International Journal of Emerging Trends in Science and Technology
6 (2) December 2020, 38-41
http://www.publishingindia.com/ijetst

# Adaptive Word Embedding to Reduce the Dimensionality of the Document to Vector Representation

M. Gunasekar[1*], M. Dhayalan[2], N. Pradeep[3], S. Sakthivel[4] and R. Venkatesh[5]
[1]Assistant Professor, Department of Information Technology, M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India. Email: guna18it@gmail.com
[2]UG Scholar, Department of Information Technology, M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India. Email: ramdhayalan3@gmail.com
[3]UG Scholar, Department of Information Technology, M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India. Email: pradeepnrnp10@gmail.com
[4]UG Scholar, Department of Information Technology, M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India. Email: sakthivels@gmail.com
[5]UG Scholar, Department of Information Technology, M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India. Email: venkatrajarvvg@gmail.com
*Corresponding Author

**Abstract: Sentiment Analysis is a methodology of detecting the emotions from the text. It is an application of Natural Language Processing (NLP) methodology. The NLP enables us to know the common day to day language of the people. This will helps to decipher the sentiments of the users and hence explain liking and disliking of the people. The traditional bag-of-words models lack the accuracy of sentiment classifications. The intention of this project is to improve the accuracy of the sentiment classification by employing the concept of dimensionality reduction. Reducing the dimensionality of a large document helps to reduce the computational cost and increase efficiency. Word embedding methods capture the context of a word in a document which helps to reduce the dimensionality of text data. Vector representation of the words using a technique like Word2Vector proves to be very effective in interpreting the meaning and hence the sentiments. The words in the document will be converted into vectors. Each word is assigned a unique value (vectors) such that these vectors represent its context, meaning, and semantics. The resulting word vectors are wont to train machine learning algorithms within the sort of classifiers for sentiment classification. We use the Machine Learning classifier Naive Bayes to analyze the sentiment from the given pre-processed dataset (word vectors). Our experiments on real-world datasets show the improvement in the accuracy of sentiment classification using the word embedding techniques.**

**Keywords: Dimensionality reduction, Sentiment analysis, Vector representation, Word embedding.**

## I. Introduction

Human Language is notably complex and very different. We express our nature in many ways, particularly in verbal and writing. We are using different languages to express ourselves, within each language there will be a unique set of grammar and syntax rules, terms and slang are available. When we write we often makes mistakes. When we speak, we have accents that are corresponding to the region and even we do activities like mumbling, stuttering, and mixed terms from other languages. Machine learning algorithms are mostly used to make computers to understand human language. But the human languages have more syntactic and semantic information that are hard for computers to understand. Each language has their own properties that are not necessarily present in these machine learning approaches. NLP resolves the vagueness in language and helps for many applications like speech recognition, sentiment prediction, etc. Building a machine learning model for analyzing the text helps in finding the emotions of the people. But these models are lack in accuracy because analyzing large text data corpse at a single time cant able to get greater accuracy. This accuracy can be increased by using word embedding techniques along with the machine learning model. We take a real-world dataset for training the model and produce the emotions that are expressed in the dataset.

## II. Related Work

The Emotional classification is based on three categories. They are Rule-based analysis, unsupervised classification and supervised classification.

The *Rule-based Analysis* is performed based on the rules of human that are made specifically for the situation and it reflects human intelligence. The rule based systems requires source data, set of rules. The rules will be like IF statement. Kmaps proposed a method to measure the distance to determine the semantic polarity of the adjectives in the English language which are based on the meaning graph model. Zhu *et al.* introduced a method to get the semantic direction of Chinese words which depends on HowNet Chinese Lexicon. Pan *et al.* identified six various kinds of emotion which are conveyed by Weibo with the help of a lexicon-based method. But the method based on lexical analysis has low accuracy and the quality of the classification will be limited by the lexicon and also this method ignores contextual information.

*Unsupervised Classification* analysis depends on the statistical properties of a documentation NLP process and predefined vocabulary which contains emotional or polarizing tendency. It does not use tagged documents for the classification. Turney presented an unsupervised learning algorithm for classifying movie reviews. Ling proposed a novel on the probabilistic modeling framework which is depend on Dirichlet allocation (DLA), it is also called as the Joint sentiment/topic model (JST), which can able to detect sentiments from the text reviews. They also explored different ways to get required information for the improvement of accuracy of emotional detection. Yili Wang and Hee Yong Youn used a feature weighting approach for the sentimental analysis to increase the accuracy of the analysis.

*Supervised Classification* is a classification model which produces emotions with the help of labeled data. The labeled data consists of words that have been marked as negative or positive. We will use bag of words for the labeled words, it provide a simplest representation of documents. But these methods have limitation during the analysis which leads to lower accuracy and lacks of performance. In order to increase the accuracy and performance we go for word embedding, it will reduce the dimensions and increase the performance of the model by creating a low dimensional word vectors.

The word vectors are produced for each individual word and each vector will be unique. These vectors are used in encoding of linguistic patterns. Linguistic patterns are necessary for the classification of the words. But these vectors are not alone necessary to produce a prediction from the reviews. We also need other information like auxiliary information which helps to improve the performance as well as the accuracy of the model.

## III. Proposed System

The intention of this project is to analyze a large number of movie reviews and categorize the reviews into three categories: positive, negative. There are several methods are available for sentimental classification but when there is large sized dataset and reviews from various languages the system fails to give accuracy. Large dataset takes more computational cost and also the system produce results with less accuracy. In order to give more accuracy we need to reduce the dimensionality of the documents. We use word embedding techniques like Word2Vector to change the words into vectors which helps to analyze the reviews easily. Each words in the review is assigned to a numerical value which is unique for the faster analyze of the dataset. Our model will be trained with the predefined dataset which has all reviews as vectors. By using word embedding along with sentimental classifier we can be able to produce results with more accuracy.

## IV. Implementation

The implementation involves preprocessing the dataset and analyzing of the dataset.

### *A. Dataset*

The dataset contains 65536 reviews these reviews are based on different kind of situation on all topics. The dataset contains collections of review which have all positive, negative review are grouped together. We will use the trained model and analysis them. It helps to determine whether the reviews are negative or positive.

### *B. Algorithm*

We have used Principal Component Analysis algorithm for the proposed scenario. PCA is applied to reduce the dimension of the dataset when the dimensions of the input are large and the components are correlated.

### *C. Module Description*

The following system flow (Fig. 1) describes the functional flow of our analysis.

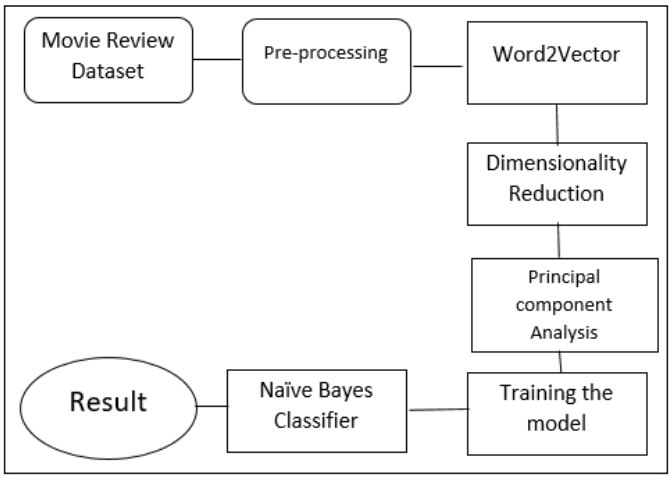


Fig. 1: Work Flow

*i. Word Embedding-Word2Vector*

Word2Vec vectors are generated for each word in the reviews dataset. We will be using average over all the vectors of words in a review sentence from the dataset. The generated vectors are stored in csv file. It can be done directly in data frame but when there is a large amount of data it is a best option to write on file. It will help if the code breaks we can start from the point where it had broken.

*ii. Dimensionality Reduction*

The reviews in the document have more features which lead to a high dimensional space. The main function of dimensionality reduction is to reduce the dimension in feature space and maintain the accuracy. It leads to the removal of irrelevant features and result in easier analysis of sentimental classification after reduction. Sentimental analysis processes the extracted tokens from reviews and stop words are removed.

*Components*

Dimensionality reduction consists of two main components:

- *Feature Selection:* It is used to remove or cutoff additional attributes that are available while building the model. Reducing the attribute will helps to reduce the dimensions.
- *Feature Extraction:* It extracts the feature from the model and helps to reduce the data from a high dimensional space to a lower dimensional space.

*iii. Classifier*

We are using Naive Bayes classifier for the classification.

Naïve Bayes classifier is a machine learning technique which uses Bayes theorem. It predict the probabilities by classifies text into classes. The text will be predicted and put into separate classes. Then the class with higher probability will be considered. Each word will be fall into any category here we have negative and positive class. When the word has a higher probability in any of the classes will be considered as output.

We take two class as positive and negative without the analysis of the input. It shows that the text is positive or negative. Each class will be counted for relative frequencies and the final calculation is done.

Consider a review such as “The movie is awesome”. We are going to use our model to predict the sentence nature. We will use the Naïve Bayes classifier for the classification.

The bag of words contains both negative and positive words as well as the words frequency counts. We have to take each word in the sentence and compare it with the negative and positive words in the bag of words. The probability for both the negative and positive will be calculated. The highest probability that the words gets will take into considerations.

If the probability of the positive words is high then the review will be positive.

P (“The movie is very good to watch”) = P (the)* P (movie)* P (is)* P(very)* P(good)* P(to)* P(watch).

By removing stop words, the words will be reduced.

- P (the | positive) * P (movie | positive) * P (is | positive) * P (very | positive) * P (good | positive) * P (to | positive) * P (watch | positive)
- After the calculation we get the probability for positive words
- P (the | negative) * P (movie | negative) * P (is | negative) * P (very | negative) * P (good | negative) * P (to | negative) * P (watch | negative)

We have also calculated the probability for negative words. Once we get both the probabilities we need compare them. The words with higher probability will be taken. After that we can able to say that the given review is positive or negative.

## V. Results

We present the analysis of the results obtained from the experiments. We have proposed a model for sentiment analysis of product reviews using Word2Vector and Multinomial Naive Bayes. It can be observed by Fig. 2. The statistical outcome of analyzing the document and shows the positive and negative level from the dataset.

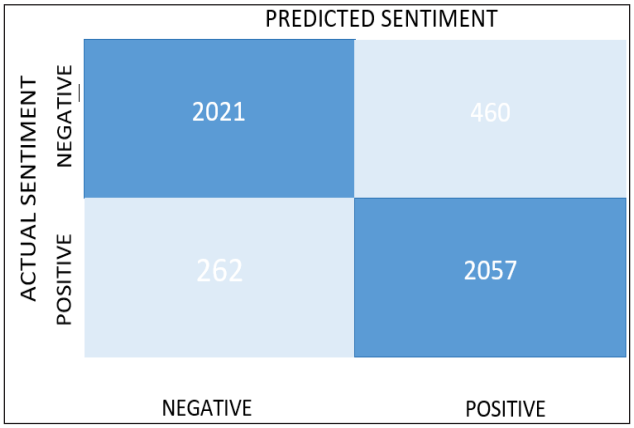


Fig. 2: Result

## VI. Conclusion and Future Work

The proposed system results to the successful development of adaptive word embedding with large datasets and identify the likes and dislikes with better accuracy. The results are higher than expectations. The system is also fast enough and even better than the existing system. Finally the results of the experiments show greater accuracy when compared to traditional models. In future, our work will aim to show the functionality of the flexibility and efficiency of the model and also focusing on improving the performance of the model. The model will be trained with other languages to understand more and to learn new syntactic and semantic ways of languages. It helps to combine different language word and increase the efficiency.

## References


[1] B. Pang, and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.

[2] A. Pak, and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010,* Valletta, Malta, May 17-23, 2010.

[3] J. Khimar, and M. Kinikar, "Machine learning algorithms for opinion mining and sentiment classification," *International Journal of Scientific and Research Publications*, vol. 3, no. 6, pp. 1-6, Jun. 2013.

[4] R. Mehra, M. K. Bedi, G. Singh, R. Arora, T. Bala, and S. Saxena, "Sentimental analysis using fuzzy and Naïve Bayes," *2017 International Conference on Computing Methodologies and Communication (ICCMC),* Jul. 2017.

[5] B. Liu, E. Blasch, Y. Chen, D. Shen, and G. Chen, "Scalable sentiment classification for big data analysis using Naïve Bayes classifier," *2013 IEEE International Conference on Big Data*, Silicon Valley, CA, USA, Oct. 6-9, 2013.

[6] S. Rana, and A. Singh, "Comparative analysis of sentiment orientation using SVM and Naïve Bayes techniques," *2016 2nd International Conference on Next Generation Computing Technologies (NGCT),* Dehradun, India, Oct. 14-16, 2016.

[7] A. Goel, J. Gautam, and S. Kumar, "Real-time sentiment analysis of tweets using Naïve Bayes," *2016 2nd International Conference on Next Generation Computing Technologies (NGCT),* Dehradun, India, Oct. 14-16, 2016.

[8] H. Parveen, and S. Pandey "Sentiment analysis on twitter data-set using Naïve Bayes algorithm," *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT),* Bangalore, India, Jul. 21-23, 2016.

[9] V. Vryniotis, "Machine learning tutorial: The multinomial logistic regression (Softmax Regression)," 2013.

[10] N. Zainuddin, and A. Selamat, "Sentiment analysis using support vector machine," *2014 International Conference on Computer, Communications, and Control Technology (I4CT),* Langkawi, Malaysia, Sept. 2-4, 2014.

[11] T. Gunasekhar, and K. T. Rao, "EBCM: Single encryption, multiple decryptions," *International Journal of Applied Engineering Research*, vol. 9, no. 19, pp. 5885-5893, 2014.

[12] K. T. Rao, P. S. Kiran, and L. S. S. Reddy, "High-level architecture to provide cloud services using green data center," *Advances in Wireless and Mobile Communications (AWMC)*, 2014.

[13] K. T. Rao, P. S. Kiran, D. L. S. S. Reddy, V. K. Reddy, and B. T. Rao, "Genetic algorithm for energy placement of virtual machines in cloud environment," *Proceedings of the IEEE International Conference on Future Information Technology*, 2012.

[14] W. P. Ramadhan, S. T. M. T. Astri Novianty, and S. T. M. T. Casi Setianingsih, "Sentiment analysis using multinomial logistic regression," *2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)*, Yogyakarta, Indonesia, Sept. 26-28, 2017.

[15] V. A. Kharde, and S. Sonawane, "Sentiment analysis of twitter data: A survey of techniques," *International Journal of Computer Applications*, vol. 139, no. 11, pp. 5-15, Apr. 2016.

[16] P. V. V. Kishore, S. R. C. Kishore, and M. V. D. Prasad, "Conglomeration of hand shapes and texture information for recognizing gestures of Indian sign language using feed forward neural networks," *International Journal of Engineering and Technology*, vol. 5, no. 5, pp. 3742-3756, 2013.