SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ DERGİSİ

Sakarya University Journal of Science
SAUJS

e-ISSN 2147-835X | Period Bimonthly | Founded: 1997 | Publisher Sakarya University |
http://www.saujs.sakarya.edu.tr/en/

Title: Sentiment Classification Performance Analysis Based on Glove Word Embedding

Authors: Yasin KIRELLİ, Şebnem ÖZDEMİR

Recieved: 2021-02-25 17:30:10

Accepted: 2021-04-02 09:21:04

Article Type: Research Article

Volume: 25
Issue: 3
Month: June
Year: 2021
Pages: 639-646

How to cite
Yasin KIRELLİ, Şebnem ÖZDEMİR; (2021), Sentiment Classification Performance Analysis Based on Glove Word Embedding. Sakarya University Journal of Science, 25(3), 639-646, DOI: https://doi.org/10.16984/saufenbilder.886583
Access link
http://www.saujs.sakarya.edu.tr/en/pub/issue/62736/886583

Sakarya University Journal of Science 25(3), 639-646, 2021

Sentiment Classification Performance Analysis Based on Glove Word Embedding

Yasin KIRELLİ*1, Şebnem ÖZDEMİR1

Abstract

Representation of words in mathematical expressions is an essential issue in natural language processing. In this study, data sets in different categories are classified as positive or negative according to their content. Using the Glove (Global Vector for Word Representation) method, which is one of the word embedding methods, the effect of the vector set based on the word similarities previously calculated on the classification performance has been analyzed. In this study, the effect of pre-trained, embedded and deterministic word embedding classification performance has analyzed by using Long Short-Term Memory (LSTM). The proposed LSTM based deep learning model has been tested on three different data sets and the results have been evaluated.

Keywords: Sentiment classification, word embedding, word weight, glove word embedding.

1. INTRODUCTION

As a result of the widespread use of digital platforms, users share their feedback and opinions on social platforms, shopping sites, blogs, and many other channels. Sentiment analysis has been a research subject of interest in extracting the outputs produced from texts in this respect not only in data mining but also in natural language processing. The perspective of users can be expressed mathematically by using their own shared words. Those statements can be positive, negative or neutral, and the theme of any outcome can be calculated in this sense [1]. For this reason, sentiment analysis, idea mining and text analysis can often be used interchangeably in this field. The main research subject of sentiment analysis is to make sentiment classification according to the stored text data on the platforms and to discover the idea based information accordingly. Not every comment gives a positive or negative opinion, finding the appropriate data set for better analysis is an important detail at this point. It is crucial to analyze opinions as areas of use: It is frequently preferred in primary sectors such as e-commerce, banking, advertising management [2,3].

In Natural language process research, words are represented vectorially with word embedding [4]. Word embedding allows words with the same meaning to be represented in a similar way. The aim is to create a coordinate system in which the relevant words are shown closer to each other. Word embedding is generally divided into two

*Corresponding author: yasin.kirelli@istinye.edu.tr
1İstinye University, Faculty of Economics, Administrative and Social Sciences, Management Information Systems Department, İstanbul.
E-Mail: sebnem.ozdemir@istinye.edu.tr
ORCID: https://orcid.org/0000-0002-3605-8621, https://orcid.org/0000-0001-6668-6285

different groups, frequency-based and prediction-based. It has been observed that probability based methods give more successful results in expressing similar words with close vectors compared to deterministic methods in word vector determination. Glove (Global Vectors for Word Representation) is a predictive word embedding techniques and is an unsupervised learning algorithm applied to obtain vector representations [5]. This technique has been developed by Stanford University. With this method, which consists of the synthesis of matrix factorization and skip-gram methods, a co-occurrence matrix is obtained. This matrix is based on the use of real corpus in subsequent estimates [1,6].

Long Short Term Memory is a deep learning methodology that created idea of having a memory for the learning process. Unlike CNN (Convolutional Neural Network), while extracting only the features in that state, LSTMs also remember previous entries [7,8]. It is a method that makes decisions according to this situation and is therefore more applicable in text processing [9,10]. In this study, LSTM and CNN models have been applied for the extraction of text

features. After the LSTM phase, inputs are provided to the CNN layer for convolution. Thus the emotion classification model has been established with vectorial (Word Frequency Index) and global vector word embedding methods using LSTM. Three different datasets have been used, consisting of movie, restaurant and e-commerce order comments consisting of different categories. Each dataset has 1000 comments. The performance of embedding methods on classification has been evaluated by using three datasets.

2.RELATED WORKS

Extraction of information from text has been an area that has been frequently studied recently. Abel et al. discussed the similarity concept underlying these techniques based on word similarity in word embedding techniques such as Word2Vec or Glove, which are calculated using big data corpora, as a research topic. In a similar study, Mattyws et al. have presented a patent categorization method based on word embedding and LSTM (Long Short-Term Memory) in the process of classifying patents and subgroups of patents [11]. Ashik et al. have discussed the impact results of pre-trained word embedding models in languages other than English [12].

Rafat et al. have discussed skip-gram, continuous bag-of-words (CBOW) and Glove model on 105000 Bengali articles. The effects of embedding models in different languages have been observed [13]. Wang et al. have created a new LSTM model. They have designed a model on Chinese words using the classical frequency-based word embedding technique [14].

3. MATERIAL AND METHOD

3.1. Methodology

In the study, three different experiments have designed. As shown in Figure 1, Glove and Frequency Index techniques, which are the vectors used for the representation of words, have been classified separately with three different datasets and their effect on the classification performance has been observed. In the following chapters, word embedding techniques and LSTM and CNN layers used in our model are explained.

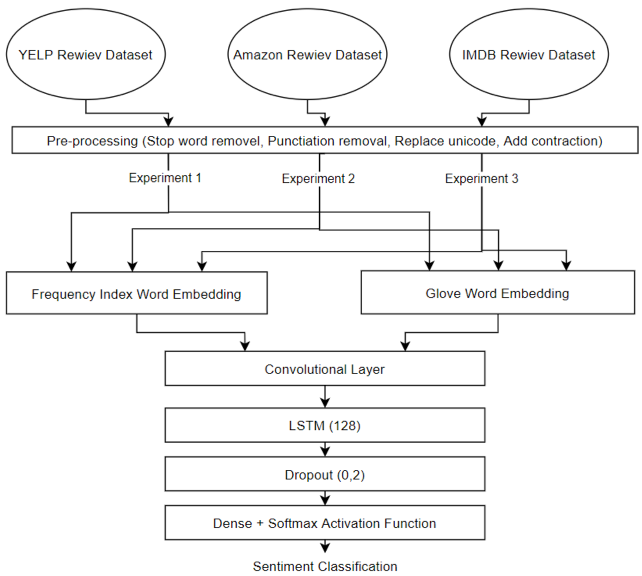
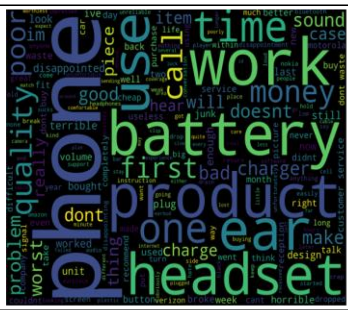
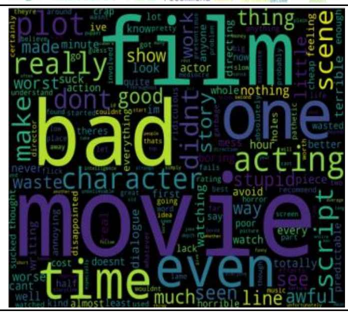


Figure 1 Experiment Model

3.1.1. Data Collecting and Pre-processing

Three different datasets are used for the model. The data have been provided over the data provided in the UCI Machine Learning Repository [15]. The data has been collected from the customer comments made by users on the websites amazon.com, imdb.com and yelp.com, respectively. It consists of positive and negative comments made by users in three different categories as shopping, movie reviews and restaurant reviews. Each dataset consists of 1000 comments and consists of 500 positive and 500 negative comment sentences. Stopwords have removed for 3 different datasets previously, tagged to sentiment labels and word clouds in Table 1 have been obtained by considering the word frequency conditions according to sentiment labels.Pre-processing of the data is a vital step that affects model performance. In the preprocessing process, capital and small letters have been checked respectively and all have been converted to lowercase letters. Punctuations have been deleted and stop words that frequently used in sentences have been removed. Thus, an optimized data set process has been prepared for the model.

Table 1 Wordcloud by sentiment label

Yelp Dataset		Amazon Dataset		IMDB Dataset
				
				

3.2. Word Embedding

With the word embedding step, words are converted to numbers expressed in vectors. In an artificial neural network, the mapping process in the hidden layer is performed for the words and the number representing. With embedding, dimension reduction is performed by using vectors expressing words and at the same time, words with the same meaning similarity are represented by close vectors [16].

3.2.1. Glove (Global Vector) Word Embedding

It is a model proposed by Stanford researchers [5]. According to this model, the vector representation of words is based on the similarity of words. This model has developed based on Skip-Gram and CBOW (Continuous Bag of Words) models. Using these two model approaches, it can be calculated more quickly and is aimed to give accurate results. Unlike the Skip-Gram and CBOW models, it also calculates the usage statistics of words. In this way, semantic relationships are also calculated. A new cost function is calculated for this problem solution [5].

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T w_j + b_i + b_j - \log X_{ij})^2 \quad (1)$$

'Xij' in Formula 1 is the number of occurrences of two words in the corpus. 'V' means our corpus. The weight function 'f (Xij)' must keep the following conditions.

1. When f (0) = 0, terms should not converge to infinity.
2. It should not decrease the weight function F (x) while giving low weight for word pairs that are less common together.
3. For large values of Xij, which is the number of occurrences of two words in the dictionary, the weight function F (x) must be smaller [5].

'X' is expressed word to word co-occurence count matrix. 'Xij' stores the number of occurrences of word j in the context of word 'i'. 'bi' is the bias value added for 'i', symmetrically this case is added for 'j' as 'bj'. It is known that how to calculate according to each word and gradient descent can be applied to minimize the cost function.

3.2.2. Frequency Based Word Embedding

If we consider Corpus (word dictionary) C and each sentence as expression {d1, d2…, dn}, we get N unique words (tokens) extracted from each document. N tokens refer to our dictionary. The dimension of M, which we accept as the Count Vector matrix, appears as DXN. Each row in the M matrix in Figure 2 shows the number of repetitions of each document (Di).

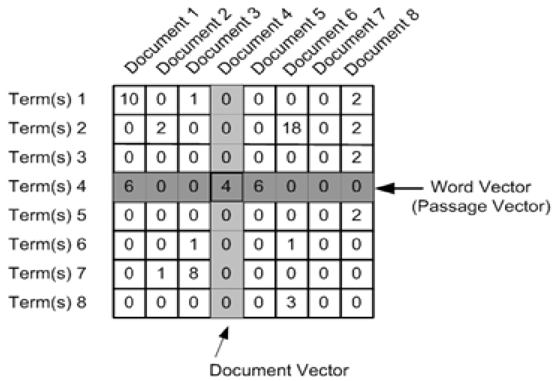


Figure 2 Word vector-matrix reference [17].

3.3. Experiments and Analysis

In this study, the model using three independent datasets, which are also specified in the flow diagram as Figure 1, has been implemented using two different word embedding techniques. All models have word embedding, text feature extraction and emotion extraction main components. In order to build a well organized classification by using LSTM based model, Python 3.8 and Keras deep learning library have been used. Unlike traditional neural networks, LSTM architecture can relate previous data to the existing one. It is a type of RNN (Recurrent Neural Network**)** that can learn long-term addictions and is widely used [18-19]. It has a repetitive neural network loop and each loop allows information to go back into the next neural network.

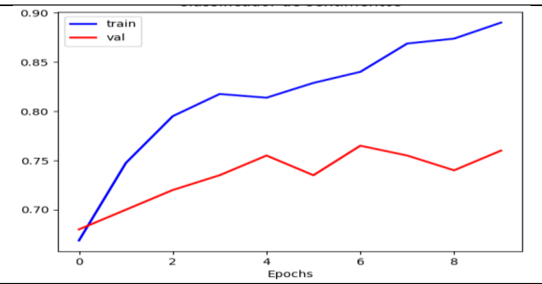
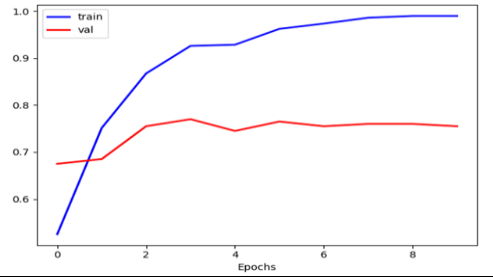
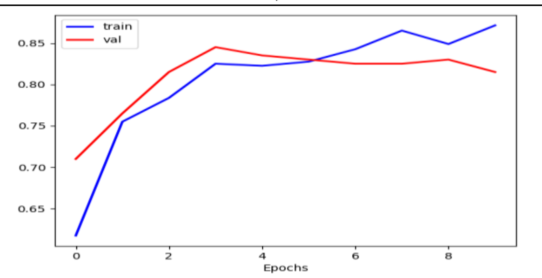
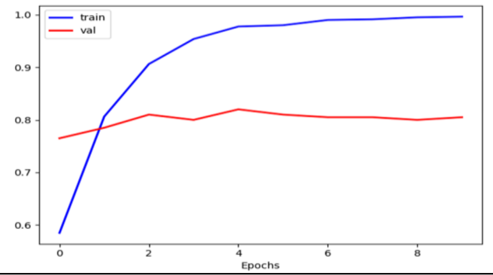
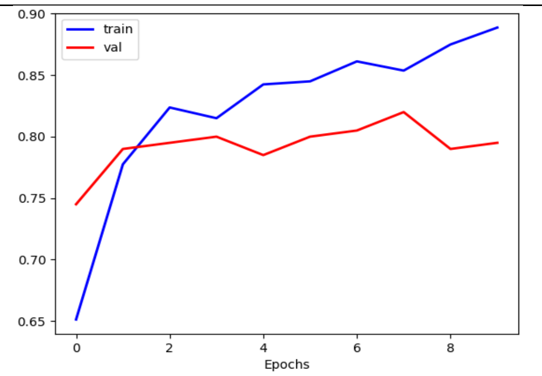
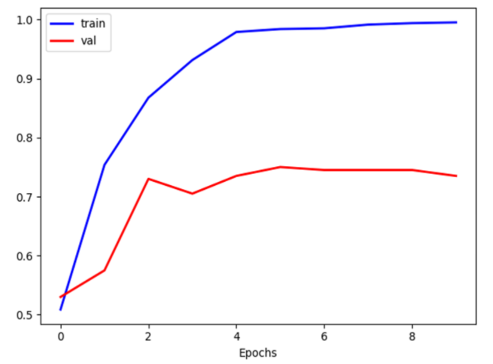
Table 2 Number of parameters in a model

Word Embedding Method	Glove Word Embedding	
Dataset	Layer (type)	Parameter Count
Yelp	Embedding	196700
	LSTM	80400
	Dense	101
	Total params	277,201
IMDB	Embedding	305300
	LSTM	80400
	Dense	101
	Total params	385,801
Amazon	Embedding	179900
	LSTM	80400
	Dense	101
	Total params	260,401
Word Embedding Method	Frequency Based Word Embedding	
Dataset	Layer (type)	Param

Yelp	Embedding LSTM Dense Total params	2000000 80400 101 2,080,501
IMDB	Embedding LSTM Dense Total params	2000000 80400 101 2,080,501
Amazon	Embedding LSTM Dense Total params	2000000 80400 101 2,080,501

Our model has subjected to a classification process for 10 epochs on 3 different datasets with 1000 pre-determined sentiment classes. Sentiment classification model, %80-%20 test and train set separation process has been performed. In three independent experiments with the Grove word embedding technique, higher accuracy results have obtained compared to the frequency index based word technique. The number of parameters in each model is shown in Table 2. The results showing the effect of the proposed models on accuracy are shown in Table 3. Glove word embedding technique has been observed accuracy results compared to frequency index based word embedding technique in 3 independent experiments. In addition, overfitting problems have been observed at early epoch levels when using frequency index based models. The main reason for this is that in the Gloveword embedding technique, vector representations that take into account more semantic associations are obtained by making use of external and larger datasets in the vector representation of words. On the other hand, the repetition frequency of the words on the dataset studied in frequency based word embedding and the tendency to memorize is higher because vector representations are formed over a narrow dataset.

Table 3 Accuracy Epoch Graphics

Dataset	Accuracy Epoch Result Graphic Using Glove	Accuracy Epoch Result Graphic Using Frequency
Yelp		
Amazon		
IMDB		

The vector representation of words in Natural Language Processing is an important factor in successful sentiment classification. In our study, the classification success of 2 different word embedding techniques has been observed with the LSTM convulutional model. In all three classification models, a maximum performance difference of 8% has been observed in classification with the proposed model according to the Glove word embedding and Frequency Based Word Embedding technique.

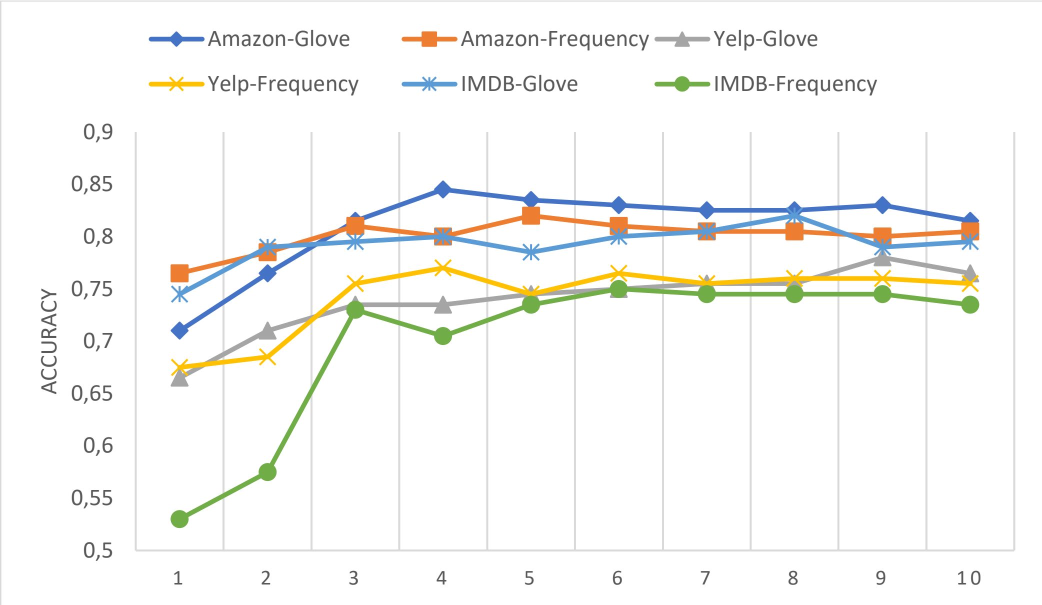


Figure 3 Comparative results by embedding model

In addition, when the Frequency model results have been observed, there is a high tendency to overfitting in the early steps, depending on the dataset and embedding technique. Although close classification rates are observed in Amazon and Yelp datasets as seen in Figure 3, a difference of 8% has observed for imdb dataset and also classification outcomes are as in Table 4.

Table 4 Precision, Recall and F-Score Rates

	Precision	Recall	F-Score
Amazon-Glove	0.8172	0.7917	0.8042
Amazon-Frequency	0.7849	0.7935	0.7892
Yelp-Glove	0.7292	0.7609	0.7447
Yelp-Frequency	0.7813	0.7282	0.7538
Imdb-Glove	0.8681	0.7315	0.794
Imdb-Frequency	0.6703	0.7262	0.6971

4. CONCLUSION

The size and diversity of the corpus with the proposed model affect the classification performance and the pre-trained vectorized techniques that increase the success in classification. For this purpose, the effect of pre-trained and model-dependent train word embedding methods on sentiment analysis has been observed in this study.

Despite the different datasets with high word density, better results have been obtained with the Glove word embedding technique, which emphasises word similarity. Besides, the model can be trained faster and scalable for the huge corpora. Since it contains a pre-trained vector matrix, it creates a disadvantage with high memory consumption. We are trying to expand our findings in our future work further. In particular, we want to examine the results with different predictive models such as "fastText" and "word2vec".

Funding

The author has no received any financial support for the research, authorship or publication of this study.

Contribution of the Authors

The authors have contributed equally to the work.

Ethics Committee Approval Notice

This study does not require ethics committee approval or any special permission.

Research and Publication Ethics Statement

In the writing process of this study, international scientific, ethical and citation rules have been followed and no falsification has been made on the collected data. Sakarya University Journal of Science and its editorial board have no responsibility for all ethical violations. All responsibility belongs to the responsible author and this study has not been evaluated in any academic publication environment other than Sakarya University Journal of Science.

REFERENCES

[1] H. Imaduddin, Widyawan and S. Fauziati, "Word Embedding Comparison for Indonesian Language Sentiment Analysis," 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, 2019, pp. 426-430, doi: 10.1109/ICAIIT.2019.8834536.

[2] Eligüzel, N., Çetinkaya, C., & Dereli, T. (2020). Comparison of different machine learning techniques on location extraction by utilizing geo-tagged tweets: A case study. Advanced Engineering Informatics, 46, 101151. https://doi.org/10.1016/j.aei.2020.101151

[3] Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., & Hassanien, A. E. (2020). Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. Applied Soft Computing, 97, 106754. https://doi.org/10.1016/j.asoc.2020.106754

[4] Ayata, Deger & Saraclar, Murat & Ozgur, Arzucan. (2017). Turkish tweet sentiment analysis with word embedding and machine learning. 1-4. 10.1109/SIU.2017.7960195.

[5] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[6] Y. Sharma, G. Agrawal, P. Jain and T. Kumar, "Vector representation of words for sentiment analysis using GloVe," 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, 2017, pp. 279-284, doi: 10.1109/INTELCCT.2017.8324059.

[7] R. Ni and H. Cao, "Sentiment Analysis based on GloVe and LSTM-GRU," 2020 39th Chinese Control Conference (CCC), Shenyang, China, 2020, pp. 7492-7497, doi: 10.23919/CCC50068.2020.9188578.

[8] Li, Dan & Qian, Jiang. (2016). Text sentiment analysis based on long short-term memory. 471-475. 10.1109/CCI.2016.7778967.

[9] Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., Kim, K. H., & Kwak, K.-S. (2019). Transportation sentiment analysis using word embedding and

ontology-based topic modeling. Knowledge-Based Systems, 174, 27–42. https://doi.org/10.1016/j.knosys.2019.02.033

[10] Q. Wang, L. Sun and Z. Chen, "Sentiment Analysis of Reviews Based on Deep Learning Model," 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 2019, pp. 258-261, doi: 10.1109/ICIS46139.2019.8940267.

[11] M. F. Grawe, C. A. Martins and A. G. Bonfante, "Automated Patent Classification Using Word Embedding," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, 2017, pp. 408-411, doi: 10.1109/ICMLA.2017.0-127.

[12] A. A. A. Rafat, M. Salehin, F. R. Khan, S. A. Hossain and S. Abujar, "Vector Representation of Bengali Word Using Various Word Embedding Model," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 2019, pp. 27-30, doi: 10.1109/SMART46866.2019.9117386.

[13] A. A. A. Rafat, M. Salehin, F. R. Khan, S. A. Hossain and S. Abujar, "Vector Representation of Bengali Word Using Various Word Embedding Model," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 2019, pp. 27-30, doi: 10.1109/SMART46866.2019.9117386.

[14] J. Wang and Z. Cao, "Chinese text sentiment analysis using LSTM network based on L2 and Nadam," *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, Chengdu, China, 2017, pp. 1891-1895, doi: 10.1109/ICCT.2017.8359958.

[15] Dimitrios Kotzias, Misha Denil, Nando de Freitas, and Padhraic Smyth. 2015. From Group to Individual Labels Using Deep Features. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). Association for Computing Machinery, New York, NY, USA, 597–606. DOI:https://doi.org/10.1145/2783258.2783380

[16] Bhoir, Snehal & Ghorpade, Tushar & Mane, Vanita. (2017). Comparative analysis of different word embedding models. 1-4. 10.1109/ICAC3.2017.8318770.

[17] N. (2020, October 19). An Intuitive Understanding of Word Embeddings: From Count Vectors to Word2Vec. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

[18] K. Zhang, W. Song, L. Liu, X. Zhao and C. Du, "Bidirectional Long Short-Term Memory for Sentiment Analysis of Chinese Product Reviews," 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 2019, pp. 1-4, doi: 10.1109/ICEIEC.2019.8784560.

[19] Chen, Nan & Wang, Peikang. (2018). Advanced Combined LSTM-CNN Model for Twitter Sentiment Analysis. 684-687. 10.1109/CCIS.2018.8691381.