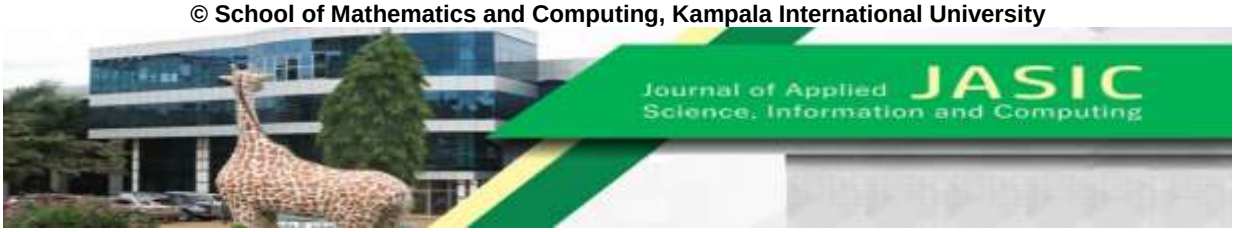
Journal of Applied Sciences, Information, and Computing
Vol. 1, No. 2 (December, 2020)



ISSN: 1813-3509 **https://jasic.kiu.ac.ug**

AN IMPROVED MULTI-LABELED LSTM TOXIC COMMENT CLASSIFICATION

Muhammad Abubakar* [1], Aminu Tukur[2] , Usman Bukar Usman

[1,2,3] Faculty of Computer Science and Information Technology,
Bayero University Kano, Nigeria
[1]Muhammadabubakar08155@gmail.com [2]Tukuraminu85@gmail.com

Abstract

The origin of text classification was far back to the early '60s. Text classification classified text into different predefined classifications. One of the techniques used for text classification long short term memory, which is an artificial recurrent neural network architecture. Today, all around the world people are expressing themselves with their opinions and also discuss among others via the media. In such a setup, it is quite observable that discussions may arise due to differences in opinion. These discussions might take a dirty side and which may further result in combats over the social media platforms and may lead to offensive language termed as toxic comments. To identify online hate speech, a large number of scientific studies have been devoted to using Natural Language Processing in combination with Machine Learning and Deep Learning methods. Among the challenges of toxic comment, classifiers are the Out-of-vocabulary words problem, which is the occurrence of words that are not present in the training data. Long-Range Dependencies are also a challenge to toxic comment classification. Which is a situation whereby the toxicity of comments often depends on expressions made in the early parts of the comment. This is especially problematic for longer comments. Another challenge is the low accuracy of comment classification techniques. Epoch was used in improving the accuracy of long short term memory. Epoch tends to improve the accuracy of the classifier since it positively affects the speed and quality of the learning process. We have an improvement of 0.4068 in precision, 0.2871 in a recall, 0.2293 in F1, and 0.4291 inaccuracy.

Keywords: Epoch, Long Short Term memory, Classifier, Machine Learning.

1. INTRODUCTION

The origin of text classification was far back to the early '60s, but machine learning techniques were effectively realistic in the '90s (Kajla, Hooda, & Saini, 2020). For over a decade, social media and social networking have been growing in geometric progression. Today, all around the world people are expressing themselves with their opinions and also discuss among others via the media. In such a setup, it is quite observable that discussions may arise due to differences in opinion. But often these discussions take a dirty side and may result in combats over the social media platforms, these might result in offensive language termed as toxic comments that may be used from one side (Chakrabarty, 2012). Machine learning has unwrapped numerous doors for researchers in text analysis. Text classification is one of them which means a task of classifying text into different predefined classifications (Kajla et al, 2020); (Mozafari, Farahbakhsh, & Crespi, 2019. LSTMs were introduced by (Schmidhuber, & Hochreiter, 1997) to alleviate the disappearing gradient problem (Guggilla, Miller, & Gurevych, 2016).

Generated hateful and toxic content by a portion of users in social media is a rising phenomenon that inspired researchers to devote considerable efforts to the challenging direction of hateful content identification. We not only need an effective automatic hate speech detection model based on advanced machine learning and natural language processing, but also an adequately large amount of annotated data to train a model (Mozafari, Farahbakhsh, & Crespi, 2019). Nowadays the Internet has become the leading platform to represent our skills. Several websites allow people to use their platform to display their skills through articles, videos, and other information in different formats. Most of the websites provide a facility for commenting on any uploaded information and there is the possibility that people can use abominable language in their comments (Kajla et al, 2020). Toxic comment classification has become a dynamic research field with

many recently proposed methods van (Aken, Risch, Krestel, & Löser, 2018). These toxic comments may be threatening, obscene, insulting, or identity-based hatred. Thus, these pose the threat of abuse and harassment online. Consequently, certain individuals stop giving their views or give up seeking different opinions which results in the unhealthy and unfair discussion. As a result, different platforms find it very difficult to facilitate fair conversation and are often forced to either limit user comments or get disbanded by shutting down user comments completely. The Conversation AI team, a study group founded by Jigsaw and Google has been working on techniques for providing a healthy setting for communication (Chakrabarty, 2012).

Sentiment classification techniques have been widely used for analyzing user opinions (van Aken, Risch, Krestel, & Löser, 2018). In conventional supervised learning techniques, several hand-crafted features are needed, which involves a thorough understanding of the domain. Subsequently, social media posts are usually short, there's a lack of features for effective classification. Thus, word embedding models can be used to learn different word usages in various contexts.

The identification and tagging of offensive content have been heavily explored with different classical Natural Language Processing (NLP) and Machine Learning techniques. However, considering the constraints associated with the natural language such as word spell variations and typos, contextual ambiguity, and semantic variations, the supervised machine learning technique turns out to be the most suitable technique for the task. The vast amount of data being generated from the digital platform, daily, usually aids in favor of the supervised machine learning technique. Though offensive text classification naturally tends to be in the arena of the traditional NLP method owing to the heavy reliance on the text, the traditional NLP approach fails to meet the enormous details associated with the texting/commenting styles of different users, has not proven to be scalable. It has been established that the LSTM (Long Short-Term Memory) and the GRU (Gated Recurrent Unit) have been the state-of-the-art sequence modeling neural networks (Schmidhuber, 2015); (Babu, & Murali, N.D).

Hyperparameter has often defined as a model hyperparameters, that cannot be inferred while fitting the machine to the training set because they denote a model selection of task or algorithm or algorithm hyperparameters that in principle have no impact on the performance of the model but affect the speed and feature of the learning process (Feurer, & Hutter, 2019). Multi-label classification (MLC) aims to assign multiple labels to each sample. It can be applied in many real-world scenarios, such as text categorization and information retrieval. Due to the complex dependency between labels, a key challenge for the MLC task is how to effectively capture high-order correlations between labels (Zhang et al, 2019). Presently, people use online social media platforms such as Twitter and Facebook to share their emotions and thoughts. Detecting and analyzing the emotions expressed in social media content benefits many applications in commerce, public health, social welfare, etc. (Jabreel &Moreno, 2019). A huge number of people are involved in online social networks. The wide-open gate, for unfettered use of offensive content in digital platforms, is quite obvious. There have been multiple efforts by various stakeholders to identify and classify such posts automatically employing different algorithms (Kajla et al, 2020); (Guggilla et al, 2016); (Mozafari et al, 2019). A large number of scientific studies have been dedicated to detecting online hate speech using NLP in combination with ML and Deep Learning (DL) methods. Among the challenges of toxic comment, classifiers are Out-of-vocabulary words, which is the occurrence of words that are not present in the training data (Arel, Rose, & Karnowski, 2010). These words which include slang or misspellings, but also intentionally obfuscated content is regularly a challenge to toxic comment classification. It is a condition where the toxicity of comments often depends on expressions made in the early parts of the comment. This is especially problematic for longer comments (>50 words) (Kajla et al, 2020) However, Haralabopoulos, Anagnostopoulos, and McAuley (2020) had worked on Ensemble Deep Learning for Multilabel Binary Classification of User-Generated Content, however, the accuracy of such system was not good enough. Using the epoch approach on Long short-term memory tends to improve the accuracy of the classifier since it positively affects the speed and quality of the learning process.

2. REVIEW OF RELATED WORKS

Haralabopoulos et al (2020) worked on Ensemble Deep Learning for Multilabel Binary Classification of comments. Ensemble learning combines the single-model outputs to improve predictions and generalization. The researchers noted that Ensemble learning improves upon three key aspects of learning, statistics, computation, and representation Ensemble methods reduce the risk of data miss representation, by combining multiple models. The researchers reduce the risk of employing a single model trained with biased data, while most learning algorithms search locally for solutions which in turn confines the optimal solution, ensemble methods can execute random seed searches with variable start points with less computational resources. Wikipedia dataset is used. The approach weighted ensemble outperformed the baseline stacked ensemble in 75% of cases by 1.5% to 5.4%.

The Empirical Evaluation of Temporal Convolutional Network for Offensive Text Classification (Babu, & Murali, N.D). The researchers evaluated the performance of TCN to identify and classify offensive language based on the intensity of its offensive content along with the conventional Convolutional Neural Network (CNN) and the state-of-the-art sequence modeling neural networks LSTM and GRU. Unlike LSTM and GRU, TCN exploits parallelism and can retain long-range history with dilated convolutions and residual blocks. TCN classifier was

assessed for hate speech, aggression, and harassment datasets. In three datasets, the TCN set new benchmark scores (weighted F1). CNN's were considered to extract features from video frames or regular images, while Recurrent Neural Nets (RNNs) excelled at sequence problems, which include text and speech problems, predominantly.

Major enhancements to vanilla RNN, including GRU and LSTM, boosted the long-range history contextualization but they did not turn out to be optimal, and effective in real-time. With TCN, long-range history could be exploited much better, in real-world scenarios, and the word-word correlation could also be captured effectively. From the evaluation results, it is evident that Temporal Convolution Neural Network outperforms conventional CNN LSTM and GRU models for offensive text classification. Furthermore, TCN performed better than RNN variants LSTM and GRU, concerning the macro averaged F1 scores of Toxic Comment Classification, in much lesser training time than the RNN variants: LSTM and GRU. GRU, considered to be the most effective model for use in short and medium sentences (<500 words), is trailing behind TCN and LSTM in the Toxic Comment Classification task. The researchers concluded that TCN equipped with Residual Blocks, Causal, and Dilated Convolution can model long-range sentences and can capture the long-range context much more efficiently, compared to all the other models, as observed in the study. Aken et al (2018) worked on Challenges for Toxic Comment Classification and compare different deep learning and shallow approaches on a new, large comment dataset and propose an ensemble that outperforms all individual models. Further, the researchers validate their findings on a second dataset. The results of the ensemble enable the researchers to perform an extensive error analysis, which reveals open challenges for state-of-the-art methods and directions towards pending future research. These challenges include missing paradigmatic context and inconsistent dataset labels.

The approaches make different errors and can be combined into an ensemble with improved F1- measure. The ensemble especially outperforms when there is high variance within the data and on classes with few examples. Some combinations such as shallow learners with deep neural networks are especially effective. Error analysis on results of the ensemble identified difficult subtasks of toxic comment classification. We find that a large source of errors is the lack of consistent quality of labels. Additionally, most of the unsolved challenges occur due to missing training data with highly idiosyncratic or rare vocabulary. Mozafari, Farahbakhsh, and Crespi (2019) worked on CNN- and LSTM-based Claim Classification in Online User Comments. The researchers described a supervised approach, based on deep neural networks, for classifying the claims made in online arguments and conduct experiments using convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) on two claim data sets compiled from online user comments. Using different types of distributional word embedding, but without incorporating any rich, expensive set of features, The researchers achieved a significant improvement over the state of the art for one data set (which categorizes arguments as factual vs. emotional), and performance comparable to the state of the art on the other data set (which categorizes propositions according to their verifiability). The approach has the advantages of using a generalized, simple, and effective methodology that works for claim categorization on different data sets and tasks.

Mozafari et al (2019) worked on the BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media and introduce a novel transfer learning approach based on an existing pre-trained language model called BERT (Bidirectional Encoder Representations from Transformers). More specifically, the ability of BERT at capturing hateful context within social media content was investigated by using new fine-tuning methods based on transfer learning. To evaluate the approach, the researchers use two publicly available datasets that have been annotated for racism, sexism, hate, or offensive content on Twitter. The results show that the approach obtains considerable performance on these datasets in terms of precision and recall in comparison to existing approaches. Consequently, the model can capture some biases in the data annotation and collection process and can potentially lead us to a more accurate model. For the implementation of the neural network, the PyTorchPyTorch-pretrained-bert library was used. It contains the pre-trained BERT model, text tokenizer, and pre-trained WordPiece. As the implementation environment, the researchers used Google Colaboratory tool which is a free research tool with a Tesla K80 GPU and 12G RAM. Based on the experiments, the Classifier was trained with a batch size of 32 for 3 epochs. The dropout probability is set to 0.1 for all layers. Conflating hatred content with offensive or harmless language causes online automatic hate speech detection tools to flag user-generated content incorrectly.

Liang et al (2019) worked on Evolutionary Neural AutoML for Deep Learning. It introduces an evolutionary AutoML framework called LEAF that not only optimizes hyperparameters but also network architectures and the size of the network. LEAF makes use of both state-of-the-art evolutionary algorithms (EAs) and distributed computing frameworks. Experimental results on medical image classification and natural language analysis show that the framework can be used to achieve state-of-the-art performance. In particular, LEAF demonstrates that architecture optimization provides a significant boost over hyperparameter optimization and that networks can be minimized at the same time with little drop in performance. LEAF, therefore, forms a foundation for democratizing and improving AI, as well as making AI practical in future applications.

Ibrahim, Torki, and ElNainay (2018) worked on Imbalanced Toxic Comments Classification Using Data

Augmentation and Deep Learning. In the paper, data from Wikipedia talk page edits were used to train a multi-label classifier that detects different types of toxicity in online user-generated content. The researchers present different data augmentation techniques to overcome the data imbalance problem in the Wikipedia dataset. The proposed solution by the researchers is an ensemble of three models: the convolutional neural network (CNN), bidirectional long short-term memory (LSTM), and bidirectional gated recurrent units (GRU). The classification problem was divided into two steps, to determine whether or not the input is toxic than to find the types of toxicity present in the toxic content. The evaluation results show that the proposed ensemble approach outperforms all the other considered algorithms. It achieves a 0.828 *F*1-score for toxic/nontoxic classification and 0.872 for toxicity types prediction.

Aken et al (2018) described the concept of toxicity and characterize its subclasses. Further, the researchers present various deep learning approaches, including datasets and architectures, tailored to sentiment analysis in online discussions. One way to make these approaches more comprehensible and trustworthy is fine-grained instead of binary comment classification. On the downside, more classes require more training data. Therefore, the researchers augment training data by using transfer learning and also discuss real-world applications, such as semi-automated comment moderation and troll detection. Jabreel and Moreno (2019) describe the development of a novel deep learning-based system that addresses the multiple emotion classification problems on Twitter. The researchers proposed a novel method to transform multiple emotions into a binary classification problem and exploit a deep learning approach to solve the transformed problem. The system outperforms the state-of-the-art systems, achieving an accuracy score of 0.59 on the challenging SemEval2018 Task 1:E-cmulti-label emotion classification problem. Tabassi et al (2019) worked on a Taxonomy and Terminology of 21 Adversarial Machine Learning. The document develops a taxonomy of concepts and defines terminology in the field of AML. The taxonomy, built on and integrating previous AML survey works, is arranged in a conceptual hierarchy that includes key types of attacks, defenses, and consequences. The terminology, arranged in an alphabetical glossary, defines key terms associated with the security of ML components of an AI system. Taken together, the terminology and taxonomy are intended to inform future standards and best practices for assessing and managing the security of ML components, by establishing a common language and understanding of the rapidly developing AML landscape.

Le, Ho, Lee, and Jung (2019) worked on an LSTM Approach to Short Text Sentiment Classification with Word Embeddings. To detect the sentiment polarity from short texts. The researchers explore deeper semantics of words using deep learning methods and investigate the effects of word embedding and long short-term memory (LSTM) for sentiment classification in social media. First, words in posts are converted into vectors using word embedding models. Then, the word sequence in sentences is input to LSTM to learn the long-distance contextual dependency among words. The experimental results showed that deep learning methods can effectively learn the word usage in the context of social media given enough training data. The quantity and quality of training data greatly affect the performance. The classification performance is better for movie reviews than casual comments and posts in online forums. But the performance of LSTM is still comparable to ELM and NB.
This shows the feasibility of an LSTM-based approach to short-text sentiment classification. Third, data size can also affect the classification performance. More training data can lead to better performance. Finally, special characteristics in certain online forums might lead to inferior classification performance. This behavior mismatch between user opinions in comments and user ratings reflects the sarcastic language used among the community in PTT online forum. The system furthermore can Forecast where the daily discharge and rainfall were used as input data. Moreover, characteristics of the datasets which may influence the model performance were also of interest. As a result, the Da River basin in Vietnam was chosen and two different combinations of input data sets from before 1985 (when the Hoa Binh dam was built) were used for one-day, two-day, and three-day flowrate forecasting ahead at Hoa Binh Station.

The predictive ability of the model is quite impressive. The Nash–Sutcliffe efficiency (NSE) reached 99%, 95%, and 87% corresponding to three forecasting cases, respectively. The findings of the study suggest a viable option for flood forecasting on the Da River in Vietnam, where the river basin stretches between many countries and downstream flows (Vietnam) may fluctuate suddenly due to flood discharge from upstream hydroelectric reservoirs. d'Sa, Illina, and Fohr (2020) worked on BERT and fastText Embeddings for Automatic Detection of Toxic Speech. The researchers proposed automatic classification of toxic speech using embedding representations of words and deep-learning techniques and also perform binary and multi-class classification using a Twitter corpus and study two approaches: (a) a method which consists extracting of word embeddings and then using a DNN classifier; (b) fine-tuning the pre-trained BERT model. We observed that BERT fine-tuning performed much better. The proposed methodology can be used for any other type of social media comments. The researchers observed that BERT fine-tuning performed much better than feature-based approaches on a Twitter corpus.

Table 1 RELATED WORK

SN	TITLE	REFERENCES	METHODOLOGY	WEAKNESS	STRENGTH
1.	Ensemble Deep Learning for Multilabel Binary Classification of comments.	(Haralabopoulos, Anagnostopoulos, & McAuley, 2020)	Long-short term memory	Less accuracy.	reduce the risk of employing a single model trained with biased data, While most learning Algorithms search locally for solutions which in turn confines the optimal solution, ensemble methods can execute random seed searches with variable start points with less computational resources.
2.	Empirical Evaluation of Temporal Convolutional Network for Offensive Text Classification.	(Sridharan, & Swapna, 2019)	Temporal Convolutional Network	Not effective in long sentences	TCN exploits parallelism and can retain long-range history with dilated convolutions and residual blocks.
3.	CNN- and LSTM-based Claim Classification in Online User Comments.	(Guggilla et al, 2016)	comparing the BiGRU RNN network with other neural networks such as CNN, LSTM, and Hybrid CNN+LSTM.	Not incorporating any rich, expensive set of features	convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) models.

3. METHODOLOGY

The proposed model is an epoch based LSTM, which is aimed towards it is aimed towards improving the accuracy of a multi-labeled toxic comment classification. The threshold value used is 0.4. To improve the multi-labeled toxic comment classifier, the system uses epoch approaches. The proposed system composes of the input and embedding layer, the inputs are sentences from the Wikipedia data set while the embedding layer provides a dense representation of words and their relative meanings. Pre-trained word embeddings are essentially word embeddings obtained by training a model unsupervised on a corpus. Unsupervised training in this case typically involves predicting a word based on one or more of these surrounding words. The word embedding's obtained from such a training process, along with the trained model can then be used, in many instances, for a supervised task like tagging tasks (NER, POS, etc) with labeled data this is often much lesser than what one might need to train a model from scratch - the pre-trained embedding's along with the model is fine-tuned for the specific task. We will also employ a Data augmentation strategy. Data augmentation strategy is a strategy that enables practitioners to significantly increase the diversity of data available for training models, without actually collecting new data. Data augmentation techniques such as cropping, padding, and horizontal flipping are commonly used to train large neural networks. However, most approaches used in training neural networks only use basic types of augmentation (Graves, & Schmidhuber, 2005); (Krenker, Bester, & Kos, 2011); (Qin, Pengda,Weiran Xu, & Jun Guo, 2016); (Goldberg, 2016). While neural network architectures have been investigated in-depth, less focus has been put into discovering strong types of data augmentation and data augmentation policies that capture data invariances. Setting 400 characters as the threshold included up to 80% of the data and hence appeared to be a good choice. "We had fewer words in total, but the percentage of toxic words captured were more". Like every other machine learning model, the system starts with an Input Layer that takes in the padded sequence of tokenized sentences, the layer then passes its outputs to the Embedding layer followed by the LSTM layer which then gets passed to a 1 dimensional Global Max Pooling layer which reduces the dimensionality of the data from 300 to 60 to further lessen the complexity, a dropout layer immediately follows to reduce overfitting.

The dropout layer is a regularization method where input and recurrent connections to LSTM units are probabilistically excluded from activation and weight updates while training a network. Then a dense layer is

accompanied by a Dropout Layer to further reduce the risks of overfitting. The Final Dense layer then squashes the input of 50 down to 6 which refers to the size of the classes. Since having just one Epoch leads to underfitting, we picked a value that will provide an optimum result. We found out that having an Epoch of ten, LeCun, Bengio, and Hinton (2015) provided a balanced result that was neither fitting nor Overfitting. Traditionally, the number of epochs is usually large, often hundreds or thousands. This allows the learning algorithm to run until the error from the model has been sufficiently minimized. However, it will take days to train a model with thousands of epochs on a traditional GPU. The batches are used to train LSTMs, and selecting the batch-size is a vital decision since it has a strong impact on the performance. For this model, the best results have been obtained with batch sizes of 16. However, it took twice the time needed to train with a batch size of 32, which was not far behind in performance.

In this work, epoch will be used on Long Short-Term Memory (LSTM) approach. The lost function reduces all the various good and bad aspects of a possibly complex system down to a single number, a scalar value, which allows candidate solutions to be ranked and compared. A threshold of 0.4 will be used. The activation layer at the final layer greatly determines the results of the model, so it is very crucial to select the right function that is best for the specific type of problem the model is trying to solve Since this is a multi-label problem containing six different forms of toxic language and a multi-class dataset containing six mutually exclusive classes of toxic language, we must choose an activation function that will allow us to model our results as a mutually exclusive probability distribution, sigmoid is the right choice. It is especially used for models where we have to predict the probability as an output.

3.1 MODEL ARCHITECTURE

In the proposed model architecture, we have five layers, which are the input layer, embedding layer, LSTM layer, Global Max Polling layer, and dropout layers. Below in figure 3, the system architecture of the system is presented.

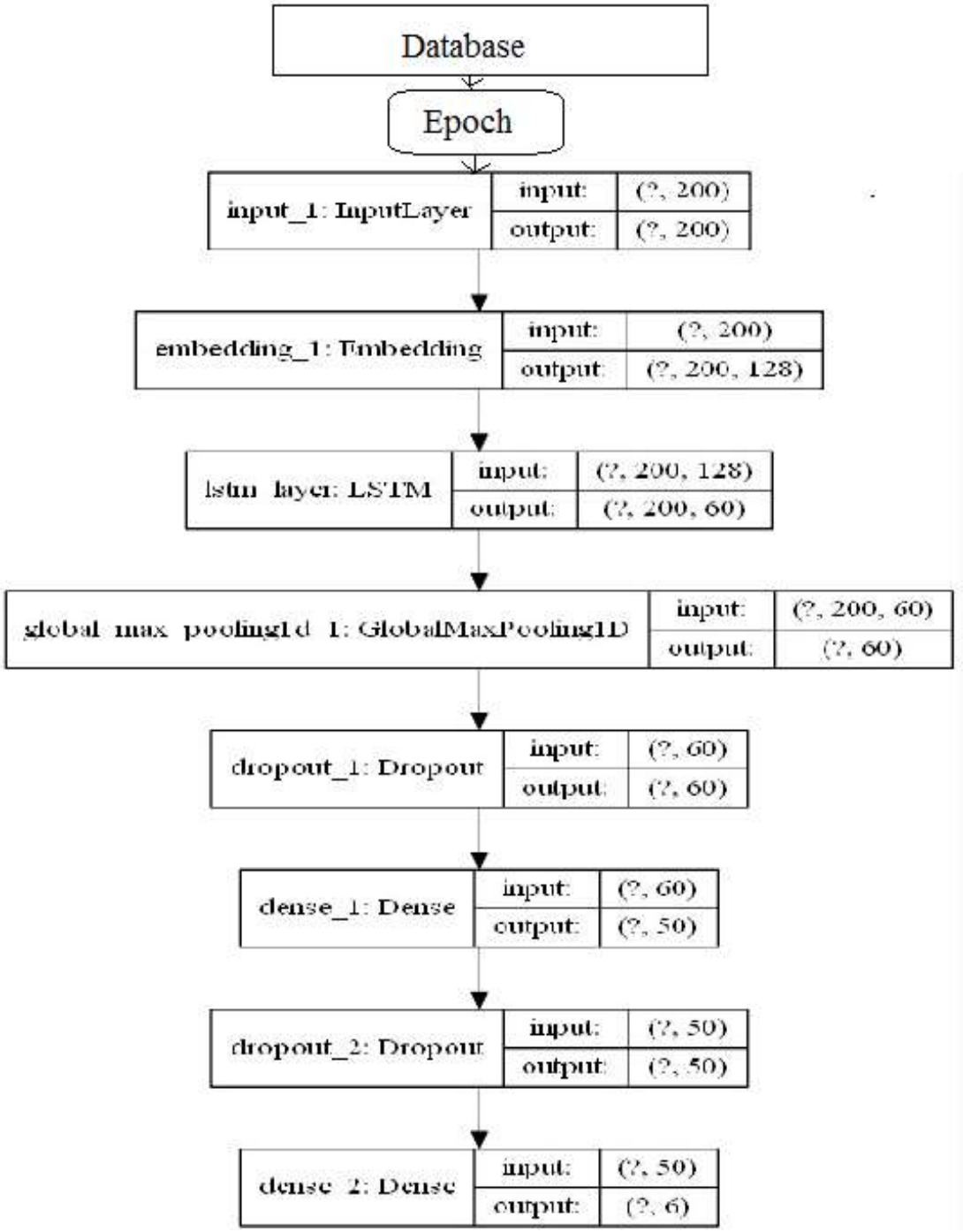


Figure 3 Architecture of the proposed system.

i. Input Layer

You always have to give a three-dimensional array or two-dimensional array as an input to your LSTM network. In this case, a two-dimensional array was used. Two hundred characters are inputted at a time. The epoch ensures that the entire database is passed forward and backward by dividing it into several smaller batches.

ii. Embedding Layer

The embedding layer provides a dense representation of words with their meanings. They are developed over

sparse representations. Word embedding can learn learned from text data. Words are represented by a dense vector. A vector represents the projections of the word into a continuous vector space. The position of a word in the learned vector space is known as its embedding.

iii. LSTM Layer

It is a stack-based memory. The stack-based LSTM has multiple hidden LSTM layers and each layer contains multiple memory cells. It has a hidden node, which learns and chooses what to forget using the long short-term memory.

iv. Global Max Pooling Layer

Global max-pooling layers are an essential part of the convolutional neural network (CNN). The use of a global max-pooling layer is to aggregate activations of spatial locations to produce a fixed-size vector in several states of the art. Global max pooling balances the contribution of all activities of a special coherent region by re-weighting all descriptors so that the impact of frequent and rare ones is equalized.

v. Dropouts Layers

Dropout serves as a regularization, especially where input and recurrent connections to LSTM units are probabilistically excluded from activation and weight updates when training a network. This has the effect of reducing overfitting and improving model performance.

EXPERIMENTAL PLAN

In our experiment, we conducted an empirical evaluation of both artificial neural networks and support vector machines as used in comment classification.

A. DATASET

- The dataset that will be used in this research is the Wikipedia dataset. In the dataset there are 6,127,462 English articles, 50,885,721 wiki pages, and 39, 555,860 users.

B. EXPERIMENTAL DESIGN

In our experimental design, we intend to perform two different experiments. The experiments are as follows;

i. Epoch based long short term memory (LSTM) approaches using Wikipedia dataset.

ii. Ensemble deep learning approach using Wikipedia dataset.

iii. EVALUATION MATRICS

Precision, recall, and f-measure (Schmidhuber, & Hochreiter, 1997); (Guggilla et al, 2016); (Mozafari et al, 2019); (Haralabopoulos, Anagnostopoulos, & McAuley, 2020); are the most popular metrics used in the evaluation. Precision is a measure of accuracy or correctness and recall is a measure of absoluteness or completeness. The formulas are described below.

$$\text{Precision}=\frac{TruePositives}{TruePositives+FalsePosittives} \quad \text{eqn 1}$$

$$\text{Recall}=\frac{True\ Positive}{True\ Positive+False\ Negative} \quad \text{eqn 2}$$

$$\text{F1}=\left(\frac{2}{recall^{-1}+precision^{-1}}\right)=2.\frac{precision.recall}{precision+recall} \quad \text{eqn 3}$$

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad \text{eqn 4}$$

Where TP=true positives, TN=true negatives, FP=false positive, FN=false negative.

iv. EVALUATION STRATEGY

To validate our approach, we have used 70% of the dataset as the training set and the remaining 30% of the dataset for testing purposes. we have experimented with the Twitter dataset.

4 . RESULT AND DISCUSSION

In the section below, the result of the toxic comment classification was presented. It was obtained that the result of Epoch based long short-term memory(LSTM) approach has higher achievement than the Ensemble Deep Learning approach. A threshold of 0.4 was used in this research. Table 4.1 below presents the result of the comparison.

Table 4.1 Result

Algorithms	Precision	Recall	F1-score	Accuracy
Ensemble Deep Learning approach	0.4213	0.5153	0.5770	0.5040
Epoch based long short term memory(LSTM)	0.8102	0.8024	0.8063	0.9331

4.1 Discussion

Pre-trained word embedding involves predicting a word based on one or more of its surrounding words. The word embedding obtained from such a training process, along with the trained model can then be used to help in a significant reduction of execution time. Embeddings are an improvement over sparse representations used in a simpler bag of word model representations in our dataset. Threshold controls the decision for converting a predicted probability or scoring in the comment classification system into a class label. It is governed by a parameter referred to as the "decision threshold," "discrimination threshold," or simply the "threshold." Hyperparameter value is used to control the learning process of the comment classification system. Epoch is used to positively affects the speed and quality of the learning process. The precision of Epoch based long short-term memory(LSTM) approach is 0.8102, while toxic comment classification with Ensemble Deep Learning approach has a precision of 0.4213. There is an improvement of 0.3889 in precision. The recall of Epoch based long short term memory(LSTM) approach is 0.8024, while toxic comment classification with Ensemble Deep Learning approach has a recall of 0.5153. There is an improvement of 0.2871 in the recall. The F1-score of Epoch based long short term memory(LSTM) approach is 0.8063, while Epoch based long short term memory(LSTM) approach has an F1-score of 0.5770. There is an improvement of 0.2293 in the F1-score. The accuracy of Epoch based long short-term memory(LSTM) approach is 0.9331, while toxic comment classification with Ensemble Deep Learning approach has an accuracy of 0.5040. There is an improvement of 0.4291 inaccuracy.

5. Conclusion and Future Work

This work aimed at improving the accuracy of comment classification by using the Epoch approach on long short term memory (LSTM). In the first section, the introduction of the whole work is presented. Section two presents related works of literature. The third section provides the methodologies of the proposed system. In the fourth section, the experimental design is described, while the result and discussion of the system are presented in the fifth section. We have an improvement of 0.4068 in precision, 0.2871 in a recall, 0.2293 in F1, and 0.4291 inaccuracy. In the future we recommend using diverse pre-trained embeds, we also suggest the epoch size to be up to 10, this will allow the model train to reduce the training errors and produce a better result.

References

Chakrabarty, J. (2012). Theory of plasticity. Elsevier.Hochreiter, Sepp, and Jurgen Schmidhuber, "Long short-term memory. *Neural computation*, *9*(8) (1997):1735-1780.

Long short-term memory. Neural Comput, 9(8), 1735-1780.

Guggilla, C., Miller, T., & Gurevych, I. (2016, December). CNN-and LSTM-based claim classification in online user comments. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 2740-2751).

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019, December). A BERT-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications* (pp. 928-940). Springer, Cham.

van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. arXiv preprint arXiv:1809.07572.Babu, B. Swapna., and G. Murali. "Performance Study on One Slope Solar Unit Conjoin with Parabolic Concentrator."

Babu, B. S., & Murali, G. Performance Study on One Slope Solar Unit Conjoin With Parabolic Concentrator.

Jabreel, M., & Moreno, A. (2019). A deep learning-based approach for multi-label emotion classification in tweets. Applied Sciences, 9(6), 1123.

Haralabopoulos, G., Anagnostopoulos, I., & McAuley, D. (2020). Ensemble Deep Learning for Multilabel Binary Classification of User-Generated Content. *Algorithms*, *13*(4), 83.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521 (7553). *DOI= http://dx. doi. org/10.1038/nature14539*, 436.

Haralabopoulos, G., Anagnostopoulos, I., & McAuley, D. (2020). Ensemble Deep Learning for Multilabel Binary Classification of User-Generated Content. *Algorithms*, *13*(4), 83.

Sridharan, M., & Swapna, T. R. (2019, June). Amrita School of Engineering-CSE at SemEval-2019 Task 6: Manipulating attention with temporal convolutional neural network for offense identification and classification. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 540-546).van Aken, Betty, et al." Challenges for toxic comment classification: An in-depth error analysis. *arXiv preprint arXiv:1809.07572(2018)*.

Van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. *arXiv preprint arXiv:1809.07572*.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019, December). A BERT-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications* (pp. 928-940). Springer, Cham.

Liang, J., Meyerson, E., Hodjat, B., Fink, D., Mutch, K., & Miikkulainen, R. (2019, July). Evolutionary neural automl for deep learning. In Proceedings of the Genetic and Evolutionary Computation Conference (pp. 401-409).

Ibrahim, M., Torki, M., & ElNainay, M. (2018, June). CNN based indoor localization using RSS time-series. In 2018 IEEE Symposium on Computers and Communications (ISCC) (pp. 01044-01049). IEEE.

d'Sa, A. G., Illina, I., & Fohr, D. (2020, February). BERT and fastText Embeddings for Automatic Detection of Toxic Speech. In SIIE 2020-Information Systems and Economic Intelligence.

Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep machine learning-a new frontier in artificial intelligence research [research frontier]. *IEEE computational intelligence magazine*, *5*(4),

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks, 18(5-6), 602-610.

Qin, P., Xu, W., & Guo, J. (2016). An empirical convolutional neural network approach for semantic relation classification. *Neurocomputing*, *190*, 1-9.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345-420.

Krenker, A., Bester, J., & Kos, A. (2011). Introduction to artificial neural networks. In Artificial neural networks-methodological advances and biomedical applications. IntechOpen.

Kajla, H., Hooda, J., & Saini, G. (2020, May). Classification of Online Toxic Comments Using Machine Learning Algorithms. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1119-1123). IEEE.

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated Machine Learning* (pp. 3-33). Springer, Cham.Zhang, X., Liao, Q., Kang, Z., Liu, B., Ou, Y., Du, J., ... & Fang, Z. (2019). Self-healing originated van der Waals homojunction with strong interlayer coupling for high-performance photodiodes. ACS Nano, 13(3), 3280-3291.

Tabassi, E., Burns, K. J., Hadjimichael, M., Molina-Markham, A. D., & Sexton, J. T. (2019). A Taxonomy and Terminology of Adversarial Machine Learning.

Wang, J. H., Liu, T. W., Luo, X., & Wang, L. (2018, October). An LSTM approach to short text sentiment classification with word embeddings. In Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 2018) (pp. 214-223).

Le, X. H., Ho, H. V., Lee, G., & Jung, S. (2019). Application of long short-term memory (LSTM) neural network for flood forecasting. *Water*, *11*(7), 1387