Enhancing the Performance of Sentiment Analysis by Using Different Feature Combinations

Nazma Iqbal [1] , *Afifa Mim Chowdhury* [1]
Dept. of Computer Science & Engineering
[1]International Islamic University, Chittagong
Chittagong- 4318, Bangladesh
[1] marwaiqbal92@gmail.com
[1] afifamim06cseiiuc@gmail.com

Tanveer Ahsan
Dept. of Computer Science & Engineering
International Islamic University, Chittagong
Chittagong- 4318, Bangladesh
tanveer.ahsan@gmail.com

***Abstract*—With the growing accessibility and acceptance of social networking epoch, Sentiment Analysis has become one of the most prominent research domain in natural language processing. Each day, millions of people share their thoughts and ideas by posting in social media or writing online reviews. This massive participation, on one hand, makes these media opinion-rich; however, on the other hand, it poses some challenges in identifying the dominant opinion. In this work, tweets and movie reviews are classified according to the polarity of the opinions by using several features in combination. Performance of several feature combinations was evaluated by feeding those in different Machine Learning algorithms (NB, SVM, MaxEnt). Hence, the goal of the work was to evaluate how the performance of a classifier is affected when different feature combinations are used in Sentiment Analysis. Experiments were done on data from two different domain namely Stanford Twitter Sentiment140 dataset and IMDb Movie Reviews dataset. Four different evaluation metrics: recall, precision, accuracy and F1 score are used for evaluating the investigational results of our system. This research demonstrates that by carefully choosing correct feature combination the classification accuracy can be increased while a random feature combination will provide little benefit.**

Keywords—social media; sentiment analysis; twitter; movie review; sentiment classification; feature combination; machine learning approach (MLA)

I. Introduction

Now-a-days, social networking sites turn out to be the most prominent opinion sharing source because of its growing popularity. So there is a need to analyze the sentiment on social media as it draws much response of those who seek to apprehend the opinions of personages. Among various types of social opinion rich resources, Twitter and Movie review blogs are very popular to share views on different issues or on specific movies. For observing and exploring people's thoughts, Sentiment Analysis has become most prominent research area as it aims to verdict hidden patterns in a large number of tweets [1], reviews or blogs with the help of Machine Learning techniques in the recent time. Modern research's spotlights are now on Sentiment Analysis on different domains as well as different languages [2]. In this paper, we have proposed a modularized polarity classification system for Sentiment Analysis using three Machine Learning techniques. To sort out the sentiment as positive or negative, we use feature extractor and classification algorithm as two independent components.

Our main contributions in this research include: (1) We have determine the best feature combination schemes for Sentiment Analysis from unigram and bigram features, most informative unigram and most informative bigram features, Single word features and Stopword filtered word features, Bigram and Stopword filtered word features and so on. (2) We have explored which Machine Learning technique best suits with which of the feature combinations mentioned above by applying the techniques on two different domains- tweet data (short length) and movie reviews (length is not limited) data. The experiment shows that the classification accuracy of proposed model using Naive Bayes, Support Vector Machine and Maximum Entropy is higher and more optimal than single feature selection model.

Our proposed system has attained better result over the baseline in terms of recall, precision, F1 score and accuracy that outperforms some well-known existing system. After conducting the experiments, we find that the Maximum Entropy is the best choice for “unigram and bigram” feature combination and achieves 90% F1 score for Movie Review data and 89% F1 score for Tweet data.

II. Related Work

In the recent time, research on Sentiment Analysis in information retrieval sphere concentrate on classifying sentiment according to their polarity either as positive or negative or neutral. Earlier researches on Sentiment Analysis are based on subjectivity or sentiment level. One standard technique was developed by [3] which outperformed human produced baselines as it analyzed the movie reviews on the basis of different Machine Learning classifiers. For many other researchers, this technique is functioned as a baseline like [4] as he intended to develop a “Distant Supervision Learning” method to handle noisy labeled tweet data. As for sentiment level detection, [5] described a system that detects the sentiment of message-level task and the sentiment of term-level task by creating two different SVM classifiers. But [2] used aspect based sentiment scrutiny technique which was applied on datasets for different languages and domains. On the other hand [1] evolved a structure to handle the target tweets. Few researchers reconnoiter features trade along with combination [4] where they developed a feature combination scheme which utilizes the sentiment lexicons and the extracted feature combinations [6] where they developed a feature combination scheme which utilizes the sentiment lexicons and the extracted tweet unigrams of high information gain by evaluating the performance. Current trend of Sentiment Analysis shows that consequences of microblogging sentiment classification are broadly used in different social media applications, including tracking sentiment towards products [7], movies [3] etc.

III. Proposed Framework

This section depicts in detail the phases of our proposed methodology for sentiment classification. The whole process of our system architecture is shown in the figure 1.

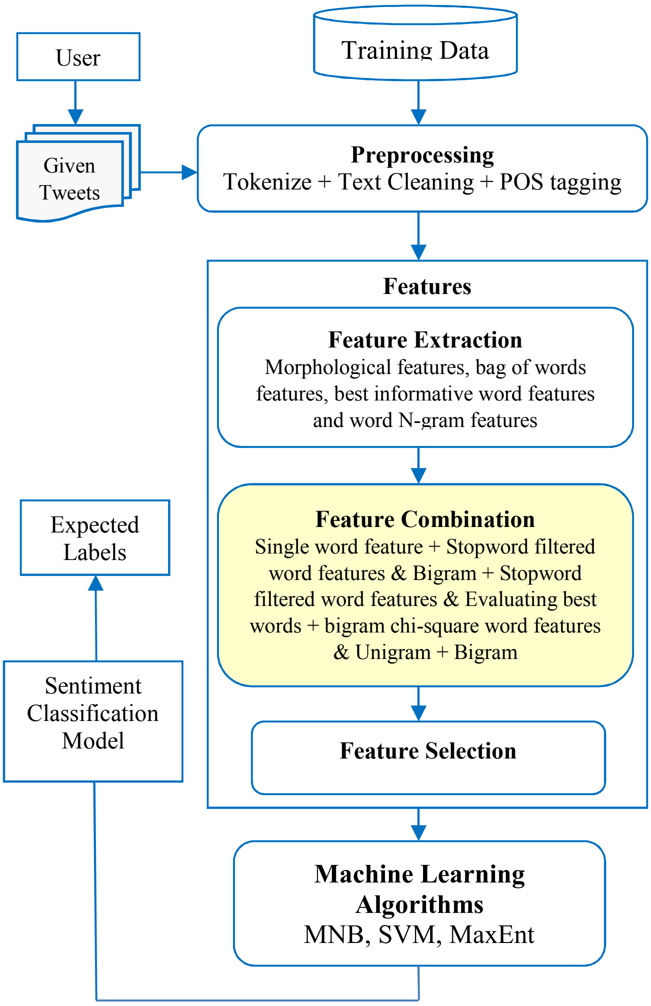
User
Training Data
Given Tweets
Preprocessing
Tokenize + Text Cleaning + POS tagging
Features
Feature Extraction
Morphological features, bag of words features, best informative word features and word N-gram features
Expected Labels
Feature Combination
Single word feature + Stopword filtered word features & Bigram + Stopword filtered word features & Evaluating best words + bigram chi-square word features & Unigram + Bigram
Sentiment Classification Model
Feature Selection
Machine Learning Algorithms
MNB, SVM, MaxEnt

Fig 1. Proposed Architecture

For developing the system, we have applied supervised learning technique as we contemplate to work with domain-specific labelled data. At the very beginning stage of our work, our system has collected raw data from two different sources and indexed the collected data for preprocessing. In preprocessing phase, we have performed several steps including text processing, text cleaning and POS Tagging. At the training phase, several kinds of features are extracted from the preprocessed data which are combined later with the aim to achieve high level of performance. Next, we have used three state-of-the-art classifiers, namely, Naïve-Bayes, Support Vector Machine and Maximum Entropy to automatically classify the text into positive or negative classes and determine the best combination for Sentiment Analysis.

A. *Data Preprocessing*

In order to enhance the performance of classifiers, we have pre-processed the extracted data before investigate. At first, we tokenize the input streams into distinct words. Usually raw texts contain special stuffs like URLs, User name, Hashtag, Punctuation and additional white space, which doesn't bear any sentiment. So we have eliminated all this unwanted stuffs. After that, we have converted all tokens to lowercase as well as eliminated stopwords, which also don't convey any sentiment, using NLTK stopword corpus[1]. After the process of data cleaning, the remaining words are then lemmatized by using WordNet Lemmatizer[2]. For Sentiment Analysis, POS taggers have been developed to classify words based on their parts of speech. Moreover, we have tagged each lemmatized word using Penn Treebank Project[3] which provides 36 different tags.

B. *Sentiment Lexical Resources*

Sentiment lexicon refers to a set of sentiment word senses which contain words like “wonderful”, “amazing”, and “terrible” with positive and negative scores. For the purpose of our research, we have used two publicly available English lexical resources namely, SentiWordNet [8] and Vader Sentiment Lexicon [9]. The positive and the negative words are extracted distinctly from the lexicons to compute the polarity of the words which are then used to train the classifier.

C. *Features*

For our system, we have extracted several kinds of features which are broadly grouped into morphological features, bag of words features, best informative word features and word N-gram features. The extracted features can be described as the following groups:

- *Morphological features*: We have used morphological features as a binary feature which analyzes the presence or absence of elongated words (such as ‘coooooool’), time and date expressions, exclamation marks and question marks. It also counts the number of elongated words, fully and partially capitalized tokens, ellipsis, exclamation and question marks etc.
- *Bag of words features*: In this model, the existence of

[1] http://www.nltk.org/data.html
[2] http://www.nltk.org/_modules/nltk/stem/wordnet.html
[3] http://www.cis.upenn.edu/~treebank/home.html

every single token is counted as a feature for learning a classifier. For our system, we have extracted stopword filtered word as a feature from the processed data. We have also handled negation features by affixing a "mark_negation" suffix after a negation word string.

- *N-Gram word features*: In [3], they have experimented with unigrams, bigrams and the combination of unigrams and bigrams as features and applied these features on various Machine Learning algorithms. Based on that, we have extracted unigrams, bigrams and trigrams features for our Sentiment Analysis approach. For these N-gram based features we have used Pipeline, a Machine Learning tool in python, which is used for feature extraction and evaluation.
- *Most informative features*: In order to accelerate processing pace, we have extracted most informative unigram and most informative bigram features for our system.

From the above mentioned extracted features, several types of features are experimented by combining to train the classifier. It is a good indication to use combined feature sets instead of single feature sets for achieving higher accuracy than solo feature selection system. For our experiment, we have combined Single word features with Stopword filtered word features (set1), Unigram with Bigram features (set2), Bigram with Stopword filtered word features (set3) and Most informative Unigram with Most informative Bigram features (set4).

To identify most relevant features, we have further focused on a supervised feature selection method for our Sentiment Analysis task. In our system, we have applied the chi square technique for selecting high informative features. More concretely, by selecting the high informative features, a model can provide enhanced performance and decline the size of the model. Python includes this technique in the BigramAssocMeasures class in metrics package. We have used FreqDist class which is used to calculate the frequency of every token. By using these frequency numbers, we have scored the tokens with the BigramAssocMeasures.chi sq function and then sorted the words by score for getting most significant unigram and bigram features.

D. *Classification Model*

In our experiment, we have used three state-of-the-art classifiers, namely, Naïve-Bayes, Support Vector Machine and Maximum Entropy. We have trained these three classifiers and performed five-fold cross-validation over the training data. After the training process, we have then applied the trained classifier on the test dataset so that the new tweets could be labeled as positive or negative.

Naive Bayes: Naive Bayes is a powerful classification model which is used to determine the probabilistic results based on Bayesian theorem [10]. In our work, we have used MultinomialNB[1] package in python which is a linear classification model used to estimate the subsequent probability of features belong to a class.

Support Vector Machines: SVM model is a non-probabilistic binary algorithm that explores training texts for classification. The elementary concept is to figure out a hyper plane which distinguishes the negative and positive classes with maximizing the boundary between two classes. For training and testing data, we have used LIBSVM[2] library with a kernel based method.

Maximum Entropy: MaxEnt is a probabilistic model which falls under the category of exponential models where features are tentatively independent [11]. It outperforms the other models because of its search-based optimization character which capitalizes the log-likelihood of the learning text.

IV. Experiment and Evaluation

A. *Data Description*

In the experiment, we have used two different publicly accessible datasets with two different fields, one is IMDB Movie Reviews dataset[3] which contains 25,000 movie reviews and another is Stanford Twitter Sentiment 140 dataset[4] which contains 1.6 million tweets with positive or negative orientation. For our Sentiment Analysis approach, first we have shuffled each dataset and 80% of the data are used as train feature set and remaining 20% are used as test feature set.

B. *Experimental Result*

At first, we have trained the classifiers by using the training datasets. We have used SentiWordnet method as a baseline for further approaches. The SentiWordnet approach have produced 0.645 F1-measure for tweet dataset and 0.677 F1-measure for movie review dataset. We have evaluated the performance by using three state-of-the-art classifiers, namely, NB, SVM and MaxEnt based on two different domains. The results of machine learning-based classifiers are evaluated in terms of four evaluation metrics. The evaluation metrics includes recall, precision, F1-measure and accuracy. It can be seen from the experimental results, using both data sets, the best result is obtained when both unigram and bigram features are combined and classified with MaxEnt classifier. Moreover, we have premised that it is useful to use feature combinations model rather than using single features model. By combining several kinds of features, we have achieved the best results by using several classifiers to estimate the sentiment label of given input stream. TABLE I shows the classification results for several extracted feature combinations. We have also provided some comparative results among some existing co-related sentiment methods and our proposed methods. Our experimental results show notably improved results compared to the available system. TABLE II. shows the comparative results based on IMDB Movie Review test data and sentiment140 test data.

[1] http://scikit-learn.org/stable/modules/naive_bayes.html

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[3] https://www.kaggle.com/c/word2vec-nlp-tutorial/data

[4] http://help.sentiment140.com/for-students/

TABLE I. CLASSIFICATION RESULTS ON BOTH DATATSETS

Data-sets	Featu re sets	NB				SVM				MaxEnt			
		Recall	*Preci-sion*	*F1-measure*	*Accura-cy*	*Recall*	*Preci-sion*	*F1-measure*	*Accura-cy*	*Recall*	*Preci-sion*	*F1-measure*	*Accura-cy*
Tweet dataset	Set1	0.79	0.83	0.78	0.79	0.84	0.84	0.84	0.84	0.78	0.83	0.77	0.78
	Set2	0.89	0.86	0.87	0.86	0.88	0.86	0.87	0.85	0.90	0.88	**0.89**	**0.88**
	Set3	0.85	0.87	0.84	0.85	0.85	0.85	0.85	0.85	0.82	0.85	0.81	0.82
	Set4	0.82	0.86	0.81	0.82	0.79	0.84	0.78	0.84	0.84	0.84	0.84	0.79
Movie Review dataset	Set1	0.81	0.85	0.80	0.81	0.85	0.86	0.85	0.86	0.80	0.84	0.79	0.80
	Set2	0.86	0.90	0.88	0.89	0.90	0.89	0.89	0.89	0.91	0.89	**0.90**	**0.90**
	Set3	0.85	0.87	0.85	0.85	0.86	0.85	0.86	0.82	0.82	0.85	0.82	0.86
	Set4	0.83	0.87	0.83	0.84	0.86	0.86	0.86	0.86	0.79	0.83	0.77	0.78

TABLE II. COMPARISON WITH CO-RELATED WORKS

Author	Data Source	F1	Accuracy
Fang *et. al.* [5]	Tweeter API	85%	N/A
Alec *et. al.*[2]	Twitter API, Sentiment 140	N/A	85%
Our System	IMDb Movie Reviews dataset	**90%**	**90%**
	Stanford Twitter Sentiment 140	**89%**	**88%**

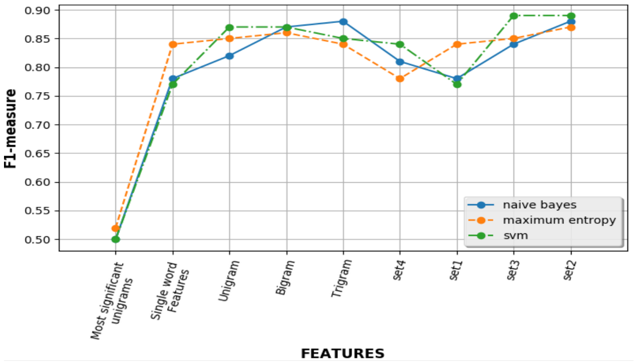

Fig 2. Feature sets illustration of Tweet data

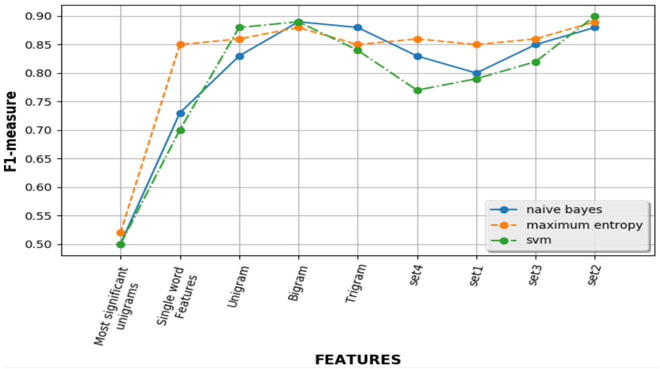

Fig 3. Feature sets illustration of Movie Review data

Fig. 2 and Fig.3 is the graphical representation of the classification results of Feature sets for both datasets.

CONCLUSION AND FUTURE DIRECTION

This paper addresses the task of document-level and sentence-level Sentiment Analysis in two different domains by developing a modularized polarity classification system using Machine Learning algorithms. Our proposed system analyses the microblogging messages based on several feature combinations schemes to determine the best combination sets for Sentiment Analysis. Illustrations of experimental results indorse that our method attained the significant enhancements over any single method, outperforming state-of-the-art methods by more than 2-5% Accuracy and F1 points in the sentiment analysis task. In future, our planned research studies center of attention will be on recognizing and identifying explicit as well as implicit opinions with the help of their explicit and implicit features along with neutral sentiment.

REFERENCES

[1] Rosenthal, Sara, Noura Farra, and Preslav Nakov. "SemEval-2017 task 4: Sentiment analysis in Twitter." *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 2017.

[2] Pontiki, Maria, et al. "SemEval-2016 task 5: Aspect based sentiment analysis." *ProWorkshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, 2016.

[3] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002.

[4] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." *CS224N Project Report, Stanford* 1.2009 (2009)

[5] Mohammad, Saif M., Svetlana Kiritchenko, and Xiaodan Zhu. "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets." *arXiv preprint arXiv:1308.6242* (2013).

[6] Yang, Ang, et al. "Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination." *Security and Privacy in Social Networks and Big Data (SocialSec), 2015 International Symposium on*. IEEE, 2015.

[7] Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." *Journal of Big Data* 2.1 (2015): 5.

[8] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In *LREC*, vol. 10, pp. 2200-2204. 2010.

[9] Hutto, Clayton J., and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In *Eighth international AAAI conference on weblogs and social media*. 2014.

[10] G. H. John and P. Langley, “Estimating continuous distributions inbayesian classifiers,” in *Proceedings of the Eleventh conference onUncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

[11] Zhu, Shenghuo, et al. "Multi-labelled classification using maximum entropy method."*Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*.ACM, 2005.