International Journal For Technological Research In Engineering
Volume 8, Issue 1, September-2020 *ISSN (Online): 2347 - 4718*

TO INVESTIGATE THE FEATURE SELECTION METHOD TO IMPROVE THE PERFORMANCE OF THE CLASSIFIER USING SENTIMENT CLASSIFIER

Sana Anzar[1], Dr. Pritaj Yadav[2], Mr. Mukesh Kumar[3]
[1]M.Tech Scholar, [2,3]Associate Professor,
Rabindranath Tagore University, Raisen

Abstract: Sentiment analysis deals with the processing of opinion text to extract and categorize opinions from the documents. The sentiment is expressed in terms of positive or negative opinion. The information technology field have entered our lives and it is impossible to imagine without Internet. To ensure efficient classification, it is important to implement an algorithm that performs well in featre selection. Therefore, the main goal of the research study is to investigate algorithms that can be applied to opinion estimation. To that extent, data preprocessing and multiple experiments are performed. The classifier is trained and tested on two different datasets with two different classifiers. The impact of training data on classifier efficiency is measured and found better outcome as compared the earlier study of sentiment classifier.
Keywords: Sentiment analysis, Sentiment, Feature selection, Convolutional Neural Network.

I. INTRODUCTION

The difference between people and machine; is that people can express personal thoughts and the dream behind artificial intelligence is to make machines behave like humans. The fields of computer linguistics that analyze opinions are called opinion mining or it is also called sentiment analysis. Opinion mining is part of natural language processing that relates to the opinion of analysis about products, services, and even people. Based on sentiment analysis and opinion mining focus primarily on ideas that express positive or negative emotion. To perform an analysis of opinions, opinions have to be extracted [1].

Today, the retrieval of opinions became easy as individuals share their views about various topics through social networks like Twitter, Facebook or they leave comments and reviews about products on a particular website. Microblogging can be considered a rich source of messages with different opinions that can be collected and further used to extract emotions. The analysis of opinion plays an important role in all science fields (politics, economics, and social life). For example, in marketing, if the seller is aware of the satisfaction of the particular product of the customer then he can estimate the demand for the product. Same for politicians, they will come to know if people support them or not [2,3,4].

II. LITERATURE REVIEW

The sentiment analysis deals with the processing of opinionated text to extract and categorize opinions from certain documents. The polarity of sentiment usually expressed in terms of positive or negative opinion [5,6]. However, it can be multi-class classification [7,9,10], hence sentiment may have a neutral label or even broadened variation of labels like very positive, positive, neutral, negative, very negative, also labels can be associated with emotions like anger, sad, fearful, happy, etc.

Sentiment analysis can be carried out at the following levels:

- Document-level. At this level, the main task is to define the opinion of the whole document (opinion should be expressed about one topic).
- Sentence level. Here every sentence is considered as a short document that can be subjective or objective. The subjective (opinionated) sentence expresses sentiment.
- Aspect level. Allows extracting opinions towards aspects of entities.

There are following methods are used in sentiment analysis :

2.1 Lexicon-based method

The technique used for SA is the lexicon-based methodology. It uses a word that contains words with a corresponding emotion score for each word. This word may be associated with a single word, phrase or idiom [8]. The sentiment is defined based on the presence or absence of words in the lexicon. Lexicon-based approaches include corpus-based approaches and dictionary-based approaches which are discussed further [11].

2.2 Dictionary-based method

The central idea behind the dictionary-based approach is to use lexical databases with opinion words to extract sentiment from the document. Based on [12], [13], a set of seed sentiment words with their polarities is collected by hand. In the beginning, this initial set does not have to be large, 40 opinion words are enough. The next step is to use the polar words to enrich a set by looking up for respective synonyms and antonyms in a lexical database. At each iteration, the algorithm takes an updated set of words and does search again until there will be no new words to include. In the end, a set of sentiment words can be reviewed to delete errors [14].

2.3 Corpus-based method

The corpus-based approach can be applied in two cases. The first case is an identification of opinion words and their polarities in the domain corpus using a given set of opinion words. The second case is for building a new lexicon within the particular domain from another lexicon using a domain corpus. The findings suggest that even if opinion words are

domain-dependent it can happen that the same word will have opposite orientation depending on the context.

2.4 Machine learning method

The technique that can be used for sentimental analysis is machine learning that includes unsupervised and supervised machine learning methods.

2.4.1 Unsupervised machine learning

An unsupervised learning approach uses unlabeled datasets to discover the structure and find similar patterns from the input data. An unsupervised method is usually used when a collection of a reliable annotated dataset is difficult, but the collecting of unlabeled data is easier. It does not cause any difficulties when new domain-dependent data have to be retrieved [15].

2.4.2 Supervised machine learning

The supervised machine learning methods assume the presence of labeled training data that are used for the learning process. As the training data set, labeled documents have to be used. Usually, bag-of-words model is employed to represent a document as a feature vector. To convert the training dataset to a feature vector, vocabulary with unique words has to be created from the training data.

III. PROBLEM STATEMENT

The focus of this paper is to conduct sentiment analysis on movie reviews and Twitter messages by identifying positive and negative ones. We decided to investigate two approaches in detail in this research study: Naïve Bayes and Convolutional neural network. Classification is performed on tweets, where each tweet is labeled as positive or negative according to the opinion expressed in it.

IV. PROPOSED METHODOLOGY

There are a lot of websites that shows business and product reviews. Amazon is a website where customers can publish their feedback about products as well as a lookup for reviews to decide on purchasing a product. Another interesting and useful source of opinions is TripAdvisor. TripAdvisor is a website that provides dozens of opinionated information about hotels, restaurants, flights, places where to go, which is very helpful for travelers [16]. The Twitter is another way of sharing views. Information from such sources is used not only by customers, but it is also vital for different organizations.

4.1 Data and preprocessing

The two datasets are used for training classifiers. The first dataset of movie reviews is considered because such kind of reviews comprise a broad range of emotions and capture many adjectives suitable for sentiment classification. The second dataset is a dataset that contains automatically annotated tweets [17].

After training data is extracted, next step is to preprocess it in order to exclude irrelevant data from the dataset. Preprocessing includes the following:

- Removal of URLs
- Removal of usernames
- Removal of hashtags
- Removal retweets and duplicates
- Compression of elongated words
- Removal of stop words

4.2 Feature extraction

After preprocessing is completed, features have to be extracted and further used for training the classifiers. In the first experiment, the unigrams were selected as features for feeding the Naïve Bayes classifier. Sentence is split into words and represented as a set of words. Using unigrams end up in a large feature set that has to be reduced to eliminate uninformative features. The experiment was conducted using a convolutional neural network. CNN uses filters that play the role of feature detectors. The size of the movie reviews dictionary constitutes 19058 words and size of the tweets dictionary constitutes 214062 words.

V. RESULTS AND ANALYSIS

This section shows the results that were obtained after conducting the experiments using the Naïve Bayes algorithm and convolutional neural network. Naïve Bayes algorithm implemented using the NLTK library and neural network using Tensorflow. Training and testing of the system were performed on the Rocket cluster that has 125 nodes (having 2.80 GHz, 164 GB RAM, 1.5 TB Hard disk drive system) helps to speed up the execution. To evaluate the quality of the classification algorithms three main metrics are used, namely precision, recall, and F_1 score. Moreover, during training and testing stages, computational time was measured that is also used in the analysis of algorithms' performance.

5.1 Evaluation metrics of algorithms

The effectiveness of the classification algorithms is usually estimated based on such metrics as precision, recall, F_1 score, and accuracy. Consider the metrics that were used for calculation of the precision, recall, F_1 score, accuracy (Table 1). The confusion matrix contains the estimated and actual distribution of labels. Each column corresponds to the actual label and each row corresponds to the estimated the label of the sentence.

Table 1. Confusion matrix for a binary classifier.

		Actual	
		positive	negative
Estimated	positive	TP	FP
	negative	FN	TN

TP is the number of true positives: the sentence that is positive and was estimated as positive, *TN* is the number of true negatives: the sentence that is negative and was estimated as negative, *FP* is the number of false positives: the sentence that is negative but estimated as positive, *FN* is the number of false negatives: the sentence that is positive but estimated as negative.

Accuracy presents the proportion of the correct answers that are given by the classifier hence it can be estimated as:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision can be estimated using following formula:

Precision = TP / (TP + FP)

It shows how many positive answers that received from the

classifier are correct. The greater precision the less number of false hits. In order to take into account the latter recall is used:
Recall = TP / TP + FN
The Recall shows the ability of the classifier to 'guess' as many positive answers as possible out of the expected.

5.2 Performance statistics
This subsection describes the conducted experiments and provides the results of the classification as well as evaluation criteria of the algorithms.
5.2.1 Naïve Bayes classifier
For the Naïve Bayes classifier, all the experiment were conducted using the different amount of word for training the classifier. It is seen that on a small dataset (up to 400 words) all demonstrated metrics have lower values compared to the usage of the larger amount of words for training. However, it is also important to notice that as some point all metrics take the same value and then the decrease in values of all metrics can be observed [18]. The highest accuracy is reached when 4000 informative words are taken as features and it constitutes 80.00 per centage. Moreover, the classifier that is trained on 4000 of the best word also shows the highest values of recall and F_1 score. Recall equals to 86.74 percentage and F_1score is 82.6 percentage.

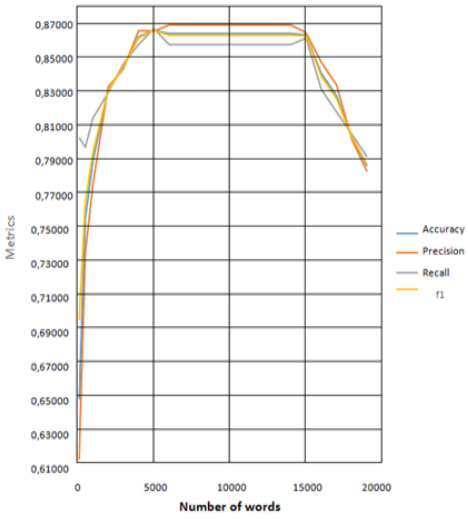

Figure 1. The evaluation of Naïve Bayes classifier that was trained on the movie reviews (the values are specified in fractions)

The highest precision is gained when 6000 words are used for learning the classifier and makes up 82.6 percentage. It is found that is a case of sentiment classification precision is more important metric because the classifier has to be precise in detecting true positive answers. Hence, the usage of 5000 words is most favorable for training the classifier on movie review in order to get the optimal performance in recognizing the positive and negative tweets.
The next test is performed using the same classifier that is trained on movie reviews, but evaluation is done on tweets. The metrics obtained after testing the classifier is illustrated in Figure 2.

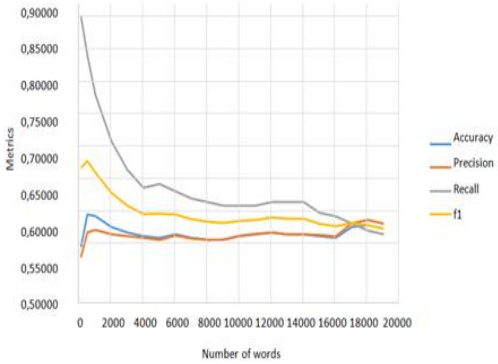

Figure 2. The evaluation of Naïve Bayes classifier that was trained on the movie reviews (the values are specified in fractions)

The figure shows all the metrics got the lower values opposed to the previous case. The highest accuracy is reached when 400 words are used for training the classifier and it equals to 58.91%. Furthermore, F_1 score gets its optimal value of 67.82 percentage if 400 words are used as features. However, the highest value of recall is gained when using only 100 words and it constitutes 90.10%. On the other hand, the optimal precision is reached when the classifier is learned from the whole dataset.

Convolutional Neural Network
The experiments are performed by employment of the convolutional neural network that has one layer and uses the randomly initialized word embeddings that are convolved with 3 different filter sizes. In the first experiment, the CNN was trained on the movie reviews and tested on tweets [19-21]. Results are shown in Table 2.

Table 2. Evaluation of the CNN that is trained on the movie reviews (the values are specified in fractions).

	Accuracy	Precision	Recall	f1
CNN movie reviews	0.53	0.63	0.52	0.57

The accuracy is 53.1% that is a bit better that what was obtained using the Naïve Bayes classifier (trained on movie reviews and tested on tweets). Therefore, the accuracy is 1.3% higher opposed to Naïve Bayes. CNN did not show great performance on movie review dataset, because usually neural network requires larger dataset for training. Hence, it is not enough data for the model to generalize well an unseen samples that leads to such insignificant results that CNN produced.

CNN model shows a 5.38% increase in accuracy compared to the Naïve Bayes classifier and it makes up 76.01%. The growth of the recall and F_1 score are also observed and they constitute 85.05% and 86.06% respectively. Hence, an improvement of recall is almost 10% and F_1 score enhancement is almost 9%. However, the slight decrease of precision is demonstrated by CNN classifier, in this case precision is 76.01%.

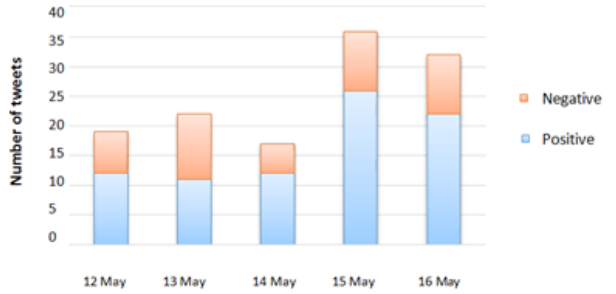

Figure 3. The Number of tweets verses time period

It can be observed that Naïve Bayes approach shows good results. Nonetheless, CNN outperforms the Naïve Bayes a bit in Figure 4. As mentioned above, when dealing with sentiment classification task, the precision is the metric that has to be high in order to define true sentiment expressed in the sentence, in this case, recall can deteriorate. Therefore, analysis of the results shows that investigated models may be further improved because metrics of the accuracy, precision, recall and F_1 score are not significant as they were expected, especially when employing CNN classifier.

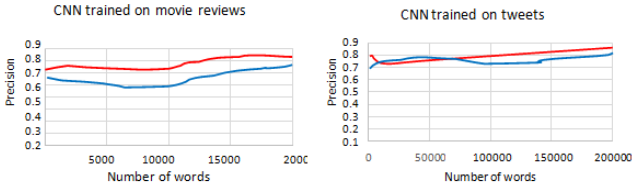

Figure 5. Comparison of Earlier Proposed system and the Current Proposed system in the form of precision value of classifier

The classifier that is trained on the diverse data with different context will highly probably be able to detect correct sentiment when it is tested across all domains. Hence, the quality of the dataset has an enormous impact on the effectiveness of the classification model. It is clearly depicted from the Figure 5 that the red line showing the proposed system performance and the blue line showing the earlier outcomes on the sentimental analysis. The proposed system gives around 91.5 percentage successful results which is around 2.5 percentage better than the earlier study.

VI. CONCLUSION

Sentiment analysis task is under research since the early 1900s and it is still in developing phase, especially the exploration of microblogs, such as Twitter. Twitter message is less informative opposed to usual review or comment and also contains a lot of noisy data that makes classification of tweets more challenging. This research study investigates the algorithms that can be used for sentiment classification. The analysis of both algorithms was carried out and their performance was estimated. The classification model was trained on two different datasets in order to study whether sentiment classification is the domain-dependent task or not. Additionally, this research work shows that in order to achieve meaningful performance of the classifier it has to be trained and tested on the same type of the dataset because the correlation exists between the classifier performance and domains, which are used for collecting training and testing samples.

REFERENCES

[1] Liu, B. (2016). Sentiment analysis and opinion mining. *Synresearch study lectures on human language technologies*, *5*(1), 1-167.

[2] Medhat, W., Hassan, A., & Korashy, H. (2018). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113.

[3] Go, A., Bhayani, R., & Huang, L. (2017). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12).

[4] Bütow, F., Schultze, F., & Strauch, L. Semantic Search: Sentiment Analysis with Machine Learning Algorithms on German News (2017).

[5] Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREc* (Vol. 10, No. 2014).

[6] Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A. (2012, December). Opinion mining and sentiment analysis on a twitter data stream. In *Advances in ICT for emerging regions (ICTer), 2016 International Conference on* (pp. 182-188). IEEE.

[7] Hallsmar, F., & Palm, J. (2016). Multi-class sentiment classification on twitter using an emoji training heuristic.

[8] Turney, P. D. (2012). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417-424). Association for Computational Linguistics.

[9] Salas-Zárate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez- García, M. Á., & Valencia-García, R. (2017). Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. *Computational and mathematical methods in medicine*, *2017*.

[10] Chiavetta, F., Bosco, G. L., & Pilato, G. (2016). A Lexicon-based Approach for Sentiment Classification of Amazon Books Reviews in Italian Language.

[11] Hailong, Z., Wenyan, G., & Bo, J. (2015). Machine learning and lexicon based methods for sentiment classification: A survey. In *Web Information System and Application Conference (WISA), 2014 11th* (pp. 262-265). IEEE.

[12] Hu, M., & Liu, B. (2014). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

[13] Miller, G. A. (2005). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

