# Hybrid Approach for Message polarity and Topic based message polarity Sentiment classification

**P.Vijayapal Reddy**
Department of Computer Science and Engineering
Matrusri Engineering College
Hyderabad, India
drpvijayapalreddy@gmail.com

**Abstract**

The approach for text sentiment classification is based on a Majority Vote scheme and combined supervised machine learning methods with classical linguistic resources, including bag-of-words and sentiment lexicon features.

***Keywords***— Lexicon, Machine Learning, Linguistic.

## 1. Introduction

For millions of users, microblogging services such as Twitter, a popular service where users can post no more than 140 characters status messages have become an elemental part of daily life. By using tools and techniques from Natural Language Processing (NLP) and machine learning, Sentiment analysis is defined as the process to identify and analyze polarity from short texts, sentences, and documents [1]. In the last few years, people from different research disciplines are interested in Sentiment Analysis Tweet Classification in either two-point or five-point scale respectively [2].

In this paper, an ensemble text sentiment classification scheme, based on an extensive empirical analysis of several classifiers and other related works, e.g. [3,4,5,6]. A voting scheme combines learning algorithms to identify and select an optimal set of base learning algorithms. These components were carefully combined and optimized to create a separate version of the system.

## 2. System Description

The system used is based on the bag-of-words representation, n-gram extraction, and usage of lexicons which have a predefined sentiment for every unigram and bi-gram. For the implementation of the system use Python's ScikitLearn [7], as well as NLTK (Natural Language Toolkit) [8].

### *A. Preprocessing*

The pre-processing steps that followed were to remove and replace strings from the tweets that do not show any sentiment, as well as to remove duplicates and Unicode strings:

- Removing duplicates: Duplicate instances are removed.
- Replacing hashtags, URLs and usernames: first remove the "#" character in front of the words and replaced the twitter oriented strings @usernames and the URLs with tags such as "AT USER" and "URL" respectively.

- Removing Unicode strings: there were many Unicode strings especially in the testing data, e.g. strings like “\u002c” and “x96”.
- Removing numbers and punctuation: preliminary experiments showed better results when removed all the numbers. Before removing punctuation, The detected useful punctuation signs such as “!” and “?” and replaced them with labels.
- Using lowercase and tokenization: the final tweets were lower-cased (after detecting words that had all of their character capitalized which were retained) and divided into tokens.
- Removing stop words: stop words are common function words with very high frequency among sentences and low content, hence remove them.
- Using stemming: stemming is the process of reducing a word to its base root form. Preliminary tests showed that stemming improves a lot the results.

*B. Feature Engineering*

The extracted features based on the lexical content of each tweet and lexicons.

TABLE 1: Number of tweets in training, development and testing for topic based Message polarity classification

| | Positive | Negative | Total |
|---|---|---|---|
| Train | 12812 (79%) | 3410 (21%) | 16222 |
| Dev | 2139 (78%) | 604 (22%) | 2743 |
| Test | 2463 (40%) | 3722 (60%) | 6185 |

- Word n-grams: the word level unigrams and bi-grams are adopted.
- Number of capitalized words
- Number of question marks, exclamation marks and the aggregation of them
- Number of elongated words: it indicates the number of elongated words in the raw text of the tweet.

Sentiment lexicons are lexical resources which are formed by a list of words without any additional information and are built by opinion words and some sentiment phrases [4].

In this system, sentiment lexicons such as Bing Liu’s lexicon [11], the NRC emotion lexicon [12], the MPQA lexicon [13] and combinations of them. The above lexicons have a sentiment tag for each word and in this approach, count the occurrences of each sentiment class for each tweet’s word. Finally, compute the overall sentiment of the tweet, by adding its words sentiments.

## 3. Experiments

In this section, after the feature extraction, analyse the classification process with the learning methods and classification algorithms that used in the proposed system.

A. Datasets

The datasets were provided by the organizers and contained all datasets of the previous years with the addition of a new. The data for development and the rest for training.

B. Evaluation Metrics

The macro-average re-call, which is the recall averaged across the three classes

$$R_{macro} = \frac{R_{pos} + R_{neu} + R_{neg}}{3}$$

Maintains the same measure, official metrics are the macro-averaged mean absolute error and the extension of macro-averaged recall for ordinal regression [2] among 5 predefined classes.

C. Learning

Using all the features described above, The first trained several classifiers to the development data in order to tune the parameters of each classifier. The main target of tuning was the metric of this specific task, which is the macro-average recall. Test a variety of classifiers that include the following:

- Ridge: an algorithm belonging to the Generalized Linear Models family that alleviates the multicollinearity amongst predictor variables.
- Logistic Regression: despite its name it is used for classification and fits a linear model. It is also known as Maximum Entropy, and uses a logistic function to model the probabilities that describe the output prediction.
- Stochastic Gradient Descent: a simple and efficient algorithm to fit linear models. It is suitable for very large number of features.
- Nearest Centroid: an algorithm that uses the center of a class, called centroid, to represent it and has no parameters.
- Bernoulli Naive Bayes: an alternative of Naive Bayes, where each term is equal to 1 if it exists in the sentence and 0 if not. Its difference from Boolean Naive Bayes is that it takes into account terms that do not appear in the sentence.
- Linear SVC: an SVM algorithm, which tries to find a set of hyperplanes that separate space into dimensions representing classes. The hyperplanes are chosen in a way to maximize the distance from the nearest data point of each class.
- Passive-Aggressive: belongs to a family of algorithms for large-scale learning, which do not require a learning rate and includes a regularization parameter C [7].

In order to vectorize the collection of raw documents, a Python's ScikitLearn [7] tfidf transformation with a max df parameter of 0.5. The value of this parameter was extracted by the tuning process and indicates that we ignore terms that have a frequency strictly higher than this threshold. The next step was to use these parameters to test the model with the help of 10-fold cross-validation on the training set.

D. Multi-class classification number

Multi-class classification problem, where each tweet has to be classified in one among three classes. The best combination for this task was the use of stemming and the three lexicons. Features like the number of exclamation marks, etc., under-performed. The three classifiers with the best results were the Bernoulli Naive Bayes, the Stochastic Gradient Descent (SGD), and the Linear SVC.

The final step was to use the majority voting classification method that combines three different classifiers and outputs the class that the majority of them agreed. Using all possible combinations of every three classifiers, the best result was with the Bernoulli Naive Bayes, SGD, and Nearest Centroid. Note that Nearest Centroid was one of the Weakest classifiers in isolation, but presented an excellent contribution when combined with other two.

TABLE 2: Experimental Results

| | $\rho$ | $F_1^{PN}$ | Acc |
|---|---|---|---|
| Polarity classification | 0.621 | 0.605 | 0.640 |
| | ($MAE^M$) | ($MAE^\mu$) | |
| Topic based | 0.895 | 0.544 | |

The topic-based classification problem, where each tweet belongs to a topic, and one has to estimate the sentiment conveyed by the tweet towards the topic on a five-point scale. The same approach for both the message polarity classification and topic based message polarity classification the best result was achieved by the combination of the Logistic Regression, the Nearest Centroid, and the Bernoulli Naıve Bayes classifiers.

## 5. Conclusions and Future Scope

By analyzing and classifying sentiments on Twitter, people can comprehend attitudes about particular topics, making Sentiment Analysis an attractive research area. An approach for Twitter sentiment analysis on two-point, three-point, and five-point scale, based on a voting classification method.
Contact with the task of sentiment analysis and compared with the top-ranked participating systems, there seems to be for us much room for improvement.

In future work, consider to focus on adding more pre-processing methods such as spelling correction and POS tagging, also consider adding more features such as emoticons, negation, character n-grams and more lexicons.

**References**

1. Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. Foundations and Trends R in Information Retrieval 2(1–2):1–135.

2. Sara Rosenthal, Noura Farra, and Preslav Nakov, 2017.Semeval-2017 task 4: Sentiment analysis in twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation(SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 501–516.

3. Alexandra Balahur. 2013. Sentiment analysis in social media texts. In 4th workshop on computational approaches to subjectivity, sentiment and social media analysis.

4. Eugenio Mart´ınez-Camara,´ Salud Mar´ıa Jimenez´-Zafra, Maite Martin, and L. Alfonso Urena Lopez. 2014. Sinai: Voting system for twitter sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pages 572–577. http://www.aclweb.org/anthology/S14-2100.

5. Georgios Balikas and Massih-Reza Amini. 2016. Twise at semeval-2016 task 4: Twitter senti-ment classification. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, San Diego, California, pages 85–91.

6.Aytu Onan, Serdar Korukolu, and Hasan Bulut. 2016.A multi objective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Systems with Applications 62:1 – 16.

7. Fabian Pedregosa, Gael¨ Varoquaux, Alexandre Gram fort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake
8. Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard´ Duch esnay. 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12:2825–2830.

9. Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural language processing with Python: analyzing text with the natural language toolkit. ” O’Reilly Media, Inc.”.

10. Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining.In LREc. volume 10.

11. Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O’Brien, Lamia Tounsi, and Mark Hughes. 2013. Sentiment analysis of political tweets: Towards an accurate classifier. Association for Computational Linguistics.

12. J Minqing Hu and Bing Liu. 2004. Mining and sum-marizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, KDD ’04, pages 168–177.

13. Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010

Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Association for Computational Linguistics, Stroudsburg, PA, USA, CAAGET '10, pages 26–34.

14. Theresa Wilson, Janyce Wiebe, and Paul Hoffmann.2005. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of theConference on Human Language Technology and Empirical Methods in Natural Language Process ing. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '05, pages 347–354.