CICBUAPnlp at SemEval-2016 Task 4-A: Discovering Twitter Polarity using Enhanced Embeddings

Helena Gómez-Adorno, Grigori Sidorov
Center for Computing Research
Instituto Politécnico Nacional
Av. Juan de Dios Bátiz
C.P. 07738, Mexico City, Mexico
helena.adorno@gmail.com
sidorov@cic.ipn.mx

Darnes Vilariño, David Pinto
Faculty of Computer Science
Benemérita Universidad Autónoma de Puebla
Av. San Claudio y 14 sur
C.P. 72570, Puebla, Mexico
darnes@cs.buap.mx
dpinto@cs.buap.mx

Abstract

This paper presents our approach for SemEval 2016 task 4: Sentiment Analysis in Twitter. We participated in Subtask A: Message Polarity Classification. The aim is to classify Twitter messages into positive, neutral, and negative polarity. We used a lexical resource for pre-processing of social media data and train a neural network model for feature representation. Our resource includes dictionaries of slang words, contractions, abbreviations, and emoticons commonly used in social media. For the classification process, we pass the features obtained in an unsupervised manner into an SVM classifier.

1 Introduction

In this paper, we describe our approach for the SemEval 2016 task 4 "Sentiment Analysis in Twitter" subtask A (Nakov et al., 2016), where the goal is to classify a tweet message as either positive, neutral, or negative. The main goal of our approach is to improve the feature representation obtained by a well-known neural network method–Doc2vec (Le and Mikolov, 2014), using dictionaries of abbreviations, contractions, slang words, and emoticons.

Approaches based on neural networks for unsupervised feature representation (or embeddings) often do not perform data cleaning (Le and Mikolov, 2014; Socher et al., 2011), considering that the network itself would solve the related problems. These approaches treat special characters such as ,.!?# and user mentions as a regular word (Le and Mikolov, 2014; Brigadir et al., 2014). Still, in some works which use embeddings a basic data cleaning process (i.e., stopwords removal, URL filtering, and removal of rare terms) improves the feature representation and, consequently, the performance of the classification task (Yan et al., 2014; Rangarajan Sridhar, 2015; Jiang et al., 2014).

The problem with the content of social media messages is that they usually have a lot of non-standard language expressions (Pinto et al., 2012; Atkinson et al., 2013). Due to the short nature of the messages, most of the users use a large vocabulary of slang words, abbreviations, and emoticons (Das and Bandyopadhyay, 2011). Slang words are not considered as a part of the standard vocabulary of a language, and they are mostly used in informal messages, while abbreviations are shortened forms of a word or name that are used in order to replace the full forms. Emoticons usually convey the current feeling of the message writer.

For this task we propose a preprocessing phase using the dictionaries that we previously built for the task of Authorship Atribution (Posadas-Durán et al., 2015). These dictionaries are useful for preprocessing and cleaning messages obtained from several social networks, such as Facebook, Google+, Instagram, etc.

The rest of the paper is structured as follows. Section 2 describes related work. Section 3 introduces the social media lexical resource used for this work. Section 4 presents our proposed approach. Section 5 presents the evaluation of the task using the neural network based feature representation. Finally, Section 6 draws the conclusions from our experiments and points out the possible directions of future work.

Proceedings of SemEval-2016, pages 145–148,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

2 Related Work

There are many works that tackle the problem of social media texts pre-processing (Baldwin, 2012; Clark and Araki, 2011; Das and Bandyopadhyay, 2011); however, to the best of our knowledge, the research based on neural network for feature representation did not consider the effect that data cleaning have on the quality of the representation (specially on social media data).

Several approaches have been proposed for vector-space distributed representations of words and phrases. These models are used mainly for predicting a word given a surrounding context. However, most of the authors indicate that distributed representations of words and phrases can also capture syntactic and semantic similarity or relatedness (Le and Mikolov, 2014; Socher et al., 2013; Mikolov et al., 2013). This particular behaviour makes these methods attractive to solve several NLP tasks, nevertheless, at the same time, it raises new issues, such as dealing with unnormalized texts, which are typically present in social media forums such as Twitter, Facebook, Instagram, among others. Researchers have proposed several pre-processing steps in order to overcome this issue, which led to an overall performance increase. Yan et al. (Yan et al., 2014) obtained almost 2% increase using standard NLP pre-processing, which consists in tokenization, lowercasing, removing stopwords and rare terms. Kumar et al. (Rangarajan Sridhar, 2015) focused on the spelling issues in social media messages, which includes repeated letters, omitted vowels, use of phonetic spellings, substitution of letters with numbers (typically syllables), use of shorthands and user created abbreviations for phrases. In a data-driven approach, Brigadir et al. (Brigadir et al., 2014) apply URL filtering combined with standard NLP preprocessing techniques.

3 Resources

We developed the dictionaries with the aim of preprocessing tweets for the author profiling task at PAN 2015 (Posadas-Durán et al., 2015). First, we reviewed the tweets present in the PAN corpus and found excessive use of shortened vocabulary, which can be divided into three categories: slang words, abbreviations, and contractions. Moreover, we came across a large number of emoticons, which are a typographic display of a facial representation.

Table 1: Number of entries of the English dictionary

Type of Dictionary	English
Abbreviations	1,346
Contractions	131
Slang words	1,249
Emoticons	482
Total	3,208

The lexical resource was originally built for 4 languages, but for the purposes of this work we only use the English dictionary. The statistics for the English dictionary are presented in Table 1. The dictionaries are freely available on our website[1].

4 Approach to Sentiment Classification

From a machine learning point of view, the Message Polarity Classification task can be considered as a supervised multi-class classification problem, where a set of tweets $\mathbf{T} = \{t_1, t_2, \ldots, t_i\}$ is given, and each sample is assigned to one of the target classes $\{positive, negative, neutral\}$. So, the problem is to build a classifier F that assigns a sentiment class to unclassified tweets.

Since the tweets are very noisy, we perform the preprocessing over each dataset (train, unlabeled and test). In the preprocessing phase, we executed the following steps:

Expand slang words and abbreviations Not all tweets use slang words and abbreviation in the same way. There are Twitter users that do not use slang words and due to this reason we expanded all slang words and abbreviations with their full meaning using the dictionaries described in section 3.

Remove url ULR do not provide information about the sentiment of the tweet and because of this reason every ULR is removed from the text.

Remove hashtags symbols Hashtags in tweets carry useful information about the topic and

[1] http://www.cic.ipn.mx/~sidorov/lexicon.zip

polarity of the message. We only remove the hashtag symbol, keeping the words.

Remove emoticons In order to obtain a distributed representation of a tweet, we used only words and punctuation symbols. So, unlike traditional preprocessing for sentiment analysis we removed the emoticons from tweets by looking up in our emoticons dictionary.

For training, a vector representation of each tweet is obtained in an unsupervised manner by a neural network based model, i.e., $v^i = \{v_1, v_2, \ldots, v_j\}$ where v^i is the vector representation of the tweet t_i. In order to obtain the vector representation of the tweets, a neural network based distributed representation model is trained using the doc2vec algorithm (Le and Mikolov, 2014). It is an unsupervised algorithm that aggregates all the words in a sentence (of variable length) into a vector of fixed length. The algorithm takes into account the contexts of words, and it is able to capture the semantics of the input texts. We used a freely available implementation of doc2vec included in the *Gensim*[2] python module. The doc2vec model is trained with both labeled and unlabeled tweets in order to learn the distributed representation. The learned vector representations have 300 dimentions, we set the windows size to 3 and minimal word frequency is set to 2. Then, a classifier is trained using the vector representations of the labeled tweets. We perform the experiments with the SVM liblinear classifier (Fan et al., 2008), especifically the *LinearSVC* algorithm the implemented in the *Scikit Learn*[3] python module with default parameters.

For the evaluation, the vector representations of the test tweets are obtained retraining the doc2vec model built in the training stage, plus the test tweets. Finally, the vector representation of the tweets are passed to the SVM model in order to assign the corresponding polarity label to each tweet.

We used the train set of SemEval-2014 Task 9: Sentiment Analysis in Twitter - subtask B (Rosenthal et al., 2014), consisting of 6124 tweets (removing the tweets with the objective class). Besides, we expanded the training set with some tweets of this year training set (the ones we could download) and with Stanford Sentiment Analysis Dataset (Go et al., 2009). So, in total we employed 11377 classified tweets for training. For the neural network based feature representation we used the 1.7 millons unlabeled tweets for training the Doc2Vec model.

[2] https://radimrehurek.com/gensim/

[3] http://scikit-learn.org/stable/index.html

5 Results

In this section we present the results obtained in the competition when various test datasets are used. The evaluation metric used in the competition is the macro-averaged F measure calculated over the positive and negative classes. Table 2 presents the overall performance of our approach for different datasets. It can be observed that our approach overcome the baseline for almost all datasets.

Table 2: Obtained results for 2016 Test and Progress

Year	Corpus	Ours	Baseline score
2013	Tweet	0.194	0.292
	SMS	0.193	0.190
2014	Tweet	0.335	0.346
	Tweet Sarcasm	0.393	0.277
	Live-Journal	0.326	0.272
2015	Tweet	0.303	0.303
2016	Tweet	0.303	0.255

6 Conclusions

We presented our results for sentiment analysis on Twitter. We rely on a supervised approach, which is based on top of a deep learning system enhanced with special preprocesing techniques using a lexical social media resource. We reported the overall accuracy for the sentiment classification task in three classes: positive, negative and neutral.

In the future, we will improve our preprocessing phase by removing the target mentions, numbers and repeated sequences of characters.

Acknowledgments

This work was done under the support of the "Red Temática en Tecnologías del Lenguaje", Mex-

ican Government (CONACYT project 240844, SNI, COFAA-IPN and SIP-IPN 20151406, 20161947).

References

John Atkinson, Alejandro Figueroa, and Claudio Pérez. 2013. A semantically-based lattice approach for assessing patterns in text mining tasks. *Computación y Sistemas*, 17(4):467–476.

Timothy Baldwin. 2012. Social media: Friend or foe of natural language processing? In *The 26th Pacific Asia Conference on Language, Information and Computation*, pages 58–59.

I. Brigadir, D. Greene, and P. Cunningham. 2014. Adaptive Representations for Tracking Breaking News on Twitter. *ArXiv e-prints*, March.

Eleanor Clark and Kenji Araki. 2011. Text normalization in social media: Progress, problems and applications for a pre-processing system of casual english. *Procedia - Social and Behavioral Sciences*, 27:2 – 11. Computational Linguistics and Related Fields.

Dipankar Das and Sivaji Bandyopadhyay. 2011. Document Level Emotion Tagging: Machine Learning and Resource Based Approach. *Computación y Sistemas*, 15:221 – 234.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June.

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

Fei Jiang, Yiqun Liu, Huanbo Luan, Min Zhang, and Shaoping Ma. 2014. Microblog sentiment analysis with emoticon space model. In *Social Media Processing*, pages 76–87. Springer.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval'16, San Diego, California, June. Association for Computational Linguistics.

David Pinto, Darnes Vilariño-Ayala, Yuridiana Alemán, Helena Gómez-Adorno, Nahun Loya, and Héctor Jiménez-Salazar. 2012. The soundex phonetic algorithm revisited for sms-based information retrieval. In *II Spanish Conference on Information Retrieval CERI 2012*.

Juan Pablo Posadas-Durán, Helena Gómez-Adorno, Ilia Markov, Grigori Sidorov, Ildar Z. Batyrshin, Alexander F. Gelbukh, and Obdulia Pichardo-Lagunas. 2015. Syntactic n-grams as features for the author profiling task: Notebook for PAN at CLEF 2015. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*

Vivek Kumar Rangarajan Sridhar. 2015. Unsupervised text normalization using distributed representations of words and phrases. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 8–16, Denver, Colorado, June. Association for Computational Linguistics.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642.

Chunwei Yan, Fan Zhang, and Lian'en Huang. 2014. Drws: A model for learning distributed representations for words and sentences. In Duc-Nghia Pham and Seong-Bae Park, editors, *PRICAI 2014: Trends in Artificial Intelligence*, volume 8862 of *Lecture Notes in Computer Science*, pages 196–207. Springer International Publishing.