DiegoLab16 at SemEval-2016 Task 4: Sentiment Analysis in Twitter using Centroids, Clusters, and Sentiment Lexicons

Abeed Sarker and Graciela Gonzalez
Department of Biomedical Informatics
Arizona State University
Scottsdale, AZ 85259, USA
{abeed.sarker,graciela.gonzalez}@asu.edu

Abstract

We present our supervised sentiment classification system which competed in SemEval-2016 Task 4: Sentiment Analysis in Twitter. Our system employs a Support Vector Machine (SVM) classifier trained using a number of features including n-grams, synset expansions, various sentiment scores, word clusters, and term centroids. Using weighted SVMs, to address the issue of class imbalance, our system obtains positive class F-scores of 0.694 and 0.650, and negative class F-scores of 0.391 and 0.493 over the training and test sets, respectively.

1 Introduction

Social media has evolved into a data source that is massive and growing rapidly. One of the most popular micro-blogging social networks, for example, is Twitter, which has over 645,750,000 users, and grows by an estimated 135,000 users every day, generating 9,100 tweets per second.[1] Users tend to use social networks to broadcast the latest events, and also to share personal opinions and experiences. Therefore, social media has become a focal point for data science research, and social media data is being actively used to perform a range of tasks from personalized advertising to public health monitoring and surveillance (Sarker et al., 2015a). Because of its importance and promise, social media data has been the subject of recent large-scale annotation projects, and shared tasks have been designed around social media for solving problems in complex domains (*e.g.*, Sarker *et al.* (2016a)) While the benefits of using a resource such as Twitter include large volumes of data and direct access to end-user sentiments, there are several obstacles associated with the use of social media data. These include the use of non-standard terminologies, misspellings, short and ambiguous posts, and data imbalance, to name a few.

In this paper, we present a supervised learning approach, using Support Vector Machines (SVMs) for the task of automatic sentiment classification of Twitter posts. Our system participated in the SemEval-2016 task *Sentiment Analysis in Twitter*, and is an extension of our system for SemEval2015 (Sarker et al., 2015b). The goal of the task was to automatically classify the polarity of a Twitter post into one of three predefined categories— positive, negative and neutral. In our approach, we apply a small set of carefully extracted lexical, semantic, and distributional features. The features are used to train a SVM learner, and the issue of data imbalance is addressed by using distinct weights for each of the three classes. The results of our system are promising, with positive class F-scores of 0.694 and 0.650, and negative class F-scores of 0.391 and 0.493 over the training and test sets, respectively.

2 Related Work

Following the pioneering work on sentiment analysis by Pang *et. al.* (2002), similar research has been carried out under various umbrella terms such as: se-

[1] http://www.statisticbrain.com/twitter-statistics/ Accessed on: 23rd December, 2015.

Proceedings of SemEval-2016, pages 209–214,
San Diego, California, June 16-17, 2016.

mantic orientation (Turney, 2002), opinion mining (Pang and Lee, 2008), polarity classification (Sarker et al., 2013), and many more. Pang *et al.* (2002) utilized machine learning models to predict sentiments in text, and their approach showed that SVM classifiers trained using bag-of-words features produced promising results. Similar approaches have been applied to texts of various granularities— documents, sentences, and phrases.

Due to the availability of vast amounts of data, there has been growing interest in utilizing social media mining for obtaining information directly from users (Liu and Zhang, 2012). However, social media sources, such as Twitter posts, present various natural language processing (NLP) and machine learning challenges. The NLP challenges arise from factors, such as, the use of informal language, frequent misspellings, creative phrases and words, abbreviations, short text lengths and others. From the perspective of machine learning, some of the key challenges include data imbalance, noise, and feature sparseness. In recent research, these challenges have received significant attention (Jansen et al., 2009; Barbosa and Feng, 2010; Davidov et al., 2010; Kouloumpis et al., 2011; Sarker and Gonzalez, 2014; Sarker et al., 2016b).

3 Methods

3.1 Data

Our training and test data consists of the data made available for SemEval 2016 task 4, and additional eligible training data from past Semeval sentiment analysis tasks. Each instance of the data set made available consisted of a tweet ID, a user ID, and a sentiment category for the tweet. For training, we downloaded all the annotated tweets that were publicly available at the time of development of the system. We obtained all the training and devtest set tweets, and also the training sets from past SemEval tasks. In total, we used over 19,000 unique tweets for training. The data is heavily imbalanced with particularly small number of negative instances.

3.2 Features

We derive a set of lexical, semantic, and distributional features from the training data. Brief descriptions are provided below. Some of these features were used in our 2015 submission to the SemEval sentiment analysis task (Sarker et al., 2015b). In short: we have removed uninformative features such as syntactic parses of tweets, and have added features learned using distributional semantics-oriented techniques.

3.2.1 Preprocessing

We perform standard preprocessing such as tokenization, lowercasing and stemming of all the terms using the Porter stemmer[2] (Porter, 1980). Our preliminary investigations suggested that stop words can play a positive effect on classifier performances by their presence in word 2-grams and 3-grams; so, we do not remove stop words from the texts.

3.2.2 N-grams

Our first feature set consists of word n-grams. A word n-gram is a sequence of contiguous n words in a text segment, and this feature enables us to represent a document using the union of its terms. We use 1-, 2-, and 3-grams as features.

3.2.3 Synset

It has been shown in past research that certain terms, because of their prior polarities, play important roles in determining the polarities of sentences (Sarker et al., 2013). Certain adjectives, and sometimes nouns and verbs, or their synonyms, are almost invariably associated with positive or non-positive polarities. For each adjective, noun or verb in a tweet, we use WordNet[3] to identify the synonyms of that term and add the synonymous terms as features.

3.2.4 Sentiment Scores

We assign three sets of scores to sentences based on three different measures of sentiment. For the first set of scores, we used the positive and negative terms list from Hu and Bing (2004). For each tweet, the numbers of positive and negative terms are counted and divided by the total number of tokens in the tweet to generate two scores.

For the second sentiment feature, we incorporate a score that attempts to represent the general sentiment of a tweet using the prior polarities of its terms.

[2]We use the implementation provided by the NLTK toolkit `http://www.nltk.org/`.

[3]`http://wordnet.princeton.edu/`. Accessed on December 13, 2015.

Each word-POS pair in a comment is assigned a score and the overall score assigned to the comment is equal to the sum of all the individual term-POS sentiment scores divided by the length of the sentence in words. For term-POS pairs with multiple senses, the score for the most common sense is chosen. To obtain a score for each term, we use the lexicon proposed by Guerini *et al.* (2013) . The lexicon contains approximately 155,000 English words associated with a sentiment score between -1 and 1. The overall score a sentence receives is therefore a floating point number with the range [-1:1].

For the last set of scores in this set, we used the Multi-Perspective Question Answering (MPQA) subjectivity lexicon (Wiebe et al., 2005). In the lexicon, tokens are assigned a polarity (positive/negative), and a strength for the subjectivity (weak/strong). We assign a score of -1 to a token for having negative subjectivity, and +1 for having positive subjectivity. Tokens having weak subjectivity are multiplied with 0.5, and the total subjectivity score of the tweet is divided by the number of tokens to generate the final score.

3.2.5 Word Cluster Features

Our past research shows that incorporating word cluster features improve classification accuracy (Nikfarjam et al., 2014). These clusters are generated from vector representations of words, which are learned from large, unlabeled data sets. For our word clusters, the vector representations were learned from over 56 million tweets, using a Hidden Markov Model-based algorithm that partitions words into a base set of 1000 clusters, and induces a hierarchy among those 1000 clusters (Owoputi et al., 2012). To generate features from these clusters, for each tweet, we identify the cluster number of each token, and use all the cluster numbers in a bag-of-words manner. Thus, every tweet is represented with a set of cluster numbers, with semantically similar tokens having the same cluster number. More information about generating the embeddings can be found in the related papers (Bengio et al., 2003; Turian et al., 2010; Mikolov et al., 2013).

3.2.6 Centroid Features

We collected a large set of automatically 'annotated' sentiment corpus (Go et al., 2009). Using the negative and positive polarity tweets separately, we generated two distributional semantics models using the Word2Vec tool.[4] We then applied K-means clustering to the two distributional models to generate 100 clusters each. Finally, we compute the centroid vectors for each of the clusters in the two sets.

Two feature vectors are generated from each tweet based on these centroid vectors. For each tweet, the centroid of the tweet is computed by averaging the individual word vectors in the tweet. The cosine similarities of the tweet centroid are then computed with each of the two sets of 100 centroid vectors. The vectors of similarities are then used as features. Our intuition is that these vectors will indicate similarities of tweets with posts of negative or positive sentiments.

3.2.7 Structural Features

We use a set of features which represent simple structural properties of the tweets. These include: length, number of sentences, and average sentence length.

3.3 Classification

Using the abovementioned features, we trained SVM classifiers for the classification task. The performance of SVMs can vary significantly based on the kernel and specific parameter values. For our work, based on past research on this type of data, we used the RBF kernel. We computed optimal values for the *cost* and γ parameters via grid-search and 10-fold cross validation over the training set. To address the problem of data imbalance, we utilized the weighted SVM feature of the LibSVM library (Chang and Lin, 2011), and we attempted to find optimal values for the weights in the same way using 10-fold cross validation over the training set. We found that *cost* $= 64.0$, $\gamma = 0.0$, $\omega_1 = 1.2$, and $\omega_2 = 2.6$ to produce the best results, where ω_1 and ω_2 are the weights for the positive and negative classes, respectively.

4 Results

Table 1 presents the performance of our system on the training and test data sets. The table presents the

[4] `https://code.google.com/archive/p/word2vec/`. Accessed Feb-22-2016.

positive and negative class F-scores for the system, and the average of the two scores— the metric that is used for ranking systems in the SemEval evaluations for this task. The training set results are obtained via training on the training set and evaluating on the devtest set. The test results are the final SemEval results.

Data set	Positive F-score (P)	Negative F-score (N)	$\frac{P+N}{2}$
Training	0.694	0.391	0.542
Test	0.650	0.493	0.571

Table 1: Classification results for the DIEGOLab16 system over the training and test sets.

4.1 Feature Analysis

To assess the contribution of each feature towards the final score, we performed leave-one-out feature and single feature experiments. Tables 2 and 3 show the $\frac{P+N}{2}$ values for the training and the test sets for the two set of experiments. The first row of the tables present the results when all the features are used, and the following rows show the results when a specific feature is removed or when a single feature is used. The tables suggest that almost all the features play important roles in classification. As shown in Table 3, n-grams, word clusters, and centroids give the highest classification scores when employed individually. Table 2 illustrates similar information, by showing which features cause the largest drops in performance when removed. For all the other feature sets, the drops in the evaluation scores shown in Table 3 are very low, meaning that their contribution to the final evaluation score is quite limited. The experiments suggest that the classifier settings (*i.e.*, the parameter values and the class weights) play a more important role in our final approach, as greater deviations from the scores presented can be achieved by fine tuning the parameter values than by adding, removing, or modifying the feature sets. Further experimentation is required to identify useful features and to configure existing features to be more effective.

Feature removed	$\frac{P+N}{2}$
None	0.542
N-grams	0.540
Synsets	0.553
Sentiment Scores	0.540
Word Clusters	0.515
Centroids	0.527
Other	0.541

Table 2: Leave-one-out $\frac{P+N}{2}$ feature scores for the training and test sets.

Feature	$\frac{P+N}{2}$
All	0.542
N-grams	0.515
Synsets	0.494
Sentiment Scores	0.472
Word Clusters	0.531
Centroids	0.535
Other	0.254

Table 3: Single feature $\frac{P+N}{2}$ scores for the training and test sets.

5 Conclusions and Future Work

Our system achieved moderate performance on the SemEval sentiment analysis task utilizing very basic settings. The F-scores were particularly low for the negative class, which can be attributed to the class imbalance. Considering that the performance of our system was achieved by very basic settings, there is promise of better performance via the utilization of feature generation and engineering techniques.

We have several planned future tasks to improve the classification performance on this data set, and for social media based sentiment analysis in general. Following on from our past work on social media data (Sarker and Gonzalez, 2014; Sarker et al., 2016b), our primary goal to improve performance in the future is to employ preprocessing techniques that can normalize the texts and better prepare them for the feature generation stage. We will also attempt to optimize our distributional semantics models further.

Acknowledgments

This work was supported by NIH National Library of Medicine under grant number NIH NLM 1R01LM011176. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NLM or NIH.

References

Luciano Barbosa and Junlan Feng. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. In *Proceedings of COLING*, pages 36–44.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *Proceedings of COLING*, pages 241–249.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision. `https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf`.

Marco Guerini, Lorenzo Gatti, and Marco Turchi. 2013. Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1259–1269.

Minqing Hu and Bing. 2004. "mining and summarizing customer reviews". In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter Power: Tweets as Electronic Word of Mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 415–463.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781v3.

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. 2014. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association (JAMIA)*.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer Kevin Gimpel, and Nathan Schneider. 2012. Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances. Technical report, School of Computer Science, Carnegie Mellon University.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the ACL conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Abeed Sarker and Graciela Gonzalez. 2014. Portable Automatic Text Classification for Adverse Drug Reaction Detection via Multi-corpus Training. *Journal of Biomedical Informatics*.

Abeed Sarker, Diego Molla, and Cecile Paris. 2013. Automatic Prediction of Evidence-based Recommendations via Sentence-level Polarity Classification. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 712–718.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015a. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212.

Abeed Sarker, Azadeh Nikfarjam, Davy Weissenbacher, and Graciela Gonzalez. 2015b. Diegolab: An approach for message-level sentiment classification in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 510–514, Denver, Colorado, June. Association for Computational Linguistics.

Abeed Sarker, Azadeh Nikfarjam, and Graciela Gonzalez. 2016a. Social Media Mining Shared Task Workshop. In *Proceedings of the Pacific Symposium on Biocomputing*.

Abeed Sarker, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. 2016b. Social media mining for toxicovigi-

lance: Automatic monitoring of prescription medication abuse from twitter. *Drug Safety*, 39(3):231–240.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394.

Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions opinion and emotion in language. *Language Resources and Evaluation*, 39:165–210.