International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 3 March 2023, pp: 876-884 www.ijaem.net ISSN: 2395-5252

A Predictive Model for the Impact of Social Media Usage

[1]Bennett, E.O., [2]Omobolaji, O.S.
Department of Computer Science, Rivers State University, Port Harcourt, Nigeria

Date of Submission: 09-03-2023 Date of Acceptance: 18-03-2023

ABSTRACT: Social media has opened a new world for people around the globe. Most people are just a click away from getting a huge chunk of information. The information on social media has brought about people's opinions. This opinion has to do with the positive and negative outlooks of different individuals regarding a particular post or headline. Sometimes the posts on social media result in bullying and passing off hate comments which may lead to abuse, fights, depression and so on. Therefore, the need to analyze the impact of social media is a very crucial one, as a result this paper presents a model for the prediction of the impact level of social media contents. The paper made use of a social media datasets comprising various expressions of users. The users' comments were processed by performing text normalization, which involves selecting important features in the dataset. The selected features were used to train the model. The performance of the model in making classification into positive, negative and neutral were measured using an evaluation matrix. The model achieved an accuracy of 99.01%

Keywords – Social Media, Recurrent Neural Network, Long Short-Term Memory, Tweets

I. INTRODUCTION

The Internet has enabled an increasing amount of User-Generated Content (UGC) that potentially becomes the primary source of information for both consumers and businesses. The past decade has witnessed a dramatic change in the social network landscape with digital social media channels (such as blogs, online forums, and social networking sites) for Word-of-Mouth (WoM) supplementing traditional media channels (e.g., newspapers, television, and magazines). The rise of UGC on the Internet has fueled a fast-growing market in personal opinions [1]. More and more organizations and top executives are recognizing social media as an incredibly rich vein for gaining a better understanding of the online discussions and market opportunities, and for gaining feedback and evaluations of their own and their competitors' products and performance [2].

Social media applications host a large volume of opinions that reflect peoples' reactions to events. The reactions of social media function as user-driven data can be automatically classified in terms of their sentiment using opinion mining or machine learning techniques [3]. Analysis of social media data seems particularly useful when unexpected and potentially stressful events occur and there is need for understanding of how Internet users are making sense of them. Sentiment analysis over a large volume of user-generated data have been used in rapid reputation assessments (brand management, political marketing) or as an indication of how digital publics respond to events, associated with television shows, football games, significant news, or other meaningful events [4].

With the increasing availability of social media data sources, recent years have seen an emergence of academic and industrial research that taps into these data sources. However, the utilization of these data sources remains in an early stage, and outcomes are often mixed [5]. The major challenges are the inherent difficulties of tracking and quantifying the overwhelmingly large amount and unstructured set of data. A large body of extant research uses the quantitative summaries of UGC, such as overall valence and volume of user review ratings, to represent the users' opinions. However, recent research suggests that it is important to extract the multifaceted textual content in UGC, which highlights the need to delve deeper into the content of the online discussions. In addition, most previous studies focus purely on the effect of online UGC and social media, without considering their interactions with conventional medial sources.

Interactive tools such as visual analytic methods, data visualization, information design, morphological analysis, business decision mapping, and knowledge visualization could help make a large amount of complex information more readable and interpretable if integrated by computational approaches, as the effectiveness of

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 3 March 2023, pp: 876-884 www.ijaem.net ISSN: 2395-5252

most computational techniques is limited due to several factors. Interactive visual analytics provide intuitive ways of making sense of many posts available on social media. These techniques are now widely used in social media data and contribute to many exploratory data analysis areas [6]. Despite most social media visualization approaches that rely solely on geographical and temporal features, some systems can exploit the sentiments of the data which helps improve visualization. Besides disaster-related data management in social media, the ability to draw out important features could be used for a better and quick understanding of the situations, leading to rapid decision-making in critical situations. Moreover, the data produced by social media during disasters and events is staggering and hard for an individual to process. Therefore, visualization is needed for facilitating pattern discovery [7].

II. LITERATURE REVIEW

[8] examined the performance of many linguistic features of text documents, parts of speech, word relation and Term Frequency-Inverse Document Frequency (TF-IDF) together with some ensemble learning methods for sentiment classification. The scheme had several classification techniques such as Naïve Bayes, maximum entropy, and support vector machines being combined by a fixed-rule output combination, meta-classifier, and weighted combination rules.

[9] did a similar work by carrying out an empirical analysis of the performance of several different linguistic representations of text documents together with bigram, Uni-gram, term frequency, and the term Frequency-Inverse Document Frequency (TF-IDF) in union with three ensemble learning methods (bagging, random subspace and boosting) including five classification techniques (Naïve Bayes, Maximum entropy, decision tree, K-nearest neighbor and support vector machines). They were able to show that efficient classification models are necessary for sentiment analysis.

[10] designed a novel machine learning framework based on Recursive Auto Encoders (RAE). Auto encoders are used to efficiently learn feature encoding which are useful for sentiment classification and it incorporates the recursive interaction between context and polarity words in sentences in a unified framework while at the same time learning the basic features needed to make accurate predictions. Polarity is a binary value either positive or negative. Auto encoders are basically neural networks that learn a reduced dimensional representation of fixed-size inputs such as image patches or bag-of-word representations of text documents. The model used hierarchical structure and compositional semantics to understand sentiment instead of using bag-of-words representation.

An Aspect Based Sentiment Analysis (ABSA) model has been proposed by [11] that can predict the aspect related to a text for both in-domain and out-of-domain applications. This was achieved by the authors through the utilization of a pre-trained language model known as Bidirectional Encoder Representations from Transformers (BERT). The BERT model was fine-tuned to a sentence pair of a classification model. In order to discover a connection between an aspect and a text, the BERT model underwent some fine-tuning. This was done to teach the model to recognize when a sentiment context is presently based on the contextual representation. The model was used in classifying the aspects as well as sentiments using just a single sentence pair classification model. The results of the experiments show that the combined model achieves a higher level of success in aspect-based sentiment classification than previous state-of-the-art results.

[12] proposed a new method called Review Conventional Reading Comprehension (RCRC). They also investigated the possibility of transforming reviews into a useful resource for providing answers to user questions. They made use of the BERT model as their core model and suggested using a joint post-training approach to improve one's knowledge of both the domain and the task. They also investigated the applicability of the BERT model to two additional models based on the extraction and aspect of sentiment classification. The effectiveness of the post-training approach before the fine-tuning phase was tested, and the results showed that it achieved an accuracy of 84.26%.

[13] utilized the methods from Natural Language Processing in analyzing the feedback that customers had provided in Turkish regarding the banking services that they had received through Net Promoter Score (NPS) questionnaires. The banking industry saw the development of BERT-based sentiment classification models, which were then contrasted with more conventional methods. The efficiency of the methods was evaluated in a setting with limited resources, in which there was a limited quantity of labeled training data and none in the target domain. The results indicated that the Turkish Based BidirectionalEncoder Representationfrom Transformers (BERTurk-based

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 3 March 2023, pp: 876-884 www.ijaem.net ISSN: 2395-5252

model) performs better than the traditional models, and its performance is affected less by a reduction in the size of the training data set.

On the End to End Aspect Based Sentiment Analysis (E2E-ABSA) task, [14] investigated the modelling power of contextualized embedding derived from pre-trained language models using BERT. To address E2E-ABSA, the authors constructed a number of straightforward yet illuminating neural baselines. The experimental findings demonstrate that the proposed BERT-based architecture can outperform the works that are state-of-the-art even when using a straightforward linear classification layer. They also standardized the comparative study by always using a hold-out development datasets for model selection, which is something that was largely ignored by earlier works. This was another way that they standardized the study.

III. SYSTEM DESIGN

System design can be seen as the process of designing the elements of a system such as the architecture, modules and components, the different interfaces of those components and the data that goes through the system. The architectural design of this system is shown in Figure 1.

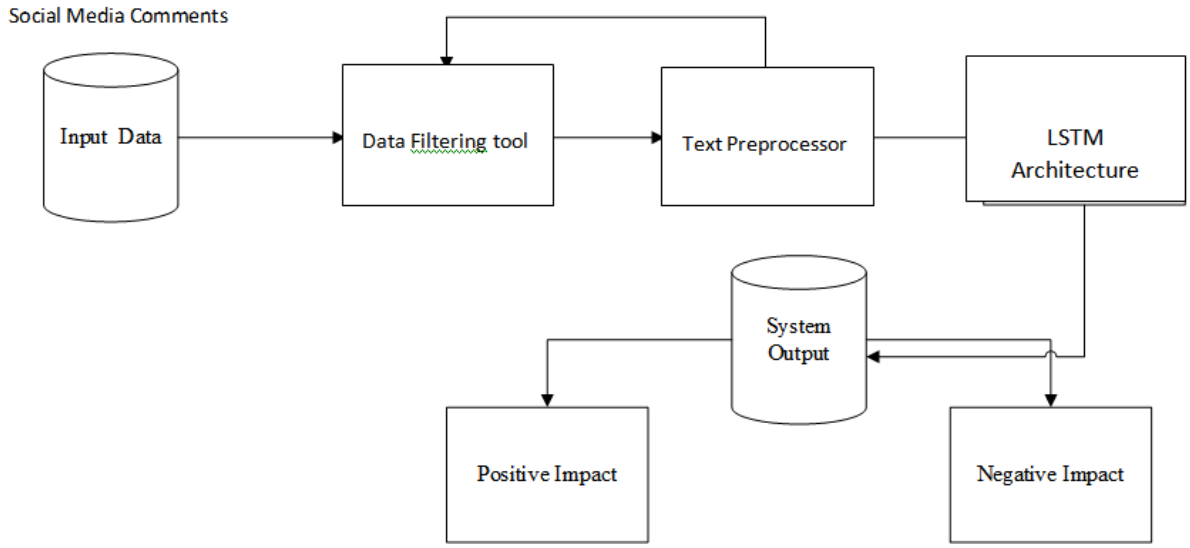

Figure 1. Architectural Design of the System

The system architecture is made up of the following components:

Input Data: These are user comments used in training the model

Data Filtering Tool: This involves using various techniques to check for completion and integrity of the data intended for use, removing empty spaces and reconciling data with their respective data types for optimum performance of the model.

Text Preprocessor: This component is responsible for transforming text into a clean and consistent format that can be fed into the model for further analysis and learning.

LSTM Architecture: Input data will be used to train the LSTM model. The algorithm for the Long Short-Term memory (LSTM) is a kind Recurrent Neural Network (RNN). Using Keras, TensorFlow framework will be used to generate the LSTM model. Our network is built using the sequential application programming interface (API) of Keras, which involves adding nodes and connecting them in a sequence.

Using Embedding technique, words (user comments) are encoded as 100-dimensional vectors. Pre-trained weights are supplied as embedding parameter. To avoid updating the embeddings, trainable may be set to false.

A layer is devoted to masking the embeddings of words for which no embedding has been previously taught and is set to zero. This layer is skipped while training Embeddings.

System Output: the output is the system's reaction to supplied data. The output might be a numerical sentiment score for each phrase usually on a scale of 0 (negative) to 1 (positive) .

Model Training and Evaluation: Accuracy, Precision, Recall and F1-score were used to assess the efficiency of the LSTM model. Using a classification report and a confusion matrix, the effectiveness of the model was dissected in further depth. This was used to explain why model

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 3 March 2023, pp: 876-884 www.ijaem.net ISSN: 2395-5252

sometimes predicted the test data accurately and sometimes erroneously. True positives, True negatives, false positives, and false negatives are all associated with accurate and inaccurate predictions.

3.1 Component Design

The component design is the breakdown of the component in the proposed system architecture. This is always needful because it shows further sub-components that were not made known in the design of the system architecture. Figure 2 shows the sub-component LSTM module.

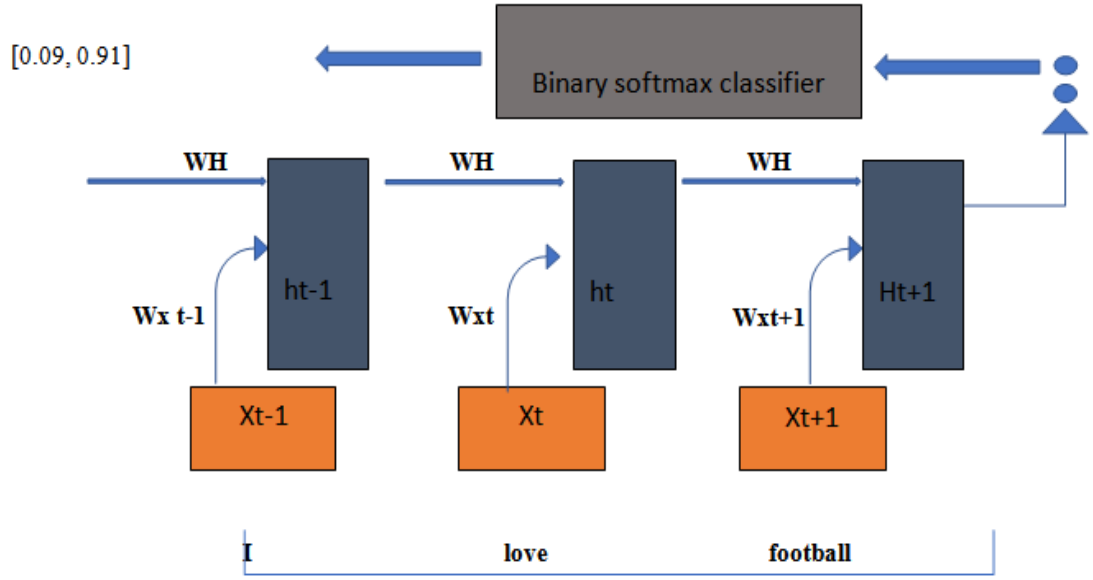

Figure 2. Component Design of our LSTM

The compact forms of the equations for the forward pass of an LSTM unit with a forget gate are:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$

$$\tilde{c}_t = \sigma_h(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \quad (5)$$

$$h_t = o_t \circ \sigma_h(c_t) \quad (6)$$

Where the initial values are $c_0=0$ and $h_0=0$ and thesubscript t indexes the time step. In this model, σ is the sigmoid activation function, tanh the hyperbolic tangent activation function, Xt the input at time t, Wi, Wc, Wf, Wo, Ui, Uc, Uf, Uo are weight matrices to regulate the input and bi, bc, bf , bo are bias vectors.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 3 March 2023, pp: 876-884 www.ijaem.net ISSN: 2395-5252

Table 1: Notations used in the Model design

S/N	Symbols	Notations	Definition
1	ft	Forget gate	The input gate decides which of the input (social media comment) is to be stored in the new state.
2	Ct	Output gate	The output gate is used to display the results(negative or positive or neutral comment)
3	WH	Weighting Matrix	This is used in measuring the weight of multiple inputs (X1……Xn)
4	ht-1	The previous hidden state	Show the value of input features before transformation.
5	xt	Input features of the social media network	The input feature can range from X1.…Xn.
6	Ct-1	The previous state	The previous states holds the input values before transforming it to the final output.
7	T	Text file from sample tweets	These are the collected data samples in form of texts.
8	S	Sentiment Strength	This is a calculated measurement that aims at the sentiment polarity of the input.
9	ti	Text token	These are tokens derived from the prepocessed raw text T.

3.2 Binary Softmax Classifier

The SoftMax function, which is used for the multi-classification model, returns the probabilities for each class as well as the likelihood that the target class will have a high probability. The equation calculates the exponential (e-power) of the specified input value as well as the total exponential value of all the input values. The output of the SoftMax function is the ratio of the input value's exponential to the sum of its exponential values. It is used in the various layers of neural networks as well as for the multi-classification job. The likelihood of the high value is higher than that of the other values. A neural network may be attempting to determine whether a dog is present in a photo. It should be able to generate a likelihood that a dog may appear in the image or not, but it may do so differently for each input. A SoftMax layer enables a multi-class function to be executed by the neural network.

3.3 Classification Criteria for model output

In order to be able to classify various inputs in their corresponding class (Positive, Negative, Neutral) the following are being considered:

- The sentiment strength S has to be >0.5 or = 1 for the output to be positive
- The Sentiment strength S has to be <=0 for the output of the input text to be Negative
- Other wise it is a neutral classification.

A Softmax activation function is used to work on the target text input in order to give a binary output and the maximum binary value that can be output by the function is 1. There are restrictions on how well SoftMax layers can determine multi-class probability. The price of SoftMax may increase as the number of classes increases. Candidate sampling may be a better remedy in some circumstances. A SoftMax layer will restrict the range of its calculations to a certain set of classes by using candidate sampling. Moreover, a SoftMax layer will not function in scenarios when an item belongs to many classes because it thinks that there is only one member per class. The option in that situation is to employ multiple logistic regressions.

3.4 LSTM network approach

The network takes the current input, the previous hidden state and the previous internal cell state, then calculates the values of the four different gates by following the below steps: -

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 3 March 2023, pp: 876-884 www.ijaem.net ISSN: 2395-5252

- For each gate, calculate the parameterized vectors for the current input and the previous hidden state by element-wise multiplication with the concerned vector with the respective weights for each gate.
- Apply the respective activation function for each gate element-wise on the parameterized vectors. Below is the list of the gates with the activation function to be applied for the gate.
- Calculate the current internal cell state by first calculating the element-wise multiplication vector of the input gate and the input modulation gate, then calculate the element-wise multiplication vector of the forget gate and the previous internal cell state and then adding the two vectors.
- Calculate the current hidden state by first taking the element-wise hyperbolic tangent of the current internal cell state vector and then performing element wise multiplication with the output gate.

Algorithm 1: Algorithm for classification

INPUT: Text File (comment or review) T, The sentiment lexicon L.
OUTPUT: Smt = {**P**, **Ng** or **N**} and strength S, where **P:** positive, Ng: Negative, N: Neutral.
INITIALIZATION: SumPos and SumNeg =0, where
SumPos: accumulates the polarity of positive tokens ti-smt in **T**
SumNeg: accumulates the polarity of negative tokens ti-smt in **T**
Begin
1. For each ti □ T do
2. Search for ti in L If ti □ Pos-list then
3. SumPos. = SumPos + ti-smt
4. Else if ti □ Neg-list then
5. SumNeg = SumNeg + ti-smt
6. End If
7. End For
8. If SumPos > |SumNeg| then
9. Smt = P
10. S = SumPos / (SumPos + SumNeg)
11. Else if SumPos < |SumNeg| then Smt = Ng
12. S = SumNeg / (SumPos + SumNeg)
13. Else
14. Smt = N
15. S = SumPos / (SumPos + SumNeg)
16. Endif
17. **END**

Note: the more the negative polarity, the lower the sentiment strength S and the more the positive polarity the more the sentiment strength S.

3.5 Text Preprocessing

3.5.1 Dataset

We train and evaluate the model using datasets from twitter. These datasets include tweets that have been classified as either good or negative in total. If it is saved on your computer as a text file, we just load it. The text is then changed to lower case, and all punctuation is removed. All of the strings are together in one enormous string. Each comment must now be broken out and stored in distinct list components. For instance [comment 1, comment 2, comment 3…, comment n]

3.5.2 Tokenization

The process of tokenizing or dividing a string of text into a list of tokens, is known as tokenization. Tokens can be thought of as little components; for example, a word may serve as a token in a sentence, and a sentence may serve as a token in a paragraph. All or any Natural Language Processing (NLP) tasks require tokenizing, which is the process of breaking a string into its intended components. Because of tokenization, there is no singular right. The right algorithm will vary depending on the application. Because that sentiment information is frequently sparsely and unusually expressed by a single cluster of

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 3 March 2023, pp: 876-884 www.ijaem.net ISSN: 2395-5252

punctuation, tokenization may be even more crucial in sentiment analysis than it is in other NLP applications. For most NLP tasks, you will create an index mapping dictionary that your commonly used terms are Lower indices are given to frequently occurring words. The Counter method from the Collections library is used to accomplish this. Using the vocabulary from all our comments, we have so far compiled a list of comments and index mapping dictionaries. All of this was done to encode comments (replace words in our comments by integers) and have now produced a list of lists. Each individual comment is saved in a single, enormous list and is composed of a list of integer or floating values.

3.5.3 Word Embedding

A text mining technique called word embedding is used to determine the relationships between words in textual data (Corpus). The context in which a word is used determines its syntactic and semantic meaning. According to the idea of distributional hypothesis, words that occur in comparable contexts have semantically similar meanings. Word embeddings occur within the deep learning frameworks such as TensorFlow, Keras and is handled by an embedding layer which stores a lookup table to map the words represented by numeric indexes to their dense vector representations. Language relationships are captured via embeddings.

IV. RESULTS AND DISCUSSION

4.1 System Requirements

For this study, the following system requirements listed are necessary for the successful implementation of the system.

4.1.2 Hardware Requirements

In order to implement the system, the following are the minimum hardware requirements:

1. A microprocessor with minimum of 2GHz Clock Signal such as Pentium IV processor.
2. 2GB of Random Access Memory (RAM)
3. Compatible mouse, keyboards.
4. Hard disk size of 250GB.

4.1.3 Software Requirements

The software requirements for the design of the proposed system are:

i. Microsoft Windows 7 and above.
ii. A web browser
iii. Anaconda (Python Distribution)

The system starts by acquiring social media data that comprises various social media posts or comments. In this model, learning is the first step, and predicting is the second. The model is trained with the datasets in the learning process, and it classifies the train datasets according to that perception. During the learning process, the model is trained with the datasets, and it then classifies the train datasets according to that perception. To avoid under fitting, we should use large datasets and a well-completed learning process. Based on the training, the model learns how to classify, and the model is then evaluated with the test set during the testing process. The LSTM model was trained on training steps of 20, batch size of 32, activation function = soft sign, activation = linear, dropout = 0.1. After training, the model achieved an accuracy result of 99.01% for the training data, and 99.10% for testing.

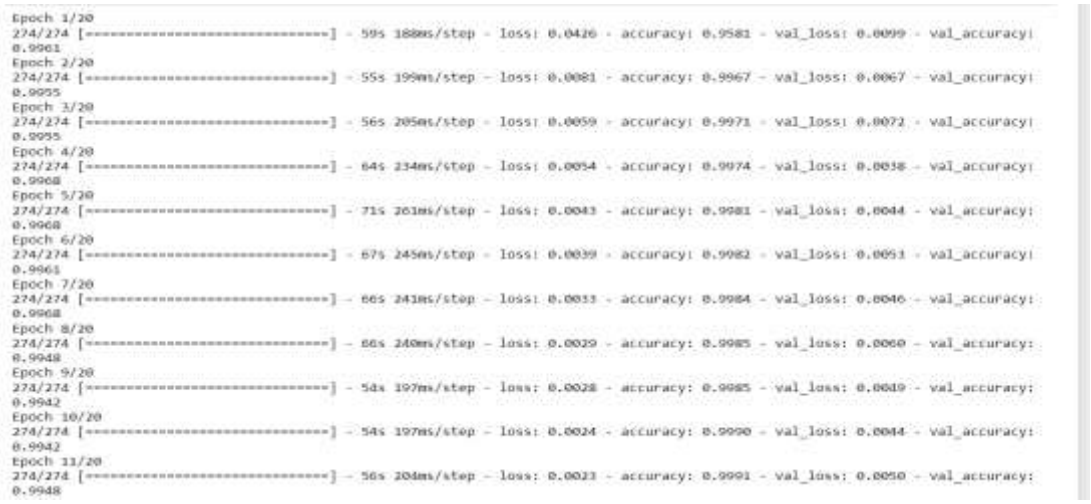

Figure 3: Training Epoch

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 3 March 2023, pp: 876-884 www.ijaem.net ISSN: 2395-5252

Figure 3 reveals the results of the learning phase. Here tensor flow is used as the back end of this model, which helps in various machine learning tasks. The model is trained using the training datasets before being put to the test in the following phases. We can see how the output improves over time as the epochs pass. The model's efficiency improves with each sequential epoch, meaning that it learns from its experience.

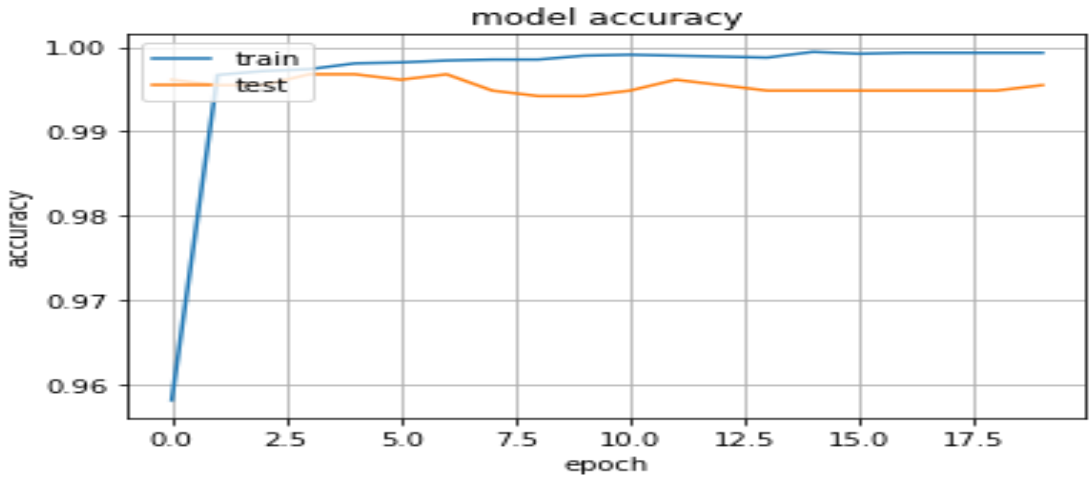

Figure 4: Accuracy for both training and test data

Figure 4 depicts the accuracy of the LSTM model for the prediction of the impact of social media usage. It shows that the model achieved 99.01% accuracy for both testing and validation.

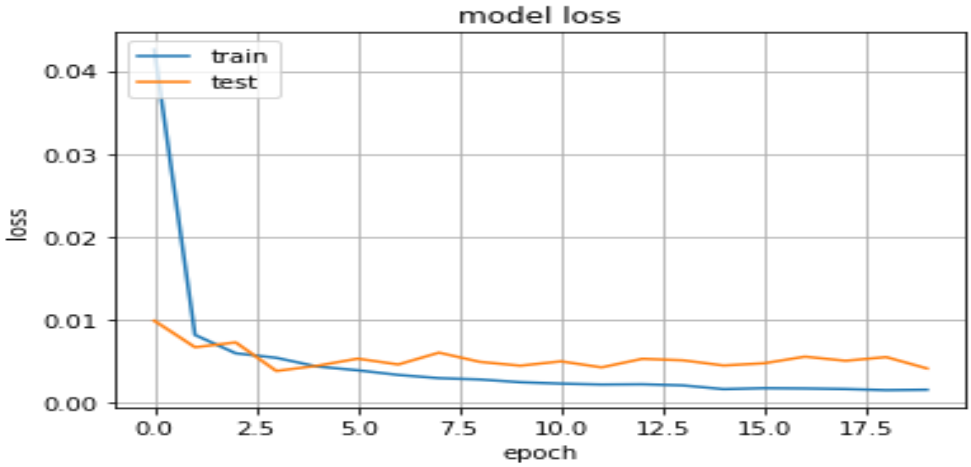

Figure 5: Loss values for both training and testing data.

Figure 5 shows that the model has both training and validation loss below 0.99%. the loss value visibly reduced as each sequential epochs went past.

V. CONCLUSION

This paper proposed a model that classifies User Generated Content (UGC) into categories of positive, negative or neutral. The proposed model used sentiment classifiers that aid in the classification of emotion in text sequences. The analysis of these user generated contents is important and can be applied in different areas like product review, social media posts, conversation regulation, and other related data by taking actionable insight of consumer views and opinions hereby leading to faster and more precise decisions. A sentiment count on various posts can be taken from the model's prediction since the model is very efficient.

REFERENCES

[1]. Archak, N., Ghose, A., Ipeirotis, P.G (2011). Deriving the pricing power of product features by mining consumer reviews, Management Science 57 (8)1485–1509.

[2]. Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. Decision Support Systems, 55(4), 919–926.

[3]. Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with Applications, 40(16), 6266-6282.

[4]. Brooker, P., Barnett, J., Cribbin, T., Lang, A.& Martin, J. (2013). User-driven data capture: locating and analyzing twitter conversation about cystic fibrosis without keywords. In SAGE research methods cases. London: SAGE Publications.

[5]. Godes, D. and Mayzlin, D. (2004). "Firm-Created Word of Mouth Communication: A Quasi-Experiment," Harvard Business School Working Paper.

[6]. Alfredo, C., Denis, P. & Jaime, N. (2015). Identifying relevant messages in a twitter-based citizen channel for natural disaster situations. In Proceedings of the 24th International Conference on World Wide Web Companion, 1189–1194,

[7]. Beigi, G., Hu, X., Maciejewski, R. & Liu, H. (2016). An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief. Studies in Computational Intelligence, 313–340.

[8]. Wu, Z., Wang, Y., Wu, J., Cao, J., & Zhang, L. (2015). Spammers detection from product reviews: A hybrid model. 2015 IEEE International Conference on Data Mining, 1039–1044.

[9]. Wang, H., Xu, Z., Fujita, H., & Liu, S. (2016). Towards felicitous decision making: An overview on challenges and trends of big data. Information Sciences, 367, 747–765.

[10]. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, Christopher D., Ng, Andrew Y., and Potts,(2015). Ch.: Recursive deep models for semantic compositionality over a sentiment tree bank.Proceedings of the 2013 conference on empirical methods in natural language processing.1631–1642.

[11]. Chouikhi, H., Chniter, H., and Jarray, F. (2021). Arabic sentiment analysis using bert model. In 13th International Conference on Computational Collective Intelligence (ICCCI). Springer, Cham.

[12]. Xu, C, Liu, Y. (2019). Word of mouth for movies: Its dynamics and impact on box office revenue. Journal of Marketing,70(3) 74–89.

[13]. Masarifoglu M, Umit T., (2021), Sentiment Analysis of Customer Comments in Banking using BERT-based Approaches. SIU 2021: 1-4.

[14]. Li, X., Hitt, L., (2008). Self-selection and information role of online product reviews. Information Systems Research, 19(4) 456–474.