CISUC-KIS: Tackling Message Polarity Classification with a Large and Diverse set of Features

João Leal, Sara Pinto, Ana Bento, Hugo Gonçalo Oliveira, Paulo Gomes
CISUC, Department of Informatics Engineering
University of Coimbra
Portugal
{jleal,sarap,arbc}@student.dei.uc.pt, {hroliv,pgomes}@dei.uc.pt

Abstract

This paper presents the approach of the CISUC-KIS team to the SemEval 2014 task on Sentiment Analysis in Twitter, more precisely subtask B - Message Polarity Classification. We followed a machine learning approach where a SVM classifier was trained from a large and diverse set of features that included lexical, syntactic, sentiment and semantic-based aspects. This led to very interesting results which, in different datasets, put us always in the top-7 scores, including second position in the LiveJournal2014 dataset.

1 Introduction

Everyday people transmit their opinion in social networks and microblogging services. Identifying the sentiment transmitted in all those shared messages is of great utility for recognizing trends and supporting decision making, key in areas such as social marketing. Sentiment Analysis deals with the computational treatment of sentiments in natural language text, often normalized to positive or negative polarities. It is a very challenging task, not only for machines, but also for humans.

SemEval 2014 is a semantic evaluation of Natural Language Processing (NLP) that comprises several tasks. This paper describes our approach to the Sentiment Analysis in Twitter task, which comprises two subtasks: (A) Contextual Polarity Disambiguation; and (B) Message Polarity Classification. We ended up addressing only task B, which is more sentence oriented, as it targets the polarity of the full messages and not individual words in those messages.

We tackled this task with a machine learning-based approach, in which we first collect several features from the analysis of the given text at several levels. The collected features are then used to learn a sentiment classification model, which can be done with different algorithms. Features were collected from several different resources, including: sentiment lexicons, dictionaries and available APIs for this task. Moreover, since microblogging text has particular characteristics that increase the difficulty of NLP, we gave special focus on text pre-processing. Regarding the tested features, they went from low-level ones, such as punctuation and emoticons, to more high-level, including topics extracted using topic modelling techniques, as well features from sentiment lexicons, some structured on plain words and others based on WordNet, and thus structured on word senses. Using the latter, we even explored word sense disambiguation. We tested several learning algorithms with all these features, but Support Vector Machines (SVM) led to the best results, so it was used for the final evaluation.

In all our runs, a model was learned from tweets, and no SMS were used for training. The model's performance was assessed with the F-Score of positive and negative classes, with 10-fold cross validation. In the official evaluation, we achieved very interesting scores, namely: 74.46% for the LiveJournal2014 (2nd place), 65.9% for the SMS2013 (7th), 67.56% for the Twitter2013 (7th), 67.95% for the Twitter2014 (4th) and 55.49% for the Twitter2014Sarcasm (4th) datasets, which ranked us always among the top-7 participations.

The next section describes the external resources exploited. Section 3 presents our approach with more detail, and is followed by section 4, where the experimental results are described. Section 5 concludes with a brief balance and the main lessons learned from our participation.

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 166–170,
Dublin, Ireland, August 23-24, 2014.

2 External resources

We have used several external resources, including not only several sentiment lexicons, but also dictionaries that helped normalizing the text of the tweets, as well as available APIs that already classify the sentiment transmitted by a piece of text.

2.1 Sentiment Lexicons

We used several public handcrafted or semi-automatically created sentiment lexicons, where English words have assigned polarity values. Those included lexicons structured in plain words, namely Bing Liu's Opinion Lexicon (Hu and Liu, 2004) (≈2,000 positive and 4,800 negative words), the AFINN list (Nielsen, 2011) (≈2,500 words with polarities between 5 and -5, 900 positive and 1,600 negative), the NRCEmoticon Lexicon (Mohammad and Turney, 2010) (≈14,000 words, their polarity, ≈2,300 positive, ≈3,300 negative, and eight basic emotions), MPQA Subjectivity Lexicon (Wilson et al., 2005) (≈2,700 positive and ≈4,900 negative words), Sentiment140 Lexicon (Mohammad et al., 2013) (≈62,000 unigrams, ≈677,000 bigrams; ≈480,000 pairs), NRC Hashtag Lexicon (Mohammad et al., 2013) (≈54,000 unigrams; ≈316,000 bigrams; ≈308,000 pairs) and labMT 1.0 (Dodds et al., 2011) (≈10,000 words).

We also used two resources with polarities assigned automatically to a subset of Princeton WordNet (Fellbaum, 1998) synsets, namely SentiWordNet 3.0 (Baccianella et al., 2010) (≈117,000 synsets with graded positive and negatives strengths between 0 and 1), and Q-WordNet (Agerri and García-Serrano, 2010) (≈7,400 positive and ≈8,100 negative senses).

2.2 Dictionaries

These included handcrafted dictionaries with the most common abbreviations, acronyms, emoticons and web slang used on the Internet and their meaning. Also, a list of regular expressions with elongated words like '*loool*' and '*loloolll*', which can be normalized to '*lol*', and a set of idiomatic expressions and their corresponding polarity.

2.3 APIs

Three public APIs were used, namely Sentiment140 (Go et al., 2009), SentimentAnalyzer[1] and SentiStrength (Thelwall et al., 2012). All of a them classify a given text snippet as positive or negative. Sentiment140 returns a value which can be 0 (negative polarity), 2 (neutral), and 4 (positive). SentimentAnalyzer returns -1 (negative) or 1 (positive), and SentiStrength a strength value between 1 and 5 (positive) or -1 and -5 (negative).

[1] http://sentimentanalyzer.appspot.com/

3 Approach

Our approach consisted of extracting lexical, syntactic, semantic and sentiment information from the tweets and using it in the form of features, for learning a sentiment classifier that would detect polarity in messages. This is a popular approach for these types of tasks, followed by other systems, including the winner of SemEval 2013 (Mohammad et al., 2013), where a variety of surface-form, semantic, and sentiment features was used. Our set of features is similar for the base classifier are similar, except that we included additional features that take advantage of word disambiguation to get the polarity of target word senses.

3.1 Features

Among the collected features, some were related to the content of the tweets and others were obtained from the sentiment lexicons.

3.1.1 Content Features

The tweets were tokenized and part-of-speech (POS) tagged with the CMU ARK Twitter NLP Tool (Gimpel et al., 2011) and Stanford CoreNLP (Toutanova and Manning, 2000). Each tweet was represented as a feature vector containing the following group of features: (i) emoticons (presence or absence, sum of all positive and negative polarities associated with each, polarity of the last emoticon of each tweet); (ii) length (total length of the tweet, average length per word, number of words per tweet); (iii) elongated words (number of all the words containing a repeated character more than two times); (iv) hashtags (total number of hashtags); (v) topic modelling (id of the corresponding topic); (vi) capital letters (number of words in which all letters are capitalized); (vii) negation (number of words that reverse polarity to a negative context, such as 'no' or 'never'); (viii) punctuation (number of punctuation sequences with only exclamation points, question marks or both, ASCII code of the most common punctuation and of the last punctuation in every

tweet); (ix) dashes and asterisks (number of words surrounded by dashes or asterisks, such as '*yay*' or '-me-'); (x) POS (number of nouns, adjectives, adverbs, verbs and interjections).

3.1.2 Lexicon Features

A wide range of features were created using the lexicons. For each tweet and for each lexicon the following set of features were generated: (i) total number of positive and negative opinion words; (ii) sum of all positive/negative polarity values in the tweet; (iii) the highest positive/negative polarity value in the tweet; and (iv) the polarity value of the last polarity word. Those features were collected for: unigrams, bigrams and pairs (only on the NRC Hashtag Lexicon and Sentiment140), nouns, adjectives, verbs, interjections, hashtags, all caps tokens (e.g 'GO AWAY'), elongated words, asterisks and dashes tokens.

Different approaches were followed to get the polarity of each word from the wordnets. From SentiWordNet, we computed combined scores of all senses, with decreasing weights for lower ranked senses, as well as the scores of the first sense only, both considering: (i) positive and negative; (ii) just positive; (iii) just negative scores. Moreover, we performed word sense disambiguation using the full WordNet 3.0 to get the previous scores for the selected sense. For this purpose, we applied the Lesk Algorithm adapted to wordnets (Banerjee and Pedersen, 2002), using all the tweet's content words as the word context, and the synset words, gloss words and words in related synsets as the synset's context. Given that SentiWordNet is aligned to WordNet 3.0, after selecting the most adequate sense of the word, we could get its polarity scores. From Q-WordNet, similar scores were computed but, since it doesn't use a graded strength and only classifies word senses as positive or negative, there were just positive or just negative scores.

3.2 Classifier

In our final approach we used a SVM (Fan et al., 2008) which is an effective solution in high dimensional spaces and proved to be the best learning algorithm for this task. We tested various kernels (e.g. PolyKernel, RBF) and their parameters with cross validation on the training data. Given the results, we confirmed that the RBF kernel, computed according to equation 1, is most effective with a $C = 4$ and a $\gamma = 0.0003$.

$$K(x^i, x^j) = \Phi(x^i)^T \Phi(x^j) = exp(-\gamma||x^i - x^j||^2) \quad (1)$$

Considering we are working on a multi-class classification problem, we implemented the "one-against-one" approach (Knerr et al., 1990) where $\#classes * (\#classes - 1)/2$ classifiers are constructed and each one trains data from classes. Due to the non-scale invariant nature of SVM algorithms, we've scaled our data on each attribute to have $\mu = 0$ and $\sigma = 1$ and took caution against class unbalance.

4 Experiments

For training the SVM classifier, we used a set of 9,634 tweets with a known polarity and also 1,281 tweets as development test to grid search the best parameters. No SMS messages were used as training or as development test. For the scorer function, we used a macro-averaged F-Score of positive and negative classes – the one made available and used by the task organizers.

4.1 Some Results

The results obtained by the system were 70.41% on the training set (using 10-Folds) and 71.03% on the development set, after train on the training set. When tested against the training set, after train in the same set, we get a score of 84.32%, which could indicate a case of underfitting. Though, our classifier generalized well, given that we got a 74.46% official score on LiveJournal2014, second in that category. On the other hand, our experiments with decision trees showed that they couldn't generalize so well, although they achieved scores of >99 on the training set. In the SMS category, our system would benefit from a specific data set in the training phase. Yet, it still managed to reach 7th place in that category. In the sarcasm category our submission ranked 4th, with a score of 58.16%, 2.69% below the best rank. On the Twitter2014 dataset, we scored 67.95% (4th), which is slightly below our prediction based on development tests. A possible explanation is that we might have over-fitted the classifier parameters when grid searching.

4.2 Features Relevance

In order to get some insights on the most relevant group of features, we did a series of experiments where each group of features were removed for

the classification, then tested against the original score. We concluded that the lexicon related features contribute highly to the performance of our system, including the set of features with n-grams and POS. Clusters, sport score, asterisks and elongated words provide little gains but, on the other hand, emoticons and hashtags showed some importance and provided enough new information for the system to learn. The API information is largely captured by some of our features and that makes it much less discriminating than what they would be on their own, but still worth using for the small gain. We also observed that it is best to create a diversified set of lexicon features with extra very specific targeted features, such as punctuation, instead of focusing on using a specific lexicon alone. Even though they usually overlap in information and may perform worse individually than a hand-refined single dictionary approach, they complement each other and that results in larger gains.

4.3 Selected Parameters

For the parameter values, we did a grid search using the development set as a test. We also found that large values of C (25) and small γ values (0.0001) performed worse than smaller values of C (4) with a slightly higher γ (0.0003) when using the development set but not when using the training set under K-Folds. For the official evaluation, we opted for the best-performing results on the development set. Using intermediate values accomplished worse results in either case.

5 Concluding Remarks

We have described the work developed for the subtask B of SemEval 2014 Sentiment Analysis in Twitter task. We followed a machine learning approach, with a diversified set of features, which tend to complemented each other. Some of the main takeaways are that the most important features are the lexicon related ones, including the n-grams and POS tags. Due to time constraints, we could not take strong conclusions on the impact of the word sense disambiguation related features alone. As those are probably the most differentiating features of our classifier, this is something we wish to target in the future.

To conclude, we have achieved very interesting results in terms of overall classification. Considering that this was our first participation in such an evaluation, we make a very positive balance. And of course, we are looking forward for upcoming editions of this task.

Acknowledgement

This work was supported by the iCIS project (CENTRO-07-ST24-FEDER-002003), co-financed by QREN, in the scope of the Mais Centro Program and European Union's FEDER.

References

Rodrigo Agerri and Ana García-Serrano. 2010. Q-wordnet: Extracting polarity from WordNet senses. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC 2010, pages 2300–2305, La Valletta, Malta. ELRA.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC 2010, pages 2200–2204, Valletta, Malta. ELRA.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, volume 2276 of *LNCS*, pages 136–145, London, UK. Springer.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, 6(12), December.

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

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, ACL 2011, pages 42–47. ACL Press.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford University.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2004, pages 168–177.

Stefan Knerr, Léon Personnaz, and Gérard Dreyfus. 1990. Single-layer learning revisited: a stepwise procedure for building and training neural network. In *Proceedings of the NATO Advanced Research Workshop on Neurocomputing, Algorithms, Architectures and Applications*, Nato ASI, Computer and Systems Sciences, pages 41–50. Springer.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 26–34, Los Angeles, CA. ACL Press.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *2nd 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. ACL Press.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, pages 93–98, May.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, January.

Kristina Toutanova and Christopher D Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, EMNLP 2000, pages 63–70. ACL Press.

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*, HLT '05, pages 347–354, Vancouver, British Columbia, Canada. ACL Press.