REACTION: A naive machine learning approach for sentiment classification

Silvio Moreira
IST/INESC-ID
Rua Alves Redol, 9
1000-029 Lisboa
Portugal
samir@inesc-id.pt

João Filgueiras
INESC-ID
Rua Alves Redol, 9
1000-029 Lisboa
Portugal
jfilgueiras@inesc-id.pt

Bruno Martins
IST/INESC-ID
Rua Alves Redol, 9
1000-029 Lisboa
Portugal
bruno.g.martins@ist.utl.pt

Francisco Couto
LASIGE - FCUL
Edifício C6 Piso 3
Campo Grande
1749 - 016 Lisboa
Portugal
fcouto@di.fc.ul.pt

Mário J. Silva
IST/INESC-ID
Rua Alves Redol, 9
1000-029 Lisboa
Portugal
mjs@inesc-id.pt

Abstract

We evaluate a naive machine learning approach to sentiment classification focused on Twitter in the context of the sentiment analysis task of SemEval-2013. We employ a classifier based on the Random Forests algorithm to determine whether a tweet expresses overall positive, negative or neutral sentiment. The classifier was trained only with the provided dataset and uses as main features word vectors and lexicon word counts. Our average F-score for all three classes on the Twitter evaluation dataset was 51.55%. The average F-score of both positive and negative classes was 45.01%. For the optional SMS evaluation dataset our overall average F-score was 58.82%. The average between positive and negative F-scores was 50.11%.

1 Introduction

Sentiment Analysis is a growing research field, especially on web social networks. In this setting, users share very diverse messages such as real-time reactions to news, events and daily experiences. The ability to tap on a vast repository of opinions, such as Twitter, where there is great diversity of topics, has become an important goal for many different applications. However, due to the nature of the text, NLP systems face additional challenges in this context. Shared messages, such as tweets, are very short and users tend to resort to highly informal an noisy speech.

Following this trend, the 2013 edition of SemEval[1] included a sentiment analysis on Twitter task (SemEval-2013 Task 2). Participants were asked to implement a system capable of determining whether a given tweet expresses positive, negative or neutral sentiment. To help in the development of the system, an annotated training corpus was released. Systems that used only the given corpus for training were considered *constrained*, while others were considered *unconstrained*. The submitted prototypes were evaluated in a dataset consisting of around 3700 tweets of several topics. The metric used was the average F-score between the positive and negative classes.

Our goal with this participation was to create a baseline system from which we can build upon and perform experiments to compare new approaches with the state-of-the-art.

2 Related Work

The last decade saw a growing interest in systems to automatically process sentiment in text. Many approaches to detect subjectivity and determine

[1] Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 490–494, Atlanta, Georgia, June 14-15, 2013.

polarity of opinions in news articles, weblogs and product reviews have been proposed (Pang et al., 2002; Pang et al., 2004; Wiebe et al., 2005; Wilson et al., 2005). This sub-field of NLP, known as Sentiment Analysis is presented in great depth in (Liu, 2012).

The emergence and proliferation of microblog platforms created a medium where people express and convey all kinds of information. In particular, these platforms are a rich source of subjective and opinionated text, which has motivated the application of similar techniques to this domain. However, in this context, messages tend to be very short and highly informal, full of typos, slang and unconventional spelling, posing additional challenges to NLP systems. In fact, early experiments in Sentiment Analysis in the context of Twitter (Barbosa et al., 2010; Davidov et al., 2010; Koulompis et al., 2011; Pak et al., 2010; Bifet et al., 2010) show that the techniques that proved effective in other domains are not sufficient in the microblog setting. In the spirit of these approaches, we included a preprocessing step, followed by feature extraction focusing on word, lexical and Twitter-specific features. Finally, we use annotated data to train an automatic classifier based on the Random Forests (Breiman, 2001) and BESTrees (Sun et al., 2011) learning algorithms.

3 Resources

Two annotated datasets were made available to participants of SemEval-2013 Task 2: one for training purposes which was to contain 8000 to 12000 tweets; and another, for development, containing 2000. The combined datasets ended up amounting to a little over 7500 tweets. The distribution of positives, negatives and neutrals for the combined datasets can be found in Table 1. Nearly half of all tweets belonged to the neutral class, and negatives represent just 15% of these datasets.

Class	Number
Positive	37%
Negative	15%
Neutral	48%

Table 1: Class distribution of annotated data.

Random examples of each class drawn from the datasets are shown in Table 2.

Positive:
1 *Louis inspired outfit on Monday and Zayn inspired outfit today..4/5 done just need Harry*
2 *waking up to a Niners win, makes Tuesday get off to a great start! 21-3 over the cards and 2 games clear in the NFC West.*
Negative:
3 *Sitting at home on a Saturday night doing absolutely nothing... Guess I'll just watch Greys Anatomy all night. #lonerproblems #greysanatomy*
4 *Life just isn't the same when there is no Pretty Little Liars on Tuesday nights.*
Neutral:
5 *Won the match #getin . Plus, tomorrow is a very busy day, with Awareness Day's and debates. Gulp. Debates*
6 *@_Nenaah oh cause my friend got something from china and they said it will take at least 6 to 8 weeks and it came in the 2nd week :P*

Table 2: Random examples of annotated tweets.

4 Approach

Given our goal of creating a baseline system, we experimented with a common set of features used in sentiment analysis. The messages were modelled as a combination of binary (or presence) unigrams, lexical features and Twitter-specific features. We decided to follow a supervised approach by learning a Random Forests classifier from the annotated data provided by the organisers of the workshop (see Section 3). In summary, the development of our system consisted of four steps: 1) preprocessing of the data, 2) feature extraction, 3) learning the classifier, and 4) applying the classifier to the test set.

4.1 Preprocessing

The lexical variation introduced by typos, abbreviations, slang and unconventional spelling, leads to very large vocabularies. The resulting

sparse vector representations with few non-zero values hamper the learning process. In order to tackle this problem, we replaced user mentions (@<username>) with a fixed tag <USER> and URLs with the tag <URL>. Then, each sentence was normalised by converting to lower-case and reducing character repetitions to at most 3 characters (e.g. *"heelloooooo!"* would be normalised to *"heellooo!"*). Finally, we performed the lemmatisation of the sentence using the Morphadorner[2] software.

4.2 Feature Extraction

After the preprocessing step, we extract a vector consisting of the top uni-grams present in the training set and represent individual messages in terms of this vector. For each message we also compute the frequency of smileys and words with prior sentiment polarity using a sentiment lexicon. Finally, we include the harmonic mean of positive and negative words. Next we explain each feature in more detail.

Word vector: a sparse word vector containing the top 25.000 most frequent words that occur in the training set. This feature aims at capturing relations between certain words and overall message polarity. The vector was extracted using the Weka toolkit (Hall et al., 2009) with the stop word list option.

Lexicon word count: positive and negative sentiment word counts. When the word is preceded by a negation particle we invert the polarity. We used Bing Liu's Opinion Lexicon[3] that includes 2006 positive and 4783 negative words and is especially tailored for social media because it considers misspellings, slang and other domain specific variations.

Smileys count: a count of positive and negative smileys that appear in the tweet. We take advantage of these constructs being especially indicative of the overall expressed sentiment in a text (Davidov et al., 2010). Although there are smiley lexicons, such as the one used on SentiStrength[4], we used regular expressions to capture most common smileys in a flexible way.

Hashtag count: a count of positive and negative hashtags. This feature also uses Bing Liu's lexicon to determine wether a word contained in an hashtag is positive or negative. The rationale behind this feature is that positive or negative words in the form of hashtags can have a stronger meaning than regular words (Davidov et al., 2010).

Positive/negative harmonic mean: harmonic mean between positive and negative token counts, including words and hashtags.

In an attempt to further reduce the dimensionality of the feature space we computed the principal components of the word vector using the Principal Components Analysis filter in Weka but observed that this yielded worse results.

4.3 Learning the classifier

To implement our classifier we used the Weka machine learning framework and experimented with two ensemble algorithms: Random Forests and BESTrees. We eventually dropped the use of BESTrees as initial results were worse.

We attempted to use most of the data while being able to effectively measure the performance of the classifier. Therefore we used the totality of both sets for training and evaluated using 10 fold cross-validation.

Since we used only the annotated dataset that was provided for this task, our approach is considered constrained.

5 Results

Our results with 10 fold cross-validation using the submitted classifier, are presented in Table 3.

Class	Precision	Recall	F-score
positive	61.0%	63.9%	62.4%
negative	54.1%	26.8%	35.8%
neutral	**64.7%**	**72.4%**	**68.3%**
average F-score (pos/neg)			49.1%

Table 3: Cross-validation results using the training set.

Task evaluation results are presented in Table 4 for tweets. Our approach ranked 44th out of 48 participants. The evaluation dataset had a similar class distribution to the annotated datasets,

[2] http://morphadorner.northwestern.edu/

[3] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

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

with almost half being neutral, and just 14% negative. Preliminary results with cross-validation were similar to those of the final evaluation for Twitter.

Class	Precision	Recall	F-score
positive	**62.52%**	55.28%	58.68%
negative	55.74%	21.80%	31.34%
neutral	56.54%	**75.43%**	**64.63%**
average F-score (pos/neg)			45.01%

Table 4: Task evaluation results for Tweets.

Also included in SemEval-2013 Task 2 was an evaluation using a SMS dataset to understand if a classifier trained using tweets could be applied to SMS messages. SMS results are shown in Table 5. In this case our approach ranked 23th out of 42 participants. The SMS evaluation dataset was composed of more than half neutral messages (58%), and similarly distributed positives (23%) and negatives (19%).

Class	Precision	Recall	F-score
positive	53.66%	59.50%	56.45%
negative	60.54%	34.26%	43.76%
neutral	**72.91%**	**79.90%**	**76.27%**
average F-score (pos/neg)			50.11%

Table 5: Task evaluation results for SMS.

6 Discussion and Conclusions

As expected, our naive approach performs poorly in the context of Twitter messages. The obtained results are in line with similar approaches described in the literature and we found that Random Forests achieve the same performance as other learning algorithms tried for the same task (Koulompis et al., 2011).

The uneven distribution of classes in the data may have also contributed to the low performance of the classifier. Although the neutral class was not considered in the evaluation, the datasets had a great predominance of neutral messages whereas the negative examples only accounted for 15% of the corpus. This suggests that it could be useful to use a minority class over-sampling method, such as SMOTE (Chawla, 2002), to reduce the effect of this imbalance on the data. We used n-grams to model the words that compose each message. However, this approach leads to very sparse representations, thus becoming important to consider techniques that reduce feature space. We experimented with PCA, without success, but we still believe that applying feature selection algorithms or denser word representations (Turian et al., 2010) could improve performance in this task.

We find that our classifier performs better on the SMS dataset. This might be explained by the fact that SMS messages tend to be more direct, whereas the same tweet can express, or show signs of, contradictory sentiments. In fact, our naive approach outperforms other systems that had better results in the Twitter dataset, but it is difficult to say why, given that we do not have access to the SMS test set annotations.

Despite the poor ranking results, we achieved our goal of performing basic experiments in the task of sentiment analysis in Twitter and developed a baseline system that will serve as a starting point for future research.

Acknowledgments

This work was partially supported by FCT (Portuguese research funding agency) under project grants UTA-Est/MAI/0006/2009 (REACTION) and PTDC/CPJ-CPO/116888/2010 (POPSTAR). FCT also supported scholarship SFRH/BD/89020/2012. This research was also funded by the PIDDAC Program funds (INESC-ID multi annual funding) and the LASIGE multi annual support.

References

Barbosa, L., and Feng, J. 2010. *Robust sentiment detection on twitter from biased and noisy data.* Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 36-44.

Bifet, A., and Frank, E. 2010. *Sentiment knowledge discovery in twitter streaming data.* Discovery Science.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1), 5-32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. 2002 *SMOTE: synthetic minority*

over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

Davidov, D., Tsur, O., and Rappoport, A. 2010 *Enhanced sentiment learning using twitter hashtags and smileys*. Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Pages 241-249. Association for Computational Linguistics.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. 2009. *The WEKA Data Mining Software: An Update* SIGKDD Explorations, Volume 11, Issue 1.

Kouloumpis, E., Wilson, T., and Moore, J. 2011. *Twitter sentiment analysis: The good the bad and the omg.* Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 538541.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining.* Synthesis Lectures on Human Language Technologies, 5(1), 1167.

Pak, A., and Paroubek, P. 2010. *Twitter as a corpus for sentiment analysis and opinion mining.* Proceedings of LREC.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. *Thumbs up?: sentiment classification using machine learning techniques.* Proceedings of the ACL-02 conference on Empirical methods in natural language processing. Volume 10, pp. 79-86. Association for Computational Linguistics.

Pang, B. and Lee, L. 2004. *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.* Proceedings of the 42nd annual meeting on Association for Computational Linguistics.

Sun, Q. and Pfahringer, B. 2011. *Bagging Ensemble Selection.* Proceedings of the 24th Australasian Joint Conference on Artificial Intelligence (AI'11), Perth, Australia, pages 251-260. Springer.

Turian, J., Ratinov, L., and Bengio, Y. 2010. *Word representations: a simple and general method for semi-supervised learning*. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 384-394). Association for Computational Linguistics.

Wiebe, J. and Riloff, E. 2005. *Creating subjective and objective sentence classifiers from unannotated texts.* Computational Linguistics and Intelligent Text Processing, pages 486-497, Springer.

Wilson, T., Wiebe, J., and Hoffmann, P. 2005. *Recognizing contextual polarity in phrase-level sentiment analysis.* Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 347-354. Association for Computational Linguistics.