the corresponding POS tags, which are necessary for some sets of features.[3] We then use a dictionary to replace any slang with the actual text.[4] We also normalize the text of each message by combining a trie data structure (De La Briandais, 1959) with an English dictionary.[5] In more detail, we replace every token of M not in the dictionary with the most similar word of the dictionary. Finally, we obtain POS tags of all the new tokens.

3.2 Sentiment lexicons

Another key attribute of our system is the use of sentiment lexicons. We have used the following:

- HL (Hu and Liu, 2004).
- SENTIWORDNET (Baccianella et al., 2010).
- SENTIWORDNET lexicon with POS tags (Baccianella et al., 2010).
- AFINN (Nielsen, 2011).
- MPQA (Wilson et al., 2005).
- NRC Emotion lexicon (Mohammad and Turney, 2013).
- NRC S140 lexicon (Mohammad et al., 2013).
- NRC Hashtag lexicon (Mohammad et al., 2013).
- The three lexicons created from the training data in (Malakasiotis et al., 2013).

Note that concerning the MPQA Lexicon we applied preprocessing similar to Malakasiotis et al. (2013) to obtain the following sub–lexicons:

S_+ : Contains strong subjective expressions with positive prior polarity.

S_- : Contains strong subjective expressions with negative prior polarity.

$S_\pm$: Contains strong subjective expressions with either positive or negative prior polarity.

S_0 : Contains strong subjective expressions with neutral prior polarity.

W_+ : Contains weak subjective expressions with positive prior polarity.

W_- : Contains weak subjective expressions with negative prior polarity.

$W_\pm$: Contains weak subjective expressions with either positive or negative prior polarity.

W_0 : Contains weak subjective expressions with neutral prior polarity.

3.3 Feature engineering

Our system employs several types of features based on morphological attributes of the messages, POS tags, and lexicons of section 3.2.[6]

3.3.1 Morphological features

- The existence of elongated tokens (e.g., "baaad").
- The number of elongated tokens.
- The existence of date references.
- The existence of time references.

[2] We used the LIBLINEAR distribution (Fan et al., 2008)

[3] Tokens could be words, emoticons, hashtags, etc. No lemmatization or stemming has been applied

[4] See `http://www.noslang.com/dictionary/`.

[5] We used the OPENOFFICE dictionary

[6] All the features are normalized to $[-1, 1]$

- The number of tokens that contain only upper case letters.
- The number of tokens that contain both upper and lower case letters.
- The number of tokens that start with an upper case letter.
- The number of exclamation marks.
- The number of question marks.
- The sum of exclamation and question marks.
- The number of tokens containing only exclamation marks.
- The number of tokens containing only question marks.
- The number of tokens containing only exclamation or question marks.
- The number of tokens containing only ellipsis (...).
- The existence of a subjective (i.e., positive or negative) emoticon at the message's end.
- The existence of an ellipsis and a link at the message's end.
- The existence of an exclamation mark at the message's end.
- The existence of a question mark at the message's end.
- The existence of a question or an exclamation mark at the message's end.
- The existence of slang.

3.3.2 POS based features

- The number of adjectives.
- The number of adverbs.
- The number of interjections.
- The number of verbs.
- The number of nouns.
- The number of proper nouns.
- The number of urls.
- The number of subjective emoticons.[7]
- The number of positive emoticons.[8]
- The number of negative emoticons.[9]
- The average, maximum and minimum F_1 scores of the message's POS bigrams for the subjective and the neutral classes.[10]
- The average, maximum and minimum F_1 scores of the message's POS bigrams for the positive and the negative classes.[11]

For a bigram b and a class c, F_1 is calculated as:

$$F_1(b,c) = \frac{2 \cdot Pre(b,c) \cdot Rec(b,c)}{Pre(b,c) + Rec(b,c)} \quad (1)$$

where:

$$Pre(b,c) = \frac{\#\text{messages of } c \text{ containing } b}{\#\text{messages containing } b} \quad (2)$$

$$Rec(b,c) = \frac{\#\text{messages of } c \text{ containing } b}{\#\text{messages of } c} \quad (3)$$

3.3.3 Sentiment lexicon based features

For each lexicon we use seven different features based on the scores provided by the lexicon for each word present in the message.[12]

- Sum of scores.
- Maximum of scores.
- Minimum of scores.
- Average of scores.
- The count of words with scores.
- The score of the last word of the message that appears in the lexicon.
- The score of the last word of the message.

[7] This feature is used only for subjectivity detection.
[8] This feature is used only for polarity detection.
[9] This feature is used only for polarity detection.
[10] This feature is used only for subjectivity detection.
[11] This feature is used only for polarity detection.
[12] If a word does not appear in the lexicon it is assigned with a score of 0 and it is not considered in the calculation of the average, maximum, minimum and count scores. Also, we have removed from SENTIWORDNET any instances having positive and negative scores that sum to zero. Moreover, the MPQA lexicon does not provide scores, so, for each word in the lexicon we assume a score equal to 1.

We also created features based on the Pre and F_1 scores of MPQA and the train data generated lexicons in a similar manner to that described in (Malakasiotis et al., 2013), with the difference that the features are stage dependent. Thus, for subjectivity detection we use the subjective and neutral classes and for polarity detection we use the positive and negative classes to compute the scores.

3.3.4 Miscellaneous features

Negation. Negation not only is a good subjectivity indicator but it also may change the polarity of a message. We therefore add 7 more features, one indicating the existence of negation, and the remaining six indicating the existence of negation that precedes words from lexicons $S_{\pm}$, S_{+}, S_{-}, $W_{\pm}$, W_{+} and W_{-}.[13] Each feature is used in the appropriate stage.[14] We have not implement this type of feature for other lexicons but it might be a good addition to the system.

Carnegie Mellon University's Twitter clusters. Owoputi et al. (2013) released a dataset of 938 clusters containing words coming from tweets. Words of the same clusters share similar attributes. We try to exploit this observation by adding 938 features, each of which indicates if a message's token appears or not in the corresponding attributes.

3.4 Feature Selection

To allow our model to better scale on unseen data we have performed feature selection. More specifically, we first merged training and development data of SEMEVAL 2013 Task 2. Then, we ranked the features with respect to their information gain (Quinlan, 1986) on this dataset. To obtain the best set of features we started with a set containing the top 50 features and we kept adding batches of 50 features until we have added all of them. At each step we evaluated the corresponding feature set on the TW_{13} and SMS_{13} datasets and chose the feature set with the best performance. This resulted in a system which used the top 900 features for Stage 1 and the top 1150 features for Stage 2.

[13] We use a list of words with negation. We assume that a token precedes a word if it is in a distance of at most 5 tokens.

[14] The features concerning $S_{\pm}$ and $W_{\pm}$ are used in subjectivity detection and the remaining four in polarity detection.

Test Set	AUEB	Median	Best
LJ_{14}	70.75	65.48	74.84
SMS_{13}	64.32	57.53	70.28
TW_{13}	63.92	62.88	72.12
TW_{14}	66.38	63.03	70.96
$TWSARC_{14}$	56.16	45.77	58.16
AVG_{all}	64.31	56.56	68.78
AVG_{14}	64.43	57.97	67.62

Table 1: $F_1(\pm)$ scores per dataset.

Test Set	Ranking
LJ_{14}	9/50
SMS_{13}	8/50
TW_{13}	21/50
TW_{14}	14/50
$TWSARC_{14}$	4/50
AVG_{all}	6/50
AVG_{14}	5/50

Table 2: Rankings of our system.

4 Experimental Results

The official measure of the Task is the average F_1 score of the positive and negative classes ($F_1(\pm)$). Table 1 illustrates the $F_1(\pm)$ score per evaluation dataset achieved by our system along with the median and best $F_1(\pm)$. In the same table AVG_{all} corresponds to the average $F_1(\pm)$ across the five datasets while AVG_{14} corresponds to the average $F_1(\pm)$ across LJ_{14}, TW_{14} and $TWSARC_{14}$. We observe that in all cases our results are above the median. Table 2 illustrates the ranking of our system according to $F_1(\pm)$. Our system ranked 6th according to AVG_{all} and 5th according to AVG_{14} among the 50 participating systems. Note that our best results were achieved on the new test sets (LJ_{14}, TW_{14}, $TWSARC_{14}$) meaning that our system has a good generalization ability.

5 Conclusion and future work

In this paper we presented our approach for the Message Polarity Classification subtask of the Sentiment Analysis in Twitter Task of SEMEVAL 2014. We proposed a 2–stage pipeline approach, which first detects sentiment and then decides about its polarity. The results indicate that our system handles well the class imbalance problem and has a good generalization ability. A possible explanation is that we do not use bag-of-words fea-

tures which often suffer from over–fitting. Nevertheless, there is still some room for improvement. A promising direction would be to improve the 1st stage (subjectivity detection) either by adding more data or by adding more features, mostly because the performance of stage 1 greatly affects that of stage 2. Finally, the addition of more data for the negative class on stage 2 might be a good improvement because it would further reduce the class imbalance of the training data for this stage.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may.

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: Posters*, COLING '10, pages 36–44, Beijing, China.

Rene De La Briandais. 1959. File searching using variable length keys. In *Papers Presented at the the March 3-5, 1959, Western Joint Computer Conference*, IRE-AIEE-ACM '59 (Western), pages 295–298, New York, NY, USA.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

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.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Prodromos Malakasiotis, Rafael Michael Karampatsis, Konstantina Makrynioti, and John Pavlopoulos. 2013. nlp.cs.aueb.gr: Two stage 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 562–567, Atlanta, Georgia, June.

Saif Mohammad and Peter Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, June.

Finn Årup Nielsen. 2011. A new anew: evaluation of a word list for sentiment analysis in microblogs. In Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, and Mariann Hardey, editors, *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98, May.

Olutobi Owoputi, Brendan OConnor, 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 NAACL*.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Barcelona, Spain.

Ross Quinlan. 1986. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March.

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.

Vladimir Vapnik. 1998. *Statistical learning theory*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354.

Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. 2013. SemEval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, June.