OPTWIMA: Comparing Knowledge-rich and Knowledge-poor Approaches for Sentiment Analysis in Short Informal Texts

Alexandra Balahur
European Commission Joint Research Centre
Via E. Fermi 2749
21027 Ispra (VA), Italy
{alexandra.balahur}@jrc.ec.europa.eu

Abstract

The fast development of Social Media made it possible for people to no loger remain mere spectators to the events that happen in the world, but become part of them, commenting on their developments and the entities involved, sharing their opinions and distributing related content. This phenomenon is of high importance to news monitoring systems, whose aim is to obtain an informative snapshot of media events and related comments.

This paper presents the strategies employed in the OPTWIMA participation to SemEval 2013 Task 2-Sentiment Analysis in Twitter. The main goal was to evaluate the best settings for a sentiment analysis component to be added to the online news monitoring system.

We describe the approaches used in the competition and the additional experiments performed combining different datasets for training, using or not slang replacement and generalizing sentiment-bearing terms by replacing them with unique labels.

The results regarding tweet classification are promising and show that sentiment generalization can be an effective approach for tweets and that SMS language is difficult to tackle, even when specific normalization resources are employed.

1 Introduction

Sentiment analysis is the Natural Language Processing (NLP) task dealing with the detection and classification of sentiments in texts. Usually, the classes considered are "positive", "negative" and "neutral", although in some cases finer-grained categories are added (e.g. "very positive" and "very negative") or only the "positive" and "negative" classes are taken into account.

This task has received a lot of interest from the research community in the past years. The work done regarded the manner in which sentiment can be classified from texts pertaining to different genres and distinct languages, in the context of various applications, using knowledge-based, semi-supervised and supervised methods [Pang and Lee, 2008]. The result of the analyses performed have shown that the different types of text require specialized methods for sentiment analysis, as, for example, sentiments are not conveyed in the same manner in newspaper articles and in blogs, reviews, forums or other types of user-generated contents [Balahur et al., 2010].

In the light of these findings, dealing with sentiment analysis in tweets and SMS (that we can generally call "short informal texts") requires an analysis of the characteristics of such texts and the design of adapted methods.

Our participation in the SemEval 2013 Task 2 [Wilson et al., 2013] had as objective to test how well our proposed methods for sentiment analysis for short informal texts (especially tweets) would perform. The two subtasks proposed in this competition were: a) the classification of sentiment from snippets from tweets and SMS marked as start and end position and b) the classification of sentiment from entire tweets and SMS. Each team could submit 2 runs for each dataset and task, one employing as training data only the data provided within the competition ("constrained") and the second em-

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 460–465, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

ploying any additional data ("unconstrained"). We submitted 2 of such runs for each of the subtasks and datasets.

The main requirements for the system we implemented were: a) not to use language-specific NLP processing tools (since our final goal is to make the present system work for many more languages); and b) to work fast, so that it can be integrated in a near real time media monitoring system.

2 Related Work and Contribution

One of the first studies on the classification of polarity in tweets was Go et al. [2009]. The authors conducted a supervised classification study on tweets in English, using the emoticons (e.g. ":)", ":(", etc.) as markers of positive and negative tweets. Read [2005] employed this method to generate a corpus of positive tweets, with positive emoticons ":)", and negative tweets with negative emoticons ":(". Subsequently, they employ different supervised approaches (SVM, Naïve Bayes and Maximum Entropy) and various sets of features and conclude that the simple use of unigrams leads to good results, but it can be slightly improved by the combination of unigrams and bigrams.

In the same line of thinking, Pak and Paroubek [2010] also generated a corpus of tweets for sentiment analysis, by selecting positive and negative tweets based on the presence of specific emoticons. Subsequently, they compare different supervised approaches with n-gram features and obtain the best results using Naïve Bayes with unigrams and part-of-speech tags.

Another approach on sentiment analysis in tweet is that of Zhang et al. [2011]. Here, the authors employ a hybrid approach, combining supervised learning with the knowledge on sentiment-bearing words, which they extract from the DAL sentiment dictionary [Whissell, 1989]. Their pre-processing stage includes the removal of retweets, translation of abbreviations into original terms and deleting of links, a tokenization process, and part-of-speech tagging. They employ various supervised learning algorithms to classify tweets into positive and negative, using n-gram features with SVM and syntactic features with Partial Tree Kernels, combined with the knowledge on the polarity of the words appearing in the tweets. The authors conclude that the most important features are those corresponding to sentiment-bearing words. Finally, Jiang et al. [2011] classify sentiment expressed on previously-given "targets" in tweets. They add information on the context of the tweet to its text (e.g. the event that it is related to). Subsequently, they employ SVM and General Inquirer and perform a three-way classification (positive, negative, neutral).

The main contributions of the approaches considered for the competition reside in the evaluation of different strategies to adapt sentiment analysis methods to the language employed in short informal texts.

The methods employed in our system are simple, work fast and efficient and can be easily adapted to other languages. The main adaptations we consider are part of a pre-processing step, in which the language in these short informal texts is normalized (brought to a dictionary form).

Finally, the methods presented are compared on different configurations and training sets, so that the conclusions drawn are relevant to the phenomena found in this type of informal texts.

3 Methods Employed by OPTWIMA in SemEval 2013 Task 2

We employ two different approaches: a) one based on supervised learning using Support Vector Machines Sequential Minimal Optimization (SVM SMO) using unigram and bigram features; and b) a hybrid approach, based on supervised learning with a SVM SMO linear kernel, on unigram and bigram features, but exploiting as features sentiment dictionaries, emoticon lists, slang lists and other social media-specific features. SVM SMO was preferred due to the computation speed. We do not employ any specific language analysis software. The aim is to be able to apply, in a straightforward manner, the same approach to as many languages as possible. The approach can be extended to other languages by using similar dictionaries that have been created in our team Steinberger et al. [2011].

The sentiment analysis process contains two stages: preprocessing and sentiment classification.

3.1 Preprocessing of Short Informal Texts

The language employed in short informal texts such as tweets and SMS is different from the one found in other types of texts, such as newspaper articles and the form of the words employed is sometimes not the one we may find in a dictionary. Further on, users writing on Twitter or SMS-ing on their cell phone employ a special "slang" (i.e. informal language, with special expressions, such as "lol", "omg"), emoticons, and often emphasize words by repeating some of their letters. Additionally, the language employed in Twitter has specific characteristics, such as the markup of tweets that were reposted by other users with "RT", the markup of topics using the "#" (hash sign) and of the users using the "@" sign.

All these aspects must be considered at the time of processing tweets and, to some extent, SMS.

As such, before applying supervised learning to classify the sentiment of the short informal texts considered, we preprocess them, to normalize the language they contain and try to abstract on the concepts that are sentiment-bearing, by replacing them with labels, according to their polarity[1]. In case of SMS messages, the slang employed, the short forms of words and the acronyms make these texts non processable without prior replacement and normalization of the slang. The preprocessing stage contains the following steps:

- Repeated punctuation sign normalization (RPSN).

  In the first step of the preprocessing, we detect repetitions of punctuation signs (".", "!" and "?"). Multiple consecutive punctuation signs are replaced with the labels "multistop", for the fullstops, "multiexclamation" in the case of exclamation sign and "multiquestion" for the question mark and spaces before and after.

- Emoticon replacement (ER).

  In the second step of the preprocessing, we employ the annotated list of emoticons from SentiStrength[2] and match the content of the tweets against this list. The emoticons found are replaced with their polarity ("positive" or "negative") and the "neutral" ones are deleted.

- Lower casing and tokenization (LCN).

  Subsequently, the tweets are lower cased and split into tokens, based on spaces and punctuation signs.

- Slang replacement (SR).

  The next step involves the normalization of the language employed. In order to be able to include the semantics of the expressions frequently used in Social Media, we employed the list of slang expressions from dedicated sites [3]. This step is especially relevant to SMS texts, whose language in their original form has little to do with language employed in ordinary texts.

- Word normalization (WN).

  At this stage, the tokens are compared to entries in Roget's Thesaurus. If no match is found, repeated letters are sequentially reduced to two or one until a match is found in the dictionary (e.g. "perrrrrrrrrrrrrrrrrrfeeect" becomes "perrfeect", "perfeect", "perrfect" and subsequently "perfect"). The words used in this form are maked as "stressed".

- Affect word matching (AWM).

  Further on, the tokens in the tweet are matched against three different sentiment lexicons: General Inquirer, LIWC and MicroWNOp, which were previously split into four different categories ("positive", "high positive", "negative" and "high negative"). Matched words are replaced with their sentiment label - i.e. "positive", "negative", "hpositive" and "hnegative".

- Modifier word matching (MWM).

  Similar to the previous step, we employ a list of expressions that negate, intensify or diminish the intensity of the sentiment expressed to detect such words in the tweets. If such a word is matched, it is replaced with "negator", "intensifier" or "diminisher", respectively.

[1] The preprocessing steps involving the use of affect dictionaries and modifier replacement are used only in one of the two methods considered

[2] http://sentistrength.wlv.ac.uk/

[3] www.noslang.com/dictionary, www.smsslang.com

- User and topic labeling (UTL).

  Finally, the users mentioned in the tweet, which are marked with "@", are replaced with "PERSON" and the topics which the tweet refers to (marked with "#") are replaced with "TOPIC".

3.2 Sentiment Classification of Short Informal Texts

Once the texts are preprocessed, they are passed on to the sentiment classification module.

We employed supervised learning using Support Vector Machines Sequential Minimal Optimization (SVM SMO) [Platt, 1998] with a linear kernel, employing boolean features - the presence or absence of unigrams and bigrams determined from the training data (tweets that were previousely preprocessed as described above) that appeared at least twice. Bigrams are used especially to spot the influence of modifiers (negations, intensifiers, diminishers) on the polarity of the sentiment-bearing words. We tested different parameters for the kernel and modified only the C constant to the best value determined on the training data (5.0)/

We tested the approach on different datasets and dataset splits, using the Weka data mining software [4]. The training models are built on a cluster of computers (4 cores, 5000MB of memory each).

4 Evaluation and Discussion

We participated in SemEval 2013 in Task 2 with two versions of the system, for each of the two subtasks (A and B). The main difference among them is the use of dictionaries for affect and modifier word matching and replacement. As such, in the first method (denoted as "Dict"), we perform all the preprocessing steps mentioned above, while the second method is applied on the data on which the AWM and MWM are not performed (i.e. words that are associated with a sentiment in a lexicon are not replaced with labels). This second method will be denoted "NoDict".

Another difference between the different evaluations we performed are the datasets employed for training. We created different models, employing:

1) For both the "Constrained" and "Unconstrained" submissions, the development and training data from the corresponding subtask (i.e. using as training the data in subtask A - the sets given as training and development together - to train a classifier for the test data in task A; the same for subtask B). In this case, the training data is marked with the corresponding subtask (i.e. training data "A", training data "B");

2) For both the "Constrained" and "Unconstrained" submissions, the development and training data from both subtasks - both training and development sets - to train one classifier which is used for both subtasks. This training set is denoted as "A+B";

3) For the "Unconstrained" submissions, we added to the joint training and development data from both subtasks the set of MySpace comments provided by [Thelwall et al., 2010]. This small set contains 1300 short texts from the MySpace social network[5]. The motivation behind this choice is that texts from this source are very similar in language and structure to tweets and (after slang replacement) SMS.

Finally, we trained different classifiers on the training sets described, with and without replacing the affective and modifier words and with and without employing the slang replacement pre-processing step.

The results are presented in Tables 1, 2, 3, 4, in terms of average F-measure of the positive and negative classes (as used by the organizers). The runs submitted in the competition are marked with an asterisk ("*"). We did not perform all the experiments for the sets of SMS without slang replacement, as the first results were very low.

As we can see from the results, our approach performed better in classifying the overall sentiment of texts than small snippets. The results were significantly better for the classification of tweets in comparison to SMS, whose language (even with slang replacement) made them difficult to tackle. We can also see that the joint use of slang replacement and dictionaries for tweets leads to significantly lower results, meaning that this step (at least with the resources we employed for slang treatment), is not necessary for the treatment of tweets. Instead, for these texts, the use of affect dictionaries and modifier lists and their generalizaton lead to better re-

[4] http://www.cs.waikato.ac.nz/ml/weka/

[5] http://www.myspace.com/

	Trained on A+B with slang replacement (Constrained)	
Test set	Dict	NoDict
Task A Tweets	0.35	0.37
Task A SMS	0.35	0.37*
Task B Tweets	0.45*	0.54
Task B SMS	0.40*	0.47

Table 1: Results obtained using A+B (train and developement data) as training set and replacing the slang.

	Trained on A+B+MySpace with slang replacement (Unconstrained)	
Test Set	Dict	NoDict
Task A Tweets	0.36	0.39*
Task A SMS	0.37*	0.37
Task B Tweets	0.46	0.54*
Task B SMS	0.40	0.37*

Table 2: Results obtained using A+B+MySpace (train and developement data) as training set and replacing the slang.

sults. This proves that such a generalization, in the context of "legible" texts, is a useful tool for sentiment analysis. Further on, the results showed that adding a small quantity of training data led to no significant growth in performance (for the data in which slang was replaced). Additional evaluations could be made to quantify the effect of this data when other methods to generalize are not applied. As an observation, our results were balanced for all three classes, with even higher scores for the neutral class. We believe this class should have been considered as well, since in real-world settings systems for sentiment analysis must also be able to classify texts pertaining to this category.

Finally, we can see that in the case of SMS, the difference between the use of slang with or without affect label generalizations is insignificant. We believe this is due to the fact that the expressions with which the slang is replaced are very infrequent in traditional sentiment dictionaries (such as the ones we employed). Even by replacing the short forms and slang with their equivalents, the texts obtained contain words that are infrequent in other types of texts, even tweets. However, we will perform additional experiments with other lists of slang and add, as much as it is possible, the informal sentiment-bearing expressions to create new affect resources for this types of texts.

5 Conclusions and Future Work

In this article, we presented and evaluated the approaches considered for our participation in the SemEval 2013 Task 2. We evaluated different combinations of features, resources and training sets and applied different methods to tackle the issues brought by the informal language used in tweets and SMS.

As future work, we would like to extend the system to more languages, using the dictionaries created by Steinberger et al. [2011] and analyze and include new features that are particular to social media - especially tweets - to improve the performance of the sentiment analysis component. Further on, we would like to quantify the influence of using linguistic processing tools to perform lemmatizing, POS-tagging and the inclusion of corresponding features on the final performance of the system. Finally, we would like to explore additional resources to deal with the issue of language informality in tweets and further explore the problems posed by the peculiar language employed in SMS.

References

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. Sentiment analysis in the news. In *Proceedings*

	Trained on data of subtask (A or B) with slang replacement	
Test Set	Dict	NoDict
Task A Tweets	0.36	0.37
Task A SMS	0.36	0.37
Task B Tweets	0.5	0.55
Task B SMS	0.49	0.53

Table 3: Results obtained using A (train and developement data) or B (train and developement data) as training set and replacing the slang.

	Trained on data of subtask (A or B), no slang replacement		Trained on A+B, no slang replacement	
Test Set	Dict	NoDict	Dict	NoDict
Task A Tweets	0.69*	0.59	0.6	0.69
Task B Tweets	0.59	0.51	0.62	0.44

Table 4: Results obtained for tweet classification using A+B or A or B as training set and not replacing the slang.

of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, may 2010.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6, 2009.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 151–160. ACL, 2011. ISBN 978-1-932432-87-9.

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta; ELRA, may 2010. ELRA. ISBN 2-9517408-6-7. 19-21.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2): 1–135, January 2008. ISSN 1554-0669.

John C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Advances in Kernel Methods - Support Vector Learning, 1998.

Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA, 2005.

J. Steinberger, P. Lenkova, M. Ebrahim, M. Ehrmann, A. Hurriyetoglu, M. Kabadjov, R. Steinberger, H. Tanev, V. Zavarella, and S. Vázquez. Creating sentiment dictionaries via triangulation. In *Proceedings of WASSA 2011*, WASSA '11, pages 28–36. ACL, 2011.

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

Cynthia Whissell. The Dictionary of Affect in Language. In Robert Plutchik and Henry Kellerman, editors, *Emotion: theory, research and experience*, volume 4, The measurement of emotions. Academic Press, London, 1989.

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

Ley Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. Technical Report HPL-2011-89, HP, 21/06/2011 2011.