Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data

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

Marco Turchi
Fondazione Bruno Kessler-IRST
Via Sommarive, 18
Povo, Trento, Italy
turchi@fbk.eu

Abstract

Sentiment analysis is currently a very dynamic field in Computational Linguistics. Research herein has concentrated on the development of methods and resources for different types of texts and various languages. Nonetheless, the implementation of a multilingual system that is able to classify sentiment expressed in various languages has not been approached so far. The main challenge this paper addresses is sentiment analysis from tweets in a multilingual setting. We first build a simple sentiment analysis system for tweets in English. Subsequently, we translate the data from English to four other languages - Italian, Spanish, French and German - using a standard machine translation system. Further on, we manually correct the test data and create Gold Standards for each of the target languages. Finally, we test the performance of the sentiment analysis classifiers for the different languages concerned and show that the joint use of training data from multiple languages (especially those pertaining to the same family of languages) significantly improves the results of the sentiment classification.

1 Introduction

Sentiment analysis is a task in Natural Language Processing whose aim is to automatically detect and classify sentiments in texts. Generally, the "positive", "negative" and "neutral" classes are considered, although other scales have also been used (e.g. from 1 to 5 "stars" - according to the reviewing systems put at the disposal of clients or users by amazon.com, booking.com, etc.; adding the "very positive" and "very negative" classes, scales from 1 to 10, etc.).

In this article, we deal with the issue of sentiment analysis in tweets, in a multilingual setting. We employ machine translation - which was shown to be at a sufficiently high level of performance (Balahur and Turchi, 2012) - to obtain data in four languages. Our goal is to test if the use of multilingual data can help to improve sentiment classification in tweets (as shown to be the case in formal texts - (Banea et al., 2010)) and if the joint use of data coming from similar languages or languages that are different in structure can influence on the final result.

The main problem when designing automatic methods for the treatment of tweets is that they are highly informal texts, i.e. they contain slang, emoticons, repetitions of letters or punctuation signs, misspellings (done on purpose or due to writing them from mobile devices), entire words in capital letters, etc.

In order to test our hypotheses, we first design a simple tweet sentiment analysis system for English, taking into account the specificity of expressions employed, but without using language-specific text processing tools. The motivation is related to the fact that: a) such a distinction would require the use of language identifiers and would need the data from the different languages to be separated; b) We would like to apply the same techniques for as many languages as possible and for some of these languages, no freely-available language processing tools exist. We test this system on the SemEval 2013 Task 2 - Sentiment Analysis in Twitter (Wilson et al., 2013) - training data and test on the development data. The choice of this test set was motivated by the fact that it contains approximately 1000 tweets, being large enough to be able to draw relevant conclusions and at the same time small enough to allow manual correction of the translations, to eliminate incorrect translations being present in both training and test data.

Proceedings of Recent Advances in Natural Language Processing, pages 49–55,
Hissar, Bulgaria, 7-13 September 2013.

Subsequently, we employ the Google machine translation system[1] to translate the SemEval 2013 training and development tweets in Italian, Spanish, German and French. We manually correct the translated development data (which we use for testing, not for parameter tuning) to produce a reliable Gold Standard.

Finally, we apply the same sentiment classification system to each of these languages and test the manner in which the combined datasets (from pairs of two languages, families of languages and all the languages together) perform. We conclude that the joint use of training data from different languages improves the classification of sentiment and that the use of training data from languages that are similar in structure helps to achieve statistically significant improvements over the results obtained on individual languages and all languages together.

The remainder of this article is structured as follows: Section 2 gives an overview of the related work. In Section 3, we present the motivations and describe the contributions of this work. In the following section, we describe in detail the process followed to pre-process the tweets, build the classification models and obtain tweets for four other languages. In Section 5, we present the results obtained on different languages and combinations thereof. Finally, Section 6 summarizes the main findings of this work and sketches the lines for future work.

2 Related Work

The work described herein is related to the development of multilingual sentiment analysis systems and sentiment classification from tweets.

2.1 Methods for Multilingual Sentiment Analysis

In order to produce multilingual resources for subjectivity analysis, Banea et al. (Banea et al., 2008) apply bootstrapping to build a subjectivity lexicon for Romanian, starting with a set of 60 words which they translate and subsequently filter using a measure of similarity to the original words, based on Latent Semantic Analysis (LSA) (Deerwester et al., 1990) scores. Another approach to mapping subjectivity lexica to other languages is proposed by Wan (2009), who uses co-training to classify un-annotated Chinese reviews using a corpus of annotated English reviews. (Kim et al., 2010) create a number of systems consisting of different subsystems, each classifying the subjectivity of texts in a different language. They translate a corpus annotated for subjectivity analysis (MPQA), the subjectivity clues (Opinion Finder) lexicon and re-train a Naive Bayes classifier that is implemented in the Opinion Finder system using the newly generated resources for all the languages considered. (Banea et al., 2010) translate the MPQA corpus into five other languages (some with a similar ethimology, others with a very different structure). Subsequently, they expand the feature space used in a Naive Bayes classifier using the same data translated to 2 or 3 other languages. Finally, (Steinberger et al., 2011a; Steinberger et al., 2011b) create sentiment dictionaries in other languages using a method called "triangulation". They translate the data, in parallel, from English and Spanish to other languages and obtain dictionaries from the intersection of these two translations.

2.2 Sentiment Classification from Tweets

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 super-

[1] http://translate.google.com/

vised 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).

3 Motivation and Contribution

The work presented herein is mainly motivated by the need to: a) develop sentiment analysis tools for a high number of languages, while minimizing the effort to create linguistic resources for each of these languages in part; b) study the manner in which the use of machine translation systems to produce multilingual data performs in the context of informal texts such as tweets; and c) evaluate the performance of sentiment classification when data from different languages is combined in the training phase. We would especially like to study the effect of using data from similar languages versus the use of data from structurally and lexically-different languages. The advantage of such an approach would be that if combined classifiers perform better, then the effort of separating tweets in different languages at the time of analysis (which in the case of streaming data is not negligeable) can be reduced or eliminated entirely.

Unlike approaches we presented in Related Work section, we employ fully-formed machine translation systems.

Bearing this in mind, the main contributions we bring in this paper are:

1. The creation of a simple tweet sentiment analysis system, that employs a pre-processing stage to normalize the language and generalize the vocabulary employed to express sentiment. At this stage, we take into account the linguistic peculiarities of tweets, regarding spelling, use of slang, punctuation, etc., and also replace the sentiment-bearing words from the training data with a unique label. In this way, the sentence "I love roses." will be equivalent to the sentence "I like roses.", because "like" and "love" are both positive words according to the GI dictionary. If example 1 is contained in the training data and example 2 is contained in the test data, replacing the sentiment-bearing word with a general label increases the chance to have example 2 classified correctly. In the same line of thought, we also replaced modifiers with unique corresponding labels.

2. The use of minimal linguistic processing, which makes the approach easily portable to other languages. We employ only tokenization and do not process texts any further. The reason behind this choice is that we would like the final system to work in a similar fashion for as many languages as possible and for some of them, little or no tools are available.

3. The use of a standard news translation system to obtain data in four other languages - Italian, Spanish, German and French;

4. The evaluation of different combinations of languages in the training phase and the effect of using languages from the same family versus the use of individual or all languages in the training phase on the overall performance of the sentiment classification performance.

We show that using the training models generated with the method described we can improve the sentiment classification performance, irrespective of the domain and distribution of the test sets.

4 Sentiment Analysis in Tweets

Our sentiment analysis system is based on a hybrid approach, which employs supervised learning with the Weka (Weka Machine Learning Project, 2008) implementation of the Support Vector Machines Sequential Minimal Optimization (Platt, 1998) linear kernel, on unigram and bigram features, but exploiting as features sentiment dictionaries, emoticon lists, slang lists and other social media-specific features. 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. They were built using the same dictionaries we employ in this work and their corrected translation to Spanish. The new sentiment dictionaries were created by simultaneously translating from these two languages to a third one and considering the intersection of the translations as correct terms. Currently, new such dictionaries have been created for 15 other languages.

The sentiment analysis process contains two stages: pre-processing and sentiment classification.

4.1 Tweet Pre-processing

The language employed in Social Media sites is different from the one found in mainstream media and the form of the words employed is sometimes not the one we may find in a dictionary. Further on, users of Social Media platforms 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. As such, before applying supervised learning to classify the sentiment of the tweets, we preprocess them, to normalize the language they contain. The pre-processing stage contains the following steps:

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

In the second step of the pre-processing, we employ the annotated list of emoticons from SentiStrength[2](Thelwall et al., 2010) 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.

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

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 from a specialized site [3].

At this stage, the tokens are compared to entries in Rogets 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. "perrrrrrrrrrrrrrrrrfeeect" becomes "perrfeect", "perfeect", "perrfect" and subsequently "perfect"). The words used in this form are maked as "stressed".

Further on, the tokens in the tweet are matched against three different sentiment lexicons: GI, 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". A version of the data without these replacements is also maintained, for comparison purposes.

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. As in the case of affective words, a version of the data without these replacements is also maintained, for comparison purposes.

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".

4.2 Sentiment Classification of Tweets

Once the tweets are pre-processed, they are passed on to the sentiment classification module. We employed supervised learning using SVM SMO with a linear kernel, based on boolean features - the presence or absence of n-grams (unigrams, bigrams and unigrams plus bigrams) determined from the training data (tweets that were previousely pre-processed as described above). Bigrams are used specifically to spot the influence

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

[3]http://www.chatslang.com/terms/social_media

of modifiers (negations, intensifiers, diminishers) on the polarity of the sentiment-bearing words.

4.3 Obtaining Multilingual Data for Sentiment Analysis in Tweets

Subsequent to the tweet normalization, we translate the Twitter data (the training and development data in the SemEval Task 2 campaign) using the Google machine translation system to four languages - Italian, Spanish, French and German. The reason for choosing the development dataset for testing is that this set is smaller and allows us to manually check and correct it, to obtain a Gold Standard (and ensure that performance results are not biased by the incorrect translation in both the training, as well as the development data).

Further on, we extract the same features as in the case of the system working for English - unigrams and bigrams - from these obtained datasets. We employ the features to train an SVM SMO classifier, in the same manner as we did for English.

5 Evaluation and Discussion

Although the different steps included to eliminate the noise in the data and the choice of features have been refined using our in-house gathered Twitter data, in order to evaluate our approach and make it comparable to other methods, we employ the data used in an established competition, allowing subsequent comparisons to be made.

5.1 Data Set

The characteristics of the training (T*) and development (test in our case) - t*- datasets employed are described in Table 1. On the last column, we also include the baseline in terms of accuracy, which is computed as the number of examples of the majoritary class over the total number of examples:

Data	#Tweet	#Pos.	#Neg.	#Neu.	Bl%
T*	6688	2450	956	3282	49%
t*	1051	386	199	466	44%

Table 1: Characteristics of the training (T*) and testing (t*) datasets employed.

5.2 Evaluation and Results

In order to test our sentiment analysis approach, we employed the datasets described above, for each of the languages individually, all the two-languages combinations, combinations of languages from the same linguistic family and all languages together.

The results are presented in Table 2. We consider the measure of accuracy and do not compare to the SemEval official results, because in the competition, the results did not take into account the "neutral" class.

Language(s)	Accuracy
English	64.75
Italian	60.12
French	62.31
German	61.32
Spanish	62.66
English + French	65.91
English + German	63.98
English + Italian	64.78
English + Spanish	68.23
Spanish + Italian	70.45
Spanish + French	67.14
Spanish + German	65.64
Italian + German	63.29
Italian + French	63.95
German + French	62.66
Italian + French + Spanish	68.53
All 5 languages	69.09

Table 2: Results obtained classifying each language individually versus on pairs and families of languages, respectively.

5.3 Discussion

From the results obtained, we can draw several conclusions.

First of all, we can see that using tweet normalization and employing machine translation, we can obtain high quality training data for sentiment analysis in many languages. The machine-translated data thus obtained can be reliably employed to build classifiers for sentiment, reaching a performance level that is similar to the results obtained for English and significatly above the baseline.

Secondly, seeing the performance of the different pairs of languages compared to individual results, we can: a) on the one hand, see that combining languages with a comparatively high difference in performance results in an increase of the lower-performing one and b) on the other hand, in

some cases, the overall performance is improved on both systems, which shows that combining this data helps to disambiguate the contextual use of specific words.

Finally, the results show that the use of all the languages together improves the overall classification of sentiment in the data. This shows that a multilingual system can simply employ joint training data from different languages in a single classifier, thus making the sentiment classification straightforward, not needing any language detection software or training different classifiers.

By manually inspecting some of the examples in the datasets, we could see that the most important causes of incorrect classification were the word orders and faulty translations in context. Another reason for incorrect sentiment classification was the different manner in which negation is constructed in the different languages considered. In order to improve on this aspect, we will include language-specific rules by adding skip-bigrams (bigrams made up of non-consecutive tokens) features in the languages where the place of the negators can vary.

6 Conclusions and Future Work

In this article, we presented a method to create a simple sentiment analysis system for English and extend it to the multilingual setting, by employing a standard news machine translation system. We showed that using twitter language normalization, we can obtain good results in target languages and that the joint use of training data from different languages helps to increase the overall performance of the classification. Finally, we showed that the joint training using translated data from languages that are similar yield significantly improved results.

In future work, we plan to evaluate the use of higher-order n-grams (3-grams) and skip-grams to extract more complex patterns of sentiment expressions and be able to identify more precisely the scope of the negation. In this sense, we plan to take into account the modifier/negation schemes typical of each of the languages, to consider (further to translation) language-specific schemes of n-grams.

We also plan to test the performance of sentiment classification using translations *to* English and employing classifiers trained on English data. In order to do this, we require lists of slang and digital dictionaries to perform normalization. We would like to study the performance of our approach in the context of tweets related to specific news, in which case these short texts can be contextualized by adding further content from other information sources. In this way, it would be interesting to make a comparative analysis of the tweets written in different languages (from the same or different regions of the globe), on the same topics.

References

Alexandra Balahur and Marco Turchi. 2012. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60, Jeju, Korea, July. Association for Computational Linguistics.

C. Banea, R. Mihalcea, and J. Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of the Conference on Language Resources and Evaluations (LREC 2008), Maraakesh, Marocco*.

C. Banea, R. Mihalcea, and J. Wiebe. 2010. Multilingual subjectivity: are more languages better? In *Proceedings of the International Conference on Computational Linguistics (COLING 2010), Beijing, China.*, pages 28–36.

S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 3(41).

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

Long Jiang, 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.

J. Kim, J.-J. Li, and J.-H. Lee. 2010. Evaluating multilanguage-comparability of subjectivity analysis systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 595–602.

Alexander Pak 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. 19-21.

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

Jonathon Read. 2005. 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. Association for Computational Linguistics.

J. Steinberger, P. Lenkova, M. Ebrahim, M. Ehrman, A. Hurriyetoglu, M. Kabadjov, R. Steinberger, H. Tanev, V. Zavarella, and S. Vazquez. 2011a. Creating sentiment dictionaries via triangulation. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Portland, Oregon.

J. Steinberger, P. Lenkova, M. Kabadjov, R. Steinberger, and E. van der Goot. 2011b. Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In *Proceedings of the Conference on Recent Advancements in Natural Language Processing (RANLP)*, Hissar, Bulgaria.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, December.

Weka Machine Learning Project. 2008. Weka. URL http://www.cs.waikato.ac.nz/˜ml/weka.

Cynthia Whissell. 1989. 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.

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.

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