'National Conference on Advancements in Computer Science and Engineering'
International Journal of Scientific Research in Computer Science,
Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

Semantic Vector Space Model for Text Classification using Progressive Learning Network Algorithm

C. Shanmuganathan, Naman Lahoti, Gairik Chakraborty, Vansh Singla
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India

ABSTRACT
Digital information is increasing day by day. Many documents require new methods and improved functionality for classifying and organizing documents. A simple method for classifying large documents is to use a hierarchical map that divides documents into hierarchical structures. Many methods have been proposed to solve the text classification problem, but most of the research has been done for English. This research presents a new classification method based on neural networks, optimized to solve existing classification problems. Both sentence level and lexical level states are integrated with layers to store deep-level information for language modelling. Dynamic routing is also optimized using context-level information from the sentence level state. Most of our models can circumvent the optimization of the node's properties and get the width of the receiving area. We also include language models to be embedded in graphs to improve extraction of non-graphical semantic information.
Keywords-digital text, neural network, text classification, graph node embedding.

I. INTRODUCTION

Electronic information retrieval systems are ubiquitous today, from instant mobile messaging applications to automated archives holding millions of documents. It creates many difficulties due to the large amount of data. However, one attempt is to identify some text so that users can more easily interpret it and convert data into patterns and knowledge creation. Organizing large amounts of electronic data in groups is of interest to many individuals and companies. The only solution to this problem is to split the text. Text classification is the process of classifying data. It uses many specializations, including machine learning, NLP, and artificial intelligence. It uses supervised learning as our method of training the model by feeding it a lot of data. Various algorithms have been proposed recently, such as SVM k-Nearest Neighbor Naive Bayes. The results show how well this method performs when it comes to classifying traditional texts. It has many applications such as text classification, modelling, sentiment analysis, curiosity and spam detection. Usually, there are only a few categories of text classification. When the classification task has many groups, we will continue with the problem of classification of the text, because they have some difficulties that require a special solution. There are many important problems in classification, which have similar characteristics in hierarchies. So hierarchical classification comes into play here. In hierarchical classification, classes are organized into categories and

subcategories, or we can say that classes are organized into class hierarchies, and when we apply this to data, it becomes the classification of the text. Hierarchical classification seems to be an important research topic in the database field in recent years. Sites like Yahoo and Wikipedia are examples of archives. There are now many applications where data is organized in hierarchies. To give a real-life example, Hierarchical text classification is similar to an accountant putting certain information into certain items.Many organizations such as IT companies, law firms, and healthcare companies benefit from data sharing. Thus, it shows that hierarchical classification is relevant to many applications and organizations.

Text classification can be done with scripting rules or supervised machine learning. The first is very easy due to changes in data or conditions, the second learns to recognize the map from the ideas to adjust the output. According to machine learning, text is often represented using a word bag model, from which features are extracted and fed to the classifier.Analyzing and analyzing this effective data helps in planning recovery operations. However, most of this user-generated content, such as tweets, is written in the local language of the disaster area. On the other hand, most of the information is focused solely on the English language. The ability to understand and interpret information written in multiple languages is critical to disaster recovery. Multilingual systems that can speak multiple languages will expand the use of these systems in rescue and disasters. There is a great need for multilingual distribution of documents that can analyze and distribute information received in the aftermath of a natural disaster. Another challenge in developing such a mechanism is the lack of sufficient registry information for damage reduction. Labeling is expensive and may not be effective during a disaster. Natural Language Processing (NLP) involves the mathematical processing and understanding of normal/human language. It has many tasks that require various measurements and data driven by mathematical methods. The most crucial job in NLP is to write a group. The main purpose of this section is distribution. The text can be a question, a sentence, or a paragraph. Text segmentation, real-world sentiment analysis, news segmentation, target segmentation, spam detection, etc. It has many applications such as Data files can be physically written to create text files. However, due to the growth of business and information on the Internet, automatic distribution of documents has become important. Three broad categories of automated text classification techniques are: rule-based, data-driven, and hybrid approaches. Using a system, a rule-based system divides data into groups. However, the registration must be completed. Alternatively, machine learning-based techniques have been developed in past few years.All the machine learning methods are done in different phases: they remove some structure from the text. Then key elements are rendered in an AI model. To classify fine points, word bag n-grams and extensions based on model time recursion and return data (TF-IDF) are effectively used. In the next step, As needed by algorithms like Naive Bayes, we combine a number of machine learning algorithms, including Support Vector Machines (SVM), Decision Trees (DT), and other techniques.

II. RELATED WORKS

J. T. Pintas et al. (2021) [1] suggests a categorization scheme for text classification methods (FS). In order to aid in the design of future experiments, it also offers a mapping of experiment settings. Trends, gaps, and the creation of a free platform for comparison and testing FS approaches for text categorization are among the primary results of the SLR.S. A. G. Shirazi and M. B. Menhaj (2006) [2] provides an innovative genetic algorithm-based method. The suggested method's effectiveness is assessed by resolving three channel

assignment issues. In compared to the other algorithms examined in the research, the results demonstrate that this technique can locate solutions with the lowest bandwidth need.M. Sebban and R. Nock (2002) [3] proposes an information theory- and statistical-based feature selection approach. The tool is effective in treating unnecessary and redundant features, even in huge feature areas, according to results. The geometrical structure to use on the training set and the optimisation criterion were both optimised. The 1-NN network offers a structure like the MST, which lessens the algorithm's complexity. Deng, Y. Li et al. (2019) [4] explains four types of cutting-edge feature selection techniques for text categorization are filter model, wrapper model, embedding model, and hybrid model. Although there are few wrapper and embedding approaches for text categorization, filter models are the most effective. To address this issue, researchers have suggested hybrid models that use filter techniques to remove unnecessary and irrelevant characteristics before feeding the chosen features to wrapper methods for further improvement. Khurana et al. (2018) [6] has been demonstrated that feature weighing can improve classification accuracy by combining feature selection with feature weighting. Performance will be enhanced with k-NN parameter adjustment. To determine the optimum average value of performance metrics, the k-NN algorithm's parameters must be tuned. The efficiency of the method takes accuracy, precision, recall, and F-measure.In contrast to employing merely feature selection using a metaheuristic approach, experimental data demonstrate that feature weighting with feature selection is a more successful method of classification. Asif et al. (2021) [7] proposes an FS approach based on SIW-APSO was suggested in this study. It starts by searching the solution space. The ideal solution is immediately determined using SIW-APSO. Typically, it takes tens of iterations to get the ideal answer. On the Reuters-21578 data set, the experimental findings demonstrate that the suggested approach beats all of its rivals by reaching 98.60% accuracy, 96.56% recall, and 97.57% F1 score.K. S. Kyaw and S. Limsiroratana (2019) [8] demonstrates that, depending on the characteristics of the training and testing datasets, swarm intelligence-based feature selection may be employed adaptatively and can provide better performance than conventional one. The computation time increases according to the number of iterations and population size, despite the fact that the testing dataset for this experiment is not very huge. To achieve the best features, swarm algorithms like NP, NI, accelerate type, and others must have a fitness function (accuracy or CCI) and adaptation parameters. K. S. Kyaw and S. Limsiroratana (2020) [9] proposed a system by drastically lowering the number of chosen features, our suggested approach offers good optimisation outcomes for the categorization of many classes of documents using a search strategy based on artificial intelligence. In terms of RMSE and BCT, the assessment process produces the best accuracy results. It may be used to categorise text documents for BBC news. However, if the behaviour of the data changes, the document categorization processing system must constantly be adjusted in volume, complexity, dimensionality, and other aspects.M. Mojaveriyan et al. (2016) [10] proposes a two-stage feature selection approach. The outcomes demonstrated that the suggested feature selection technique improves text classification effectiveness. The main advantage of the provided method is the combination of filter and wrapper approaches. The texts are classified using the K-nearest Neighbour classifier. Results indicated that, in comparison to the examined approaches, the suggested strategy can improve document classification efficiency.

Drawbacks-

- Learning to represent words contextually, ignoring their polarization, which can be problematic when used for text classification.

- It can only learn one vector for a word but cannot solve the problem that a word can mean different things in different contexts. Multiple layers increase computational cost and processing time.
- Decrease the generalizability of the model.
- Unpredictability in the training is higher.
- Inability to handle outliers completely.

III. TEXT CLASSIFICATION

The textual data can be labelled physically to perform text categorization. The three major groups of automated text categorization techniques are rule-based, data-driven, and hybrid approaches. Using a set of criteria, the rule-based technique divides textual data into several groups. Rule based methods use techniques to classify data into different categories. However, it requires complete domain knowledge. As an alternative, techniques based on machine learning have been increasingly successful in recent years. Each machine learning method involves two steps of operation: first, they extract certain manually created characteristics within the data. These features are then fed into the machine learning model. Temporal frequency and frequency conversion data (TF-IDF) based models and their modifications are often used to extract models. We employ a number of machine learning methods, including the support vector machine (SVM), decision tree (DT), and authority, in the next stage. Make use of probabilistic techniques like Naive Bayes and other ensemble approaches.

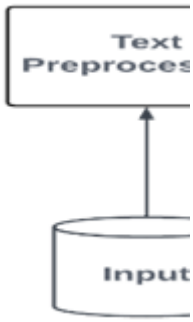

Figure 1. Text Classification

The method for classifying texts is -

A. Text Preprocessing

The preprocessing stage includes the filtering process. It shows that we've gotten rid of less useful parts of the text, such as punctuation. .!; symbols etc is removed as it reduces the overall accuracy of group operation. Removing them leads to better results than the algorithms used. The Natural Language Toolkit (NLTK) is used to complete this process.The word count is counted after the deletion of the text is complete. The selection of variables or the determination of features to create a good model is called feature selection. The purpose of this process is to obtain more accurate results.

B. Vectorizing Textual Inputs

NLP uses both character processing and word processing as its primary methods for typing. We initially concentrate on the language's brain architecture in our study, which divides letters into words. The text must first be converted into word vectors for each word before being fed into the model. Word embeddings change

every word into a distributed low-dimensional representation as opposed to a one-dimensional encoding. Based on the presumption that words with similar contexts have similar meanings, word embeddings produce word vectors that can display word similarity. Sequences need to be represented by special number vectors before text can be processed with neural networks. One way to achieve this is to use an embedding layer to learn a map from a vector of input sequences representing the search. However, multilayer training for different users in FL leads to better model performance after model aggregation in PS. Also, although large language constructs can be modified enough to retain their representation after compilation, communication through these large constructs involves a lot of exchange.

C. Text Classification

This particular network is a recurrent neural network. The network is an attempt to represent time or behaviour based on behaviour and is a neural network. The structure of the algorithm is designed to be a structure of linked data such as text. The planning process's neuron will update the state that governs its output using the input from today and the

state saved in the past. A network is a neural network that replaces our neural network layers with cell blocks. The input gate, forget gate, and output gate are a few of the parts that make up these cells. This memory gives the network the ability to acquires long-term reliance sequentially, enabling it to consider the complete context when generating a prediction. Neural Network acquires long-term reliance simply to create channels where the gradient flows for long intervals, which entails selectively remembering some information and transmitting it to the following state. A neural network unit is made up of a cell, an output gate, an input gate, and a forget gate. Neural network models have gates that provide the ability to add and remove information from the cell's state. Neural networks use a locking mechanism to control what data should be stored at a time, for how long, and when it can be read from the memory cell. A neural network has a module with memory that can learn data in real time. It is widely used in many places due to its performance in recording time. A memory module in a neural network has three multiplier units: an input port, a memory port, and an output port. These gateways control the flow of data in and out, so the network has some kind of memory. On the other hand, the network can learn a lot from these gateways.

IV. PROPOSED SYSTEM

First, we clean the main content, then we create a tree representation based on the relationship between users, extract the content in the original data from the user data and display properties, and then combine the content of the extracted content from the original data. Therefore, we get a high-dimensional feature vector. We use feature selection to avoid overfitting high-dimensional feature vectors during training and to determine the ideal small group of attributes to exclude. The algorithm's feature optimisation method not only solves the local optima problem but also accelerates the algorithm's convergence. In addition to the semantic features, the weights of the other three types of statistical results extracted from the data are too different to be used for direct classification, so the data must be made this specific first.We use Normalization to normalize properties and the property value method takes into account negative sensitivity property values. Accuracy (Acc) and F-score were chosen as evaluation criteria to test the effectiveness.We test the performance of various methods in

two datasets on various kinds of classification functions. The results prove that the word embedding neural network classifiers often beats traditional classifiers using TF-IDF.

Benefits:

- Effectively create context-sensitive representations during custom operations. The Effectiveness for distributed optimization Trustworthy and reliable, which refers to obtain explain ability.
- To acquire a quick and precise result. Can effectively guide the label assignment and boost the label confidence.
- Excellent generalization capability, optimal solution, and discriminative power.
- Capable of further reducing the required level of human effort.

A. Proposed Algorithm

The Progressive Learning Network Algorithm is a deep learning method for continuous learning and includes: learning, progress and termination. In the list of job candidates, a job is selected for study, usually using the study method. The incremental method is used to increase the capacity of the model by adding new parameters based on what has not been studied in the previous tasks, while learning from the available information for the new task of the hand without being affected by the forgotten damage.

Benefits:

- The main advantage of PLNs over ANNs is that PLNs can model datasets (e.g. time series).
- PLN can process inputs of any length.
- Weights can be segmented.

B. Dataset Description

The E-Commerce Text Dataset is used for taxonomy-based e-commerce text dataset groups - Electronic Home Books and Clothing & Accessories and covers almost one percent of all e-commerce websites. The data is in a two-column.csv format - the class name and content data are represented using columns. Content information is products and descriptions taken from e-commerce sites.

TABLE 1. POSITIVE HEADLINES LENGTH DISTRIBUTION

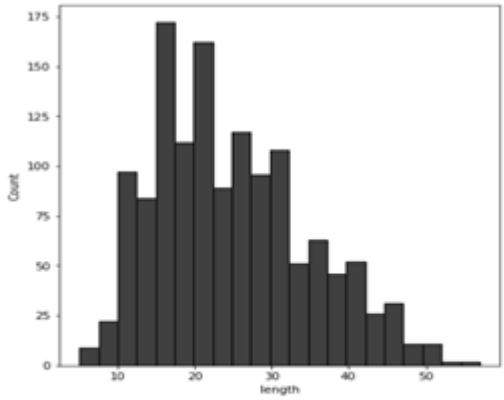

TABLE 2. NEGATIVE HEADLINES LENGTH

75%	30.0
Max	56.0

DISTRIBUTION

Count	1363.0
Mean	24.69
Std	10.06
Min	5.0
25%	17.0
50%	23.0
75%	31.0
Max	57.0

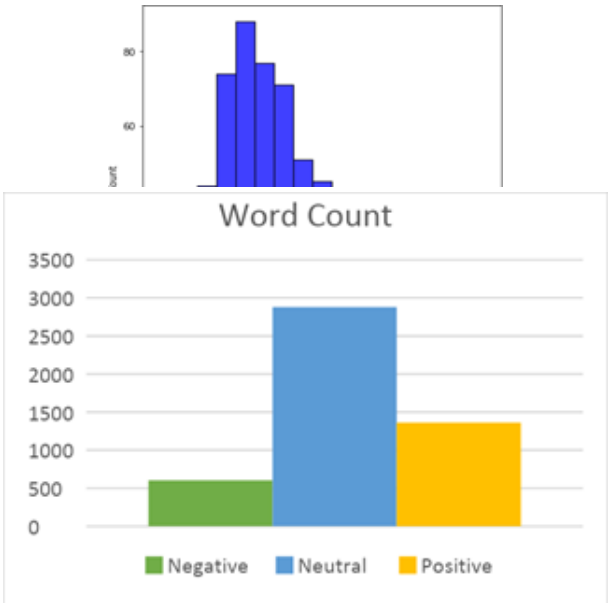

Figure 4. Statistical distribution of sentiment

C. System Architecture

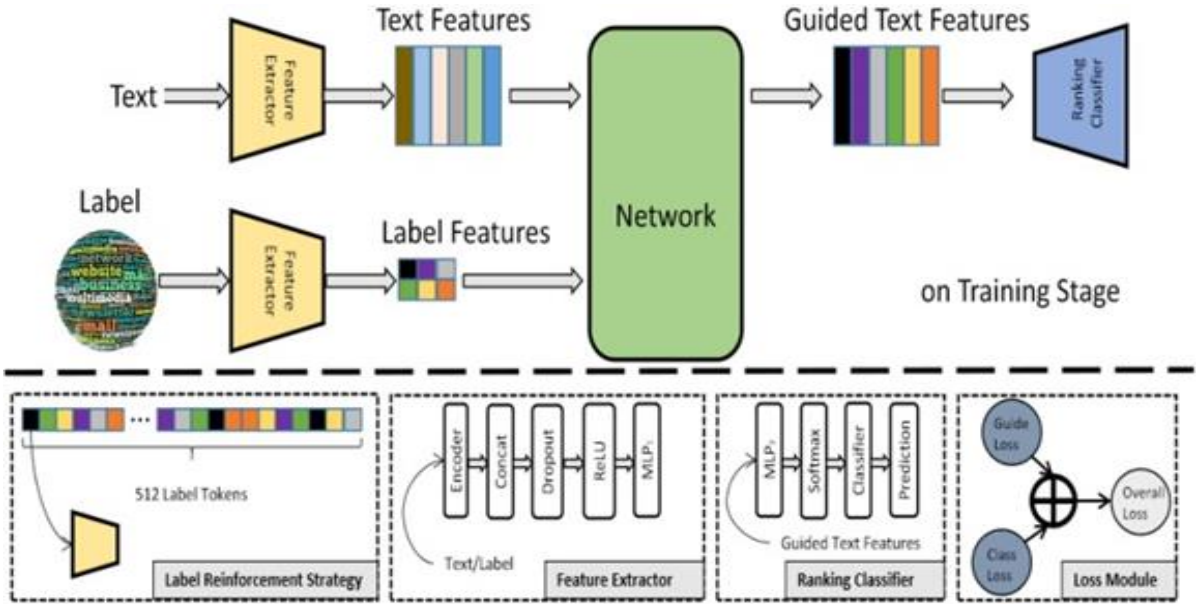

Figure 5. System Architecture

V. RESULT AND DISCUSSION

In addition to the semantic features, the other three types of statistical results extracted from the dataset are very different from the weight that should be used for direct classification, i.e., this feature file is primarily required. We use Normalization to normalize properties and the property value method considers negative sensitivity property values. Accuracy (Acc) and F-score were chosen as evaluation criteria to test the effectiveness. Evaluation of the performance of various approaches on a dataset for both multi-class classification tasks. The outcome indicates that word embedding neural network classifiers frequently beat the conventional classifier utilizing TF-IDF.

The Effectiveness for distributed optimization Trustworthy and reliable, which refers to obtain explain ability. It can create context-sensitive representations during custom tweaking. To acquire a quick and precise result. Can effectively guide the label assignment and boost the label confidence. It has excellent generalization capability, optimal solution, and discriminative power. It is Capable of further reducing the required level of human effort.

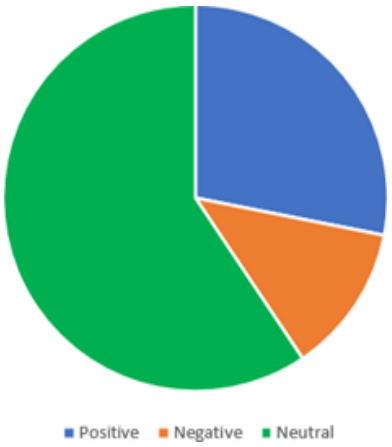

Figure 6. Pie Chart of different sentiments of headlines

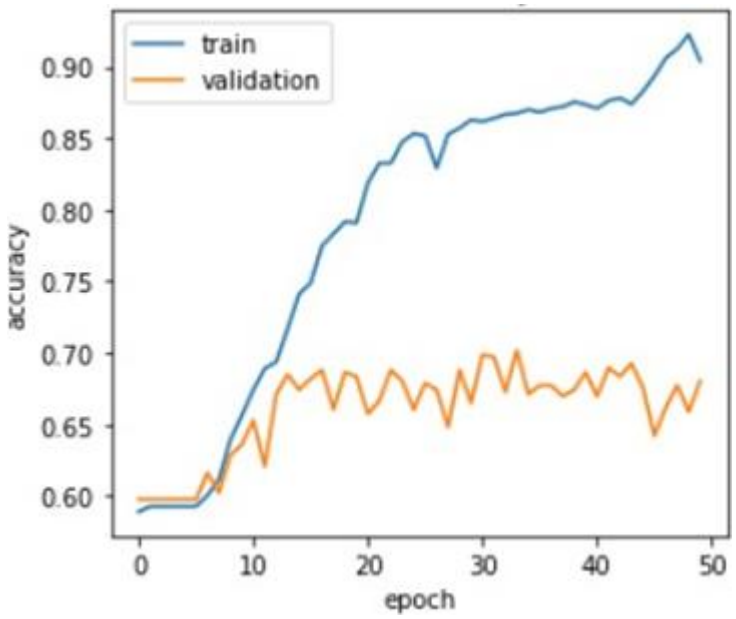

Figure 7.Accuracy Of Model

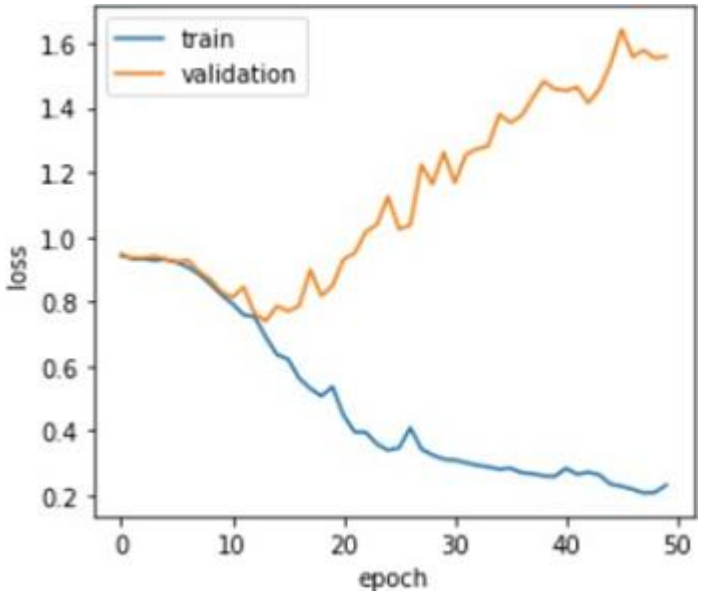

Figure 8.Loss Of Model

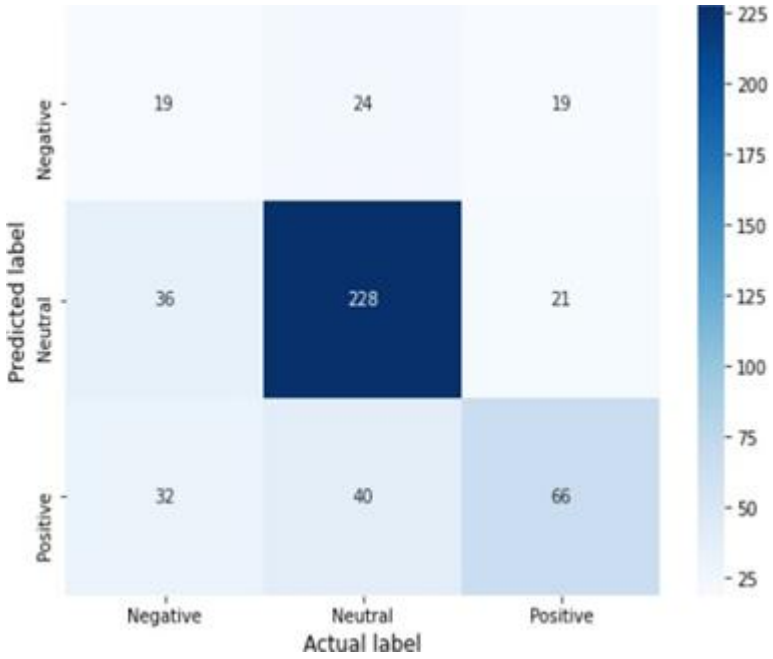

Figure 9.Confusion Matrix

TABLE 3. ACCURACY

Feature	Accuracy %			
	SVM	*NB*	*BLR*	*RFDT*
Word2vec	73.07	77.88	73.07	79.80
TF	83.65	79.80	78.84	81.73
TF-IDF	80.76	78.80	80.76	82.69

TABLE 4. WORD EMBEDDING VS TF-IDF

Feature	Precision	Recall	F1 score	Support
Word Embedding	0.956098	0.859649	0.905312	None
TF-IDF	0.985294	0.881579	0.930556	None

TABLE 5. MODEL SUMMARY

Layer(type)	Output Shape	Param #
Embedding (Embedding)	(None,60,32)	160000
Conv1d (Conv1D)	(None,60,32)	3104
Max_pooling1d (MaxPooling1D)	(None,30,32)	0
Bidirectional (Bidirectional)	(None,64)	16640
Dropout (Dropout)	(None,64)	0
Dense (Dense)	(None,3)	195
Total params: 179,939 Trainable params: 179,939 Non-trainable params: 0		

VI. CONCLUSION

In this paper we discussed the challenges and solutions associated with text classification task. While the task may be rather straightforward in some languages and preprocessing steps is a key to improving classification performance. We presented a word embedding model that captures semantic similarities between words at sub word level. Since the word embedding model uses sub words it can handle, and spelling errors caused by prior segmentation step. We proposed neural network models that utilize the word embedding model in text classification. The experimental results with multi-class classification datasets proved that the neural network models using the word embedding model consistently outperformed the baseline model using TF-IDF.

VII. REFERENCES

[1]. J. T. Pintas, L. A. Fernandes, and A. C. B. Garcia, “Feature selection methods for text classification: a systematic literature review,” Artificial Intelligence Review, Feb. 2021, doi: 10.1007/s10462-021-09970-6.

[2]. S. A. G. Shirazi and M. B. Menhaj, “A New Genetic Based Algorithm for Channel Assignment Problems,” Springer Berlin Heidelberg eBooks, Jan. 2006, doi: 10.1007/3-540-34783-6_10.

[3]. M. Sebban and R. Nock, “A hybrid filter/wrapper approach of feature selection using information theory,” Pattern Recognition, Apr. 2002, doi: 10.1016/s0031-3203(01)00084-x.

[4]. X. Deng, Y. Li, J. Weng, and J. Zhang, “Feature selection for text classification: A review,” Multimedia Tools and Applications, Feb. 2019, doi: 10.1007/s11042-018-6083-5.

[5]. B. Kitchenham, “Procedures for Performing Systematic Reviews,” Jan. 2004.

[6]. Khurana, Anshu and Verma, Om Prakash, "PSO based Optimal Text Classification using Tuned k-NN and Feature Weighting", International Journal of Information Systems & Management Science,2018, Vol. 1, No. 1, 2018, Available at SSRN: https://ssrn.com/abstract=3363570

[7]. M. Asif, A. A. Nagra, M. B. Ahmad, and K. Masood, "Feature Selection Empowered by Self-Inertia Weight Adaptive Particle Swarm Optimization for Text Classification," Applied Artificial Intelligence, Dec. 2021, doi: 10.1080/08839514.2021.2004345.

[8]. K. S. Kyaw and S. Limsiroratana, "Traditional and Swarm Intelligent Based Text Feature Selection for Document Classification," International Symposium on Communications and Information Technologies, Sep. 2019, doi: 10.1109/iscit.2019.8905200.

[9]. K. S. Kyaw and S. Limsiroratana, "An Optimization of Multi-Class Document Classification with Computational Search Policy," ECTI Transactions on Computer and Information Technology, Jun. 2020, doi: 10.37936/ecti-cit.2020142.227431.

[10]. M. Mojaveriyan, H. Ebrahimpour-Komleh, and S. J. Mousavirad, "IGICA: A Hybrid Feature Selection Approach in Text Categorization," International journal of intelligent systems and applications, Mar. 2016, doi: 10.5815/ijisa.2016.03.05.

[11]. F. Z. Kermani, E. Eslami, and F. Sadeghi, "Global Filter–Wrapper method based on class-dependent correlation for text classification," Engineering Applications of Artificial Intelligence, Oct. 2019, doi: 10.1016/j.engappai.2019.07.003.

[12]. B. Drury, L. Torgo, and J. Almeida, "Classifying news stories to estimate the direction of a stock market index," Iberian Conference on Information Systems and Technologies, Jun. 2011.

[13]. Pope, Mark W.,"Automatic Classification of Online News Headlines," 2007. https://doi.org/10.17615/arcn-py08.

[14]. Deshmukh, Dr. R. R. and Mr. D. K. Kirange. "Classifying News Headlines for Providing User Centered E-Newspaper Using SVM." (2013).

[15]. Porter, Martin F. "Snowball: A language for stemming algorithms." (2001).

[16]. Shanmuganathan, C., Boopalan, K., Elangovan, G., Sathish Kumar, P.J., Enabling security in MANETs using an efficient cluster based group key management with elliptical curve cryptography in consort with sail fish optimization algorithm, Transactions on Emerging Telecommunications Technologies, Vol.34, Issue 3, March 2023.

[17]. M. Shobana, C. Shanmuganathan, Nagendra Panini Challa, S. Ramya, An optimized hybrid deep neural network architecture for intrusion detection in real-time IoT networks, Transactions on Emerging Telecommunications Technologies, Vol.33, Issue 12, December 2022.