International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Concept of TF-IDF, Common Bag of Word and Word Embedding for Effective Sentiment Classification

Swamy L N, J V Gorabal

Abstract*: Sentiment Classification is one of the well-known and most popular domain of machine learning and natural language processing. An algorithm is developed to understand the opinion of an entity similar to human beings. This research fining article presents a similar to the mention above. Concept of natural language processing is considered for text representation. Later novel word embedding model is proposed for effective classification of the data. Tf-IDF and Common BoW representation models were considered for representation of text data. Importance of these models are discussed in the respective sections. The proposed is testing using IMDB datasets. 50% training and 50% testing with three random shuffling of the datasets are used for evaluation of the model.

***Keywords* : Sentiment Analysis, Word Embedding, Machine Learning.**

I. INTRODUCTION

Knowing others opinion about a problem is very important information for manufacturing company[1]. Before internet, people used request for recommendation or suggestion before going for a product. But internet made this things easy and fast due to application of machine learning to understand and estimate the knowledge in text data. In this directions, automatic sentiment classification can be achieved using document classification. Document classification for set of n classes required a knowledge of n different classes[2-3]. A classifier will be trained using these information. Classification function will be defined for classifying unknown samples into one of the predefined classes based on the ontology present in it. [4-5]

Generally, the sentiment about an entity which includes movie, a product, a travel destination, opinion about a person, home appliance etc can be categorized into two categories: Negative and Positive or in other manner Bad or Good[6]. A systematic study towards an entity which includes movie, a product, a travel destination, opinion about a person, home appliance etc, will help a manufacturer to approximate the extent of a product.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Swamy L N*, Research Scholar, Department of CSE, Sahyadri College of Engineering & Management, VTU-Belagavi, Mangalore, Karnataka, India. swamyln@gmail.com
Dr. J V Gorabal, Department of Computer Science & Engineering, Sahyadri College of Engineering & Management, Mangalore, Karnataka, India, jvgorabal@gmail.com

It also helps the manufacturer to understand the market acceptance about the product. If the extent of the product is acceptable, he can go for increase the production, or he can determine the strategies to improve the quality of the product. Sentiment analysis can also help politicians to understand the public sentiment with respect to a specific policy or political views.

In machine learning, a sentiment classification can be designed by training set of positive and negative reviews about a product[7]. A set of test is considered to validate the ffective of the system. The classifier used here will focus on the optimizing both class parameters value with set of binary values or with specified weights. It also requires features represented as attribute – value pairs. Most popular support vector machine construct a hyper plane by optimizing the training classes. For specific applications of sentiment analysis like mobile reviews classification requires rule based classifiers. It is because, in such application specific features need to be compared and there will be standard value for specific features. For example: Feature - mobile battery. One major drawback of rule based classification is it requires large amount of data and huge time for optimizing the classifier.

Based on the importance of sentiment analysis and application of machine learning, in this article address this with novel word embedding model. Major stages involved in this article include, pre-processing of input data. Representation of Sentiments using standard Tf/Idf and CBow Models. Construction of classification model.

It is quite clear that, text data is the common media for exchange of information. Using text data we all communicate each other, exchange our feeling etc. Input text data need to be formatted and transferred to machine understandable format. The feature representation models life Tf/Idf and CBow Models will help achieve this. Once the text data transformed into numerical data, then a classifier will be designed to for classification.

This article is organized as follows. Section two Literature details, three proposed model for sentiment analysis. Experimentation and conclusion in section four and section five respectively.

II. LITERATURE SURVEY

Many researchers have done promising contribution to the field of Sentiment Analysis towards forecasting the price of the products using machine learning algorithms. There is large amount of research data is available in this direction.

But based on aim of this article we are focusing on text representation and classification for sentiment analysis. [8]
Based on the literature survey it is clear that, the main forecasting is towards enhancing the accuracy of the system by properly detecting the false detections.
Among machine learning algorithms, Support vector machine is one of the dominant algorithm which has shown significant results in classification of sentiments. SVM's hyper planes separate the positive and negative class of information among the datasets. Finding the best hyper plane can be treated as an optimization algorithm. Many approaches have found with SVM and Information gain as a feature selection method. Generally, word frequency is considered than word probability.[9]
Another important model is based on N-gram based language model with the support of NLP for achieving good results in sentiment analysis. Here language is considered as major feature than the terms. Basically this method is derived from N-gram model. It is core concept of natural language processing used for sentiment analysis. This model also provides the probability distribution among the positive and negative classes. [10]
In association with the opinion mining application, the majority of the research indicates usage of machine learning and semantic orientation of the input text data. Majority of the algorithms are based on the supervised learning based approaches along with feature selection. Few works based on unsupervised approaches can also be found. In [] the word is closely related to finding the semantic orientation of objectives. It is based on the fact that there is a language constraints on the language meaning and its orientations of adjectives in conjunctions.
Considering the supervised algorithms on large data corpus, time complexity depends on training the data. As an alternative to it, unsupervised algorithms approaches are proposed by many researchers for as alternatives. Key observations with his approach is performance depends on domain knowledge dependencies. The task of sentiment analysis of diverse data with less data labelled. Many solutions to this problem can also be found using deep learning algorithms.[12]

III. PROPOSED MODEL

Based on the importance of sentiment analysis and application of machine learning, in this article address this with novel word embedding model. Major stages involved in this article include, pre-processing of input data. Representation of Sentiments using standard Tf/Idf and CBow Models. Construction of classification model. It is quite clear that, text data is the common media for exchange of information. Using text data we all communicate each other, exchange our feeling etc. Input text data need to be formatted and transferred to machine understandable format. The feature representation models life Tf/Idf and CBow Models will help achieve this. Once the text data transformed into numerical data, then a classifier will be designed to for classification. To achieve good results in sentiment analysis, embedding model behind the text representation plays an important role. Here novel word embedding with TF/IDF and Common BoW representation model is proposed. All the stages of the models is shown in Fig 1.

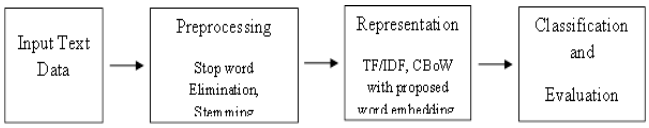

Fig. 1 Different stages of the proposed Word Embedding for Sentiment Classification

Input Data: In the early stages of the system, user's reviews are collected. All the user's reviews need to be transformed into classifier understandable format for classification. Hence, the input data will be subjected for preprocessing.
Preprocessing: In this stage, the widely used preprocessing techniques like stemming and stop word elimination are applied to make the data ready for processing.
Representation: This is one of the important stage of the system, which involves word embedding stage for sentiment classification. Hence this stage will be divided into several sub stages like word embedding stage and representation stage.
Word Embedding stage: Textual data is the common plat for expressing the information, sharing of views and sentiment exchange. Hence understanding text data for sentiment mining application is crucial task. One of the major issue we faced with understanding text data is, ways of represent the language i.e., a sentence can be represented in 'n' different ways.
Let us consider an example sentences:
S1 = 'This movie is not good'. and S2 = 'It is a bad Movie'. It is easy for us to understand the S1 and S2 but making system to understand is difficult. To tackle such situation, we are proposing our own embedding model at word level. Each word in the sentence will be processing to find the embedding weight with the support of global positive and negative dictionaries. Once the word embedding is applied, each words contribution to the polarity of the sentence will be added as a weightage to the representation model. This technique will add good significance to each word and its contribution to the respective sentence.
Let us understand the proposed word embedding model in detail. A global D_positive and D_negative dictionaries will be designed for processing of text documents. Uni-Gram representation model will be considered to obtain the bi-gram and tri-gram data from input. One of the major reasons behind selecting bi-gram and tri-gram word model is to capture the semantics of the input data at word level. Each term (bi-gram and tri-gram) is compared with D_positive and D_negative to get the positive score and negative score using Jaccard similarity.
The processing of obtaining Jaccard similarity is equated in eq. (1)

$$Jaccard\ Similarity_{WordLevel} = \frac{|Sentence1 \cap Sentence2|}{|Sentence1|+|Sentence2|-|Sentence1 \cap Sentence2|} \quad (1)$$

Jaccard Similarity is one of the popular proximity measure, used to measure how different given the two sentences are? Interest thing here is, N-Gram model will capture the context of reach word and Jaccard Similarity will provide the jaccard index for the considered terms.

It's a kind of extracting the similar words with same context in the sentences. For normalization, we have restricted the jaccard index JacIndex value between 0<JacIndex<1. As result of this stage, each word will have two JacIndex associated with D_positive and D_negative. Based on the two JacIndex from D_positive and D_negative max pooling will be applied. Hence the JacIndex with max value will be added as weightage to text vector. To create a hyper plane between positive samples and negative samples, positives indexes are added and negative indexes are subtracted during text representation. All the above processes will be illustrated in illustration - 1.

Dpositive = {term1, term2, term3, …. , termn}
Dnegative = {term1, term2, term3, …. , termn}
Input: {word1, word2, word3, word4, word5, word6, word7, word8, word9, word10}
Preprocessing: { word1, word3, word4, word5, word6, word7, word9, word10} // stop words are eliminated
Bi-gram model: {[word1, word3], [word3, word4,], [word4, word5], [word5, word6], [word6, word7], [word7, word9], [word9, word10]}
Tri-gram model : {[word1, word3, word4], [word3, word4, word5], [word4, word5, word6], [word5, word6, word7], [word6, word7, word9], [word7, word9, word10]}
Now, consider the term [word1, word3] from bi-gram. It will be subjected for obtaining JacIndex. As a result the term [word1, word3] will have two values with respect to two dictionaries.
These JacIndexes will be added / subtracted based on the polarities to the TF/IDF and CBoW vectors. Hence a proper context sensitive word embedding model can be obtained for sentiment classification problem. The proposed word embedding model is evaluated with standard state of art embedding models. The details of the experiments, details of the datasets will be presented in the respective section of the article.

IV. EXPERIMENT SETUP AND RESULTS

Any machine learning algorithm need to be evaluated for its proficiency using globally accepted datasets. Hence in this article, we are also conducted a serious of experiments, the details of these experiments can be found in current section. For the proper evaluation of the proposed word embedding model, IMDB movie review dataset. It is one of the largest dataset consists of 50,000 data samples can be used for training and testing of algorithms. For the sake of experimentation, dataset is splitted into training and testing ratios. Based on few journals in this domain, the dataset is splitted into 50% Training and 50% testing. Three sets of such samples sets prepared by shuffling the data samples among training and testing. Also the proposed model is evaluated based precision, recall and f-measures metrics over the scale between 0 to 1. Table – 1 presents the details of the experiments conducted for demonstrating the efficiency of the system. The Fig 2: F – Measure Scores of the proposed system over IMDB dataset with TF/IDF and Fig 3 : F – Measure Scores of the proposed system over IMDB dataset with CBoW

Table. 1 F – Measure Scores of the Proposed system over IMDB dataset.

IMDB DATASET WITH 50:50	TF/IDF	TF/IDF with word embeddin g	CBow	CBow with word embedding
	F-Measure	F-Measure	F-Measure	F-Measure
Train 1 50 : 50 Random Shuffle Set 1	0.8476	0.9087	0.8709	0.9395
Train 2 Random Shuffle Set 2	0.8610	0.9090	0.8916	0.9412
Train 3 Random Shuffle Set 3	0.8453	0.9157	0.8730	0.9543

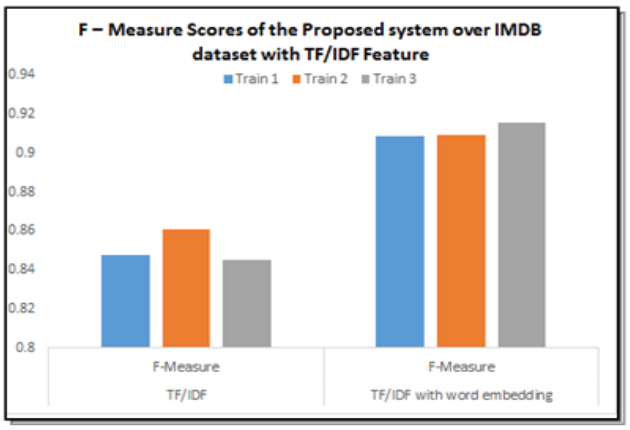

Fig1: Measure Scores of the proposed system over IMDB dataset with TF/IDF

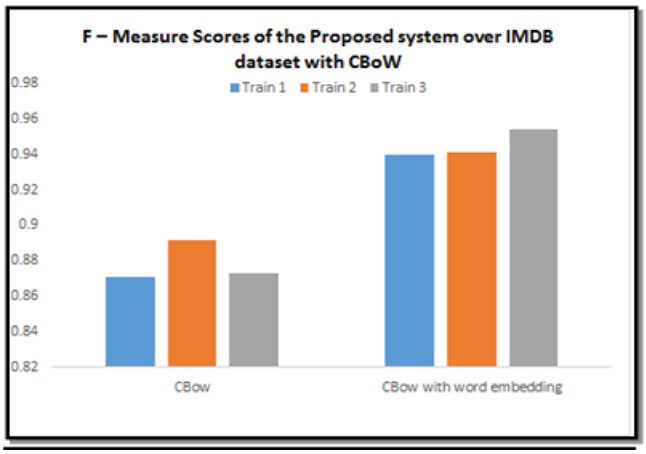

Fig 3 : F – Measure Scores of the proposed system over IMDB dataset with CBoW

V. CONCLUSIONS

This article presents the novel word embedding model for sentiment classification applications. Sentiment classification is most popular application of natural language processing which is having huge scope for improvements. Here we have made an effort to capture the knowledge with word embedding and state of the art text representation models. Four different types of experiments are carried out to present the goodness of the system.

Based on these experiments the proposed model is capable of producing f-measures as 0.8453, 0.9157, 0.8730 and 0.9543 for TF/IDF, TF/IDF with word embedding, CBow, CBow with word embedding respectively. The proposed model is having scope for enhancement in many directions. It can be enhanced in text representation and developing model particularly for the proposed model. Machine Learning is a big domain, which has having many applications. Some of them are [13-18]

REFERENCES

1. Ravi, K. and Ravi, V., 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems, 89, pp.14-46.
2. Singh, P.K., Sachdeva, A., Mahajan, D., Pande, N. and Sharma, A., 2014, September. An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites. In 2014 5th International Conference-Confluence The Next Generation Information Technology Summit (Confluence) (pp. 329-335). IEEE.
3. Sharma, R., Nigam, S. and Jain, R., 2014. Supervised opinion mining techniques: a survey. The Proceedings of the International Journal in Foundations of Computer Science & Technology (IJFCST), 4(3).
4. Bhushan, S.B. and Danti, A., 2017. Classification of text documents based on score level fusion approach. Pattern Recognition Letters, 94, pp.118-126.
5. Bhushan, S.B. and Danti, A., 2018. Classification of compressed and uncompressed text documents. Future Generation Computer Systems, 88, pp.614-623.
6. Li, Z., Liu, L. and Li, C., 2015, September. Analysis of customer satisfaction from Chinese reviews using opinion mining. In 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS) (pp. 95-99). IEEE.
7. Kang, M., Ahn, J. and Lee, K., 2018. Opinion mining using ensemble text hidden Markov models for text classification. Expert Systems with Applications, 94, pp.218-227.
8. Shelke, N.M., Deshpande, S. and Thakre, V., 2012. Survey of techniques for opinion mining. International Journal of Computer Applications, 57(13), pp.0975-8887.
9. Devi, D.N., Kumar, C.K. and Prasad, S., 2016, February. A feature based approach for sentiment analysis by using support vector machine. In 2016 IEEE 6th International Conference on Advanced Computing (IACC) (pp. 3-8). IEEE.
10. Zhang, L. and Liu, B., 2014. Aspect and entity extraction for opinion mining. In Data mining and knowledge discovery for big data (pp. 1-40). Springer, Berlin, Heidelberg.
11. Balaji, P., Haritha, D. and Nagaraju, O., 2018. An overview on opinion mining techniques and sentiment analysis. International Journal of Pure and Applied Mathematics, 118(19), pp.61-69.
12. Riaz, S., Fatima, M., Kamran, M. and Nisar, M.W., 2019. Opinion mining on large scale data using sentiment analysis and k-means clustering. Cluster Computing, 22(3), pp.7149-7164.
13. J V Gorabal and Manjaiah D H, RFID A Smart Technology, International Journal of Computer Engineering Technology (IJCET), Volume 5, Issue 8, August (2014), pp. 18-24.
14. J. V. Gorabal and Manjaiah D. H., and S. N. Bharath Bhushan Novel Implementation of multimodal Biometric Approaches to handle Privacy and security Issues of RFID Tag May 2016 Special Issue Volume 4,Issue 4 April 2017
15. Ankitha ,Manjaiah D H , J V Gorabal ," Survey on Animal Healt Care System in Veterinary Science using Internet of things ,International Journal of Emerging Research in Management and Technology. Volume-6,Issue-5 May 2017
16. Rajeshwari Banni & J V Gorabal " Citrus Leaf Disease Detection Using Image Processing Approaches : International Conference on Recent Innovations in Science, Engineering and echnology Conference held at: Hirasugar Institute of Technology,Nidasoshi-18-19 May 2018 Published in Scopus Index Joournal
17. L N Swamy and Dr J V Gorabal " Product Review analysis using LR Method" presented in third IEEE International Conference on Electrical and Electronics , Communication, Computer Technologies and optimization Techniques; 14-15 December 2018.
18. Swamy L N, Dr. J V Gorabal, "Logistic Regression based Classification for reviews analysis on E-commerce based Applications" 7th International Conference on Frontiers of Intelligent Computing: Theory And Application (FICTA 2018), Springer AISC series on November 29th - 30th 2018.

AUTHORS PROFILE

SWAMY L N is currently working as an Assistant Professor in Department of CSE (Master of Computer Applications), Post Graduate Studies – Mysuru Region, Visvesvaraya Technological University, Mysuru, Karnataka.. Currently pursuing PhD at Visvesvaraya Technological University, Belagavi. He completed his Master degree (M.Tech. in Computer Science and Engineering) from SJC Institute of Technology, Chickaballapur (VTU University, Belagavi) in the year 2010. He is a Life Member in various Professional Societies like CSI, ISTE, IEI, InSc, IAENG, IACSIT, ICSES, ASR, STRA and SDIWC.

Dr. J V Gorabal holding Ph.D in Computer Science and Engineering is working as an Professor and Head of Department of Computer Science & Engineering at Sahyadri College of Engineering & Management, Mangalore. His area of interest includes RFID A Smart Technology, Sentimental Analysis, Artificial Intelligence, Machine Learning. He published various papers research areas. He is a Life Member in various Professional Societies like Computer Society of India, Indian Society for Technical Education (ISTE).