OPAL at SemEval-2016 Task 4: the Challenge of Porting a Sentiment Analysis System to the "Real" World

Alexandra Balahur
European Commission Joint Research Centre
Institute for the Protection and Security of the Citizen
Via E. Fermi 2749, 21027 Ispra (VA), Italy
alexandra.balahur@jrc.ec.europa.eu

Abstract

Sentiment analysis has become a well-established task in Natural Language Processing. As such, a high variety of methods have been proposed to tackle it, for different types of texts, text levels, languages, domains and formality levels. Although state-of-the-art systems have obtained promising results, a big challenge that still remains is to port the systems to the "real world" – i.e. to implement systems that are running around the clock, dealing with information of heterogeneous nature, from different domains, written in different styles and diverse in formality levels. The present paper describes our efforts to implement such a system, using a variety of strategies to homogenize the input and comparing various approaches to tackle the task. Specifically, we are tackling the task using two different approaches: a) one that is unsupervised, based on dictionaries of sentiment-bearing words and heuristics to compute final polarity of the text considered; b) the second, supervised, trained on previously annotated data from different domains. For both approaches, the data is first normalized and the slang is replaced with its expanded version.

1 Introduction

Sentiment analysis is the task in Natural Language Processing (NLP) that deals with classifying opinions according to the polarity of the sentiment they express. Due to the large quantities of user-generated online contents available on different Internet sites (forums, social networks, blogs, review sites, microblogs, etc.) and the possible value they can have for different domains (Marketing, e-Rulemaking, Political Science, etc.), Sentiment Analysis has become a very popular task in the field in the past decade.

As such, a high variety of methods have been proposed to tackle it, for different types of texts, text levels, languages, domains and formality levels. Although state of the art systems have obtained promising results (most of all in priory defined datasets – domains, languages, styles, formality levels), a big challenge that still remains is to port the systems to the "real world" – i.e. systems that are running around the clock, dealing with information of heterogeneous nature, from different domains, written in different styles and diverse in formality levels.

The present paper describes our efforts to build such a system. The challenge is not straightforward to tackle, as this entails building a system that is: a) on the one hand, robust enough to obtain similar levels of performance across domains, languages, text types and formality levels; b) on the other hand, flexible enough to treat all these types of texts. Further on, this entails that such a system must have components to treat the input to make it as homogeneous as possible, so as it can be treated in an equal way (POS-tagging, lemmatizing, syntactic parsing, etc.). The methods employed have to be general enough so that they can be used for as many different languages as possible. This is especially difficult, since there are languages that are under-resourced and have little tools available (i.e. it is not possible to perform accurate syntactic

Proceedings of SemEval-2016, pages 262–265,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

parsing in all languages, lemmatizing is more difficult for some languages than others, etc.).
In order to address these issues, we proposed two approaches, which we plan to eventually combine in a unique system. The first approach is based on knowledge, taken from dictionaries of sentiment-bearing words. A variant of this system is implemented in-house to compute the tonality of entity mentions in the news. The second is based on supervised learning and is implemented in a system we are currently running in-house to classify the sentiment of tweets on different subjects. A model is trained on a set of input data and is subsequently used to classify new examples. The next sections detail on the implementation of the two methods.

2 State of the art and related approaches.

As far as tweet processing and sentiment analysis in tweets is concerned, Go et al. (2009) pioneered research proposing the use of 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 ":(". 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. These approaches employed different supervised approaches for sentiment analysis, using n-grams as features. Zhang et al. (2011) employ a hybrid approach, combining supervised learning with the knowledge on sentiment-bearing words, which they extract from the DAL sentiment dictionary (Whissell, 1989). 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 adding the context of the tweet to increase the text length.
The approaches employed are based on the system we developed in (Balahur et al., 2010) and (Balahur, 2013).

3 OPAL: Approaches in SemEval 2016 Task 4 A and B

The OPAL system participated in SemEval in Task 4, subtasks A and B.
In subtask A, the method employed was based on sentiment dictionaries and rules to compute the final polarity of the tweets.
In subtask B, the system was trained on the training data provided by the organizers using SVM and uni- and bigrams.
Before applying each of the two approaches, the texts were pre-processed in order to be transformed from informal to formal language, to be treated by traditional text processing methods. Additionally, the words are matched against our in-house (Balahur et al., 2010)[1] sentiment and modifier dictionaries.
In the next subsections, we detail the steps, methods and resources employed.

3.1. Text Pre-processing

- **Multiple punctuation sign identification.** 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. The entire context before the punctuation sign, up until the previous punctuation sign, is marked as "intensifier".
- **Emoticon replacement**. 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 score from this resource.
- **Lower casing and tokenization.** Further on, the tweets are lower cased and split into tokens, based on spaces and punctuation signs.
- **Slang replacement.** 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 the Urban Dictionary[3] and other two slang dictionaries dedicated sites[4].
- **Word normalization.** In the next step, we match the tokens against 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. "greeeeat" becomes "greeeat", "greeat" and finally "great"). The words used in this way are replaced with "intensifier" plus original word as matched from Roget's Thesaurus.

[1] The dictionaries were obtained by mixing three existing resources that have proven to be most precise, although with less recall as others: General Inquirer, LIWC and MicroWNOp.
[2] http://sentistrength.wlv.ac.uk/
[3] http://www.urbandictionary.com/
[4] www.noslang.com/dictionary, www.smsslang.com

- **Affect word matching.** Further on, the tokens in the tweet are matched against the in-house produced lexicon based on three different sentiment dictionaries: General Inquirer, LIWC and MicroWNOp and 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.** 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.
- **User and topic labeling.** 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. OPAL Task 4 A

In subtask A, the participating systems were supposed to classify a set of tweets in three classes, according to the polarity of the sentiment they conveyed: positive, negative or neutral.
To tackle this task, we used an unsupervised method, based on the identified sentiment words and modifiers. Each of the sentiment words was mapped to four categories, which were given different scores: positive (1), negative (-1), high positive (4) and high negative (-4). The dictionaries have been previously built and the process is described by Balahur et al. (2010). It is based on *WordNet Affect* (Strapparava and Valitutti, 2004), *SentiWordNet* (Esuli and Sebastiani, 2006), *MicroWNOp* (Cerini et al, 2007) and an in-house built resource entitled *JRC Tonality*. This resource has also been used to create the multilingual tonality dictionaries described by Steinberger et al. (2011) and are also implemented in the Europe Media Monitor system. Subsequently, in a window of 6 words around the identified sentiment-bearing token, the following rules were applied:

- When an "intensifier" was present, the value was multiplied with 1.5.
- When a "diminisher" was identified, the value was multiplied with 0.5.
- When a "negator" was identified, the value was multiplied with -1.

Finally, the partial scores obtained for the sentiment contexts were added and normalized by the number of contexts. A positive score led to the text being classified as "positive", a negative score to its being classified as "negative" and a score of 0 labeled as "neutral".

3.3. OPAL Task 4 B

In subtask B, the participating systems were tasked to classify a set of tweets in two classes: positive or negative.
In this task, we employed supervised learning using Support Vector Machines Sequential Minimal Optimization (Platt, 1998) with a binomial kernel, employing boolean features - the presence or absence of unigrams and bigrams determined from the training data (tweets that were previously 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 trained and tested the approach using the Weka data mining software[5], on the data provided for training by the organizers.

4 Evaluation

In the SemEval 2016 Task 4, the OPAL system obtained the following results (Nakov et al., 2016):
In subtask A, 0.50521 average F1, 0.56020 average Recall and 0.54122 accuracy.
In subtask B, OPAL scored 0.61617 average Recall, 0.63316 average F1 and 0.79215 accuracy.

5 Conclusions and Future Work

In this paper, we presented the two approaches to classify tweets according to their polarity, being simple enough to be ported to different languages, domains and deal with documents written in different styles and diverse in formality levels. The performance levels are promising given the simplicity of the implementations. As such, our next challenge resides in adding and evaluating new and simple processing components that can be added for specific languages and domains, in order to increase the classification performance while at the same time keeping the wide usability and reliability of the system.

References

1. Balahur, A. (2013). Sentiment analysis in social media texts. Proceedings of the 4th workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 120-128, Association for Computational Linguistics.
2. Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, Halkia, M.,

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

Pouliquen, B. Belyaeva, J. (2010) Sentiment Analysis in the News, Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 2216-2220. Valletta, Malta, 19-21 May 2010.

3. Cerini, S. , V. Compagnoni, A. Demontis, M. Formentelli and G. Gandini. (2007). Language resources and linguistic theory: Typology, second language acquisition, English linguistics, chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano, IT. 2007.
4. Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available resource for opinion mining. In Proceedings of the 6th International Conference on Language Resources and Evaluation. ELRA.
5. Go, A., Richa Bhayani, and Lei Huang (2009). Twitter sentiment classification using distant supervision. Processing, pages 1–6.
6. Jiang, L., Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao (2011). Target-dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – Volume 1, HLT '11, pages 151–160, Stroudsburg, PA, USA. Association for Computational Linguistics.
7. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V. (2016) SemEval-2016 Task 4: Sentiment Analysis in Twitter. Proceedings of SemEval 2016, San Diego, USA, 2016.
8. Pak, A. and Patrick Paroubek (2010). Twitter as a corpus 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 conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta; ELRA, may. European Language Resources Association, pp. 19-21.
9. Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. In Advances in kernel methods, Eds. Bernhard Schölkopf, Christopher J. C. Burges, Alexander J. Smola, ISBN 0-262-19416-3, pp. 185-208, MIT Press.
10. Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. MIT Press, Cambridge, MA, USA, pp. 185–208. URL http://dl.acm.org/citation.cfm?id=299094.299105
11. Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL Student Research Workshop, ACL student '05, pages 43–48, Stroudsburg, PA, USA. Association for Computational Linguistics.
12. Steinberger, J., Lenkova, P., Kabadjov, M., Steinberger, R., Van Der Goot, E. (2011) Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora. In Proceedings of the conference on Recent Advances in Natural Language Processing (RANLP 2011), pages 770-775.
13. Strapparava, C. and Valitutti, A. (2004) WordNet-Affect: an affective extension of WordNet. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May 2004, pp. 1083-1086.