International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

TEXT CLASSIFICATION USING DEEP LEARNING

[1]E. Moukthika, [2]E. Sai Vatsalya, [3] J. Tejah Phani, [4]G. Sahithi Reddy,

[1,2,3,4]Department of AIML

School of Engineering.

Malla Reddy University, Hyderabad

[5]Sujit Das

Assistant Professor

[5]Department of AIML

Email: sujit.das@mallareddyuniversity.ac.in

School of Engineering, Malla Reddy University.

ABSTRACT

Our daily usage of the internet generates vast volumes of text, much of which is unfiltered. Unstructured data must typically be categorised in order to increase the speed at which a particular text is interpreted. Unstructured text data can be distinguished using a branch of natural language processing called text classification. Because machine learning can generate intricate prediction functions dynamically, it is frequently employed in the categorization of textual data. Similar to how statistical models may explain the relationship between two or more random variables, textual data is frequently classified using statistical models. The challenge of performing sentiment analysis in an e-commerce setting is often difficult. Neural network methods for machine learning Recent studies on text categorization using neural network-based applications have demonstrated promising results. The model still finds it difficult to think about regional characteristics and words that depend on the information in the phrase. This research suggested using deep learning to create more exact sentences that use the classification of earlier texts. A recurrent neural network (RNN) with the architecture of long short-term memory (LSTM) is one of the deep learning techniques employed. As a result of how we use the internet on a daily basis, huge volumes of text are created regularly, and the majority of this text is unfiltered. Unstructured data must typically be categorised in order to increase the speed at which a particular text is interpreted. Unstructured text data can be distinguished using a branch of natural language processing called text classification. Because machine learning can generate intricate prediction functions dynamically, it is frequently employed in the categorization of textual data. Similar to how statistical models may explain the relationship between two or more random variables, textual data is frequently classified using statistical models. The challenge of performing sentiment analysis in an e-commerce setting is often difficult. The performance of Naive Bayes and Decision Tree machine learning approaches is limited when it comes to sentiment analysis. A comparison of recurrent neural networks (RNN) is done in this paper.

Support Vector Machine (SVM) is used to categorise consumer product review data according to whether the remarks are favourable or negative. In order to attain the best results, this study prefers to apply Long Short-Term Memory (LSTM) to improve the conventional RNN. The research's findings demonstrate that RNN, with an accuracy of 87.57.86%, outperforms state-of-the-art SVM, which has an accuracy of 59.67%.

KEYWORD: Text classification, Recurrent Neural Networks, Long Short-Term Memory.

INTRODUCTION:

One of the key components of Natural Language Processing (NLP) research, text categorization has applications in a number of areas, including sentiment analysis, document classification, text categorization, and information retrieval. Traditional machine learning techniques, including Nave Bayes, K-Nearest Neighbour, Support Vector Machine, and Logistic Regression, have been suggested in prior text categorization research. Traditional machine learning algorithms have been successful in classifying text, but they have drawbacks when handling multi-label data and huge datasets. It is an often-preferred method for classifying textual material. Text is often seen as a recorded or spoken piece of material in its raw form. Text may alternatively be characterised as any language that a reader can comprehend. It might be as basic as one or two words or as complicated as a rationally connected series of phrases. On the other hand, classification presents a problem in identifying the groups with which a new finding is connected, based on a training collection of data that includes specific observations that indicate a group member. The qualifying procedure in the supervised learning model is input into groups of similar classes in the unlabelled test dataset. The class group is carefully documented and the models are correctly trained for classic SVM and RNN classification tasks to guarantee that the data is properly assigned to the right class. Both linear and nonlinear classification may be done using support vector machines. Unsupervised learning refers to machine learning techniques that analyse text by grouping the text's format into distinct clusters without providing a labelled response or output. On the other hand, no training data are given to the machine. Two earlier methods make up the machine learning technique. Researchers have discovered a small number of unlabelled details that might improve accuracy, leading to the creation of semi-supervised learning methodologies. Classified data is scarce in the application industry, whereas unmarked datasets are easily accessible and inexpensive. Because seasoned programmers are required to identify unknown data patterns, labelling instances is particularly challenging. The algorithm for semi-supervised learning addresses the issue and acts as a bridge between supervised and unsupervised learning. It has been argued that semi-supervised learning can help alleviate these issues since it allows a testing group to identify unknown test data using only a small quantity of training data. According to the existing literature, the research gap was based on two main issues: a lack of a sufficient dataset and a failure to apply the appropriate method to a text classification task, such as using CNN for a task where statistical methods like Naive Bayes, SVM, and Deep Learning Techniques exist. found that it is comparatively easier to construct categories that broadly describe the data included in the data collections they used. Despite the multi-task learning model described by unique approach's better performance when compared to MBOW, MV-RNN, RNTN, DCNN, and PV the performance still falls short of that of the unique LSTM model with the greatest performance.

AIM: Identifying the predictive algorithm of deep learning for text classification.

OBJECT: The main objective of text classification using deep learning is to Automatically assigning predetermined categories or labels to a given text document or fragment of text is the goal of text categorization using deep learning. The underlying patterns and characteristics in text data may be learned by training deep learning models, especially those based on neural networks, allowing them to categorise text documents into different groups or categories. Text classification seeks to automate the process of classifying text documents by utilising deep learning techniques, enabling effective information retrieval, document organisation, sentiment analysis, spam filtering, and language identification across a variety of applications and sectors.

LITERATURE REVIEW:

Using very little label data, Shan et al. Developed SSL for sentiment classification, trained a semi-supervised deep neural network with a different configuration, and compared the results to the original, a supervised deep neural network trained with an equivalent number of labelled results. The training and test datasets, which were tagged, were divided into two portions for the research. As the labelled training dataset shrinks, the performance of the classifier dataset degrades, according to the research. The reduction is computed using binary cross-entropy, and the Adam process is employed for optimisation. The unmasked datasets were selected at random, which allowed the researchers to achieve the desired outcome.

Boiy and Moens developed a machine learning method for sentiment analysis in multilingual online content. Observational studies of public opinion in online reviews, blogs, and group texts produced in French, Dutch, and English They are chosen from a sample of sentences and phrases that are carefully divided into positive, neutral, and negative remarks about a certain incident. For texts in French and Dutch, the study was able to reach precisions of 68% and 70%, respectively, and 83% for those in English. The study only included a few languages.

A rapid miner was used by Chauhan to compare and contrast supervised machine learning algorithms. In this study, four supervised machine learning methods—Neural Network, Naive Bayes, Support Vector Machine, and Decision Tree—were compared. These methods are used to analyse emotions based on different output functions. The study's findings demonstrated that the support vector machine outperforms the other three supervised machine learning algorithms in terms of efficiency. They came to the conclusion that Support Vector Machine has a high score of 68.29% compared to Decision Tree and Naive Bay with 61.11% and 57.08%, respectively, based on the analysis of the various data for all emotion classification algorithms. SVM fared better, although its accuracy measured in percentage terms was poor.

To enhance the user experience, Kumar & Zymbler developed a machine learning approach for researching tweets. Using the word embedding in the Glove dictionary framework and the n-gramme approach, features were extracted from tweets.

The tweets were mapped to positive and negative categories using SVM, ANN, and CNN classification models. Support vector machines and artificial neural network variations have been shown to perform worse than convolutional neural networks. After 2700 iterations on the validation range, it had an accuracy of 87.3%, which is respectable compared to the artificial neural network model's accuracy of 69.16%. It is also obvious that CNN is able to process text data more precisely and is more effective than the AN

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

METHODS:

AG News is a four-category subject classification system for online news stories that includes names and descriptions.

classes: business, sports, entertainment, and world.

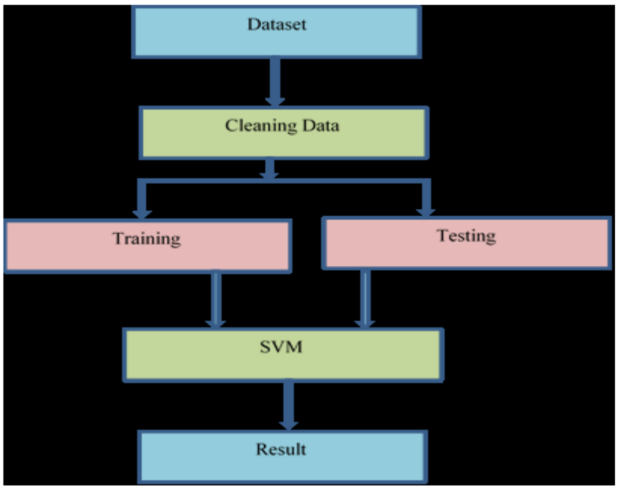
The dataset is already divided into the train (120000 text examples) and test (7600 text examples) sets.

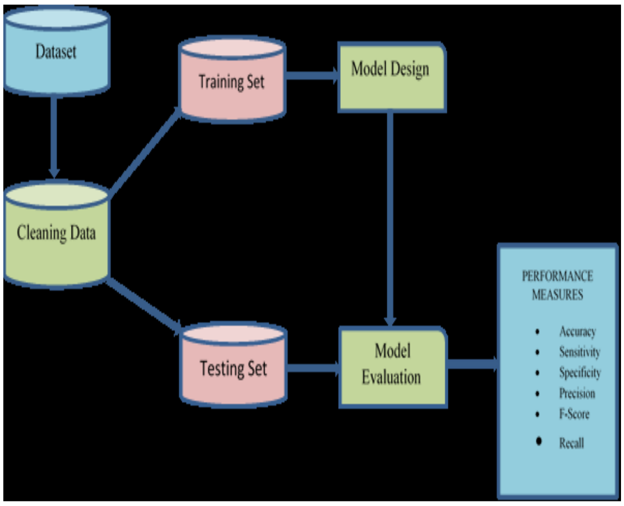
Overview of Research Method

Our dataset is first loaded into the development environment in this study, after which it undergoes a cleaning procedure. The data is divided into training and testing categories. Each classifier has been designed and its performance metrics have been assessed.

Long Short-Term Memory

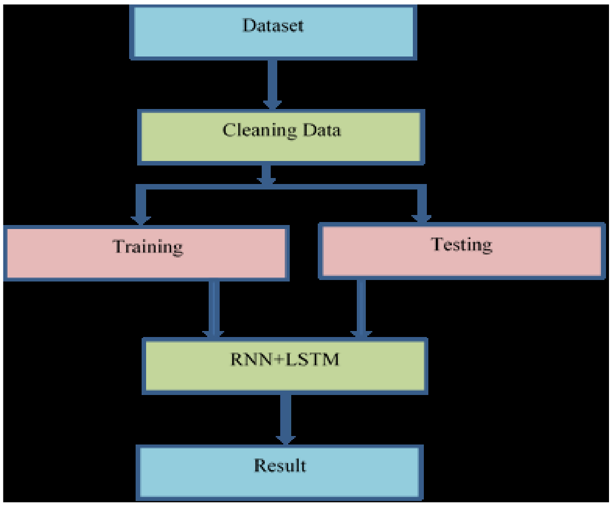
One of the RNN designs, long-short-term memory (LSTM), has gained popularity among NLP researchers for its superior capacity to model and learn from sequential input. Models In several areas, including language modelling, tagging issues, and sequence-to-sequence predictions, LSTM has demonstrated exceptional results. The goal of LSTM is to address the RNN issue known as gradient disappearing and exploding. Recurrent neural networks' hidden vectors are swapped out for memory blocks with gates in LSTM. The input





gate, forget gate, and output gate are the three levels that the LSTM gates execute.

FIG: Workflow Diagram



Performance Metrics

Similar assessment criteria may be used to FIG: Workflow Diagram of LSTM Model

assess the effectiveness of deep learning-

based text categorization models on a business dataset. However, there are a few other measures that can offer precise insights into the model's performance in the context of a business dataset.

These pertinent assessment metrics are listed below:

Precision, recall, and F1 Score per Class: Text classification tasks frequently involve several classes or categories in commercial datasets. It's critical to assess each class's accuracy, recall, and F1 score separately. This enables you to comprehend the model's performance in particular categories and spot potential development areas.

Accuracy per Class: In addition to measuring overall accuracy, determining accuracy per class may be used to evaluate how well the model works for each distinct business category. It assists in finding possible biases or imbalances in the forecasts made for the various classes.

The precision-recall curve is equally important in business datasets, even if the AUC-ROC curve is a typical assessment measure. It aids in the analysis of the trade-off between recall and accuracy at different probability thresholds. This is especially helpful when the business activity places a higher value on recall or precision than accuracy.

RECURRENT NEURAL NETWORK: The gradient disappearing and exploding during training affects classic RNNs. RNN is a category of deep learning since it processes data. Without defining characteristics, automatically. In the course of learning, RNN does not simply ignore prior knowledge. This sets RNN apart from conventional artificial neural networks. A

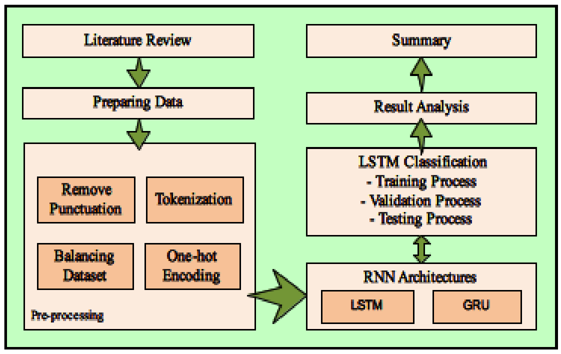
component of the neural network used to analyse sequential data is the RNN. Due to looping, which is a feature of RNN design, information from the past may be automatically kept. RNN can process the FIG: Workflow Diagram of RNN.

input sequence using its internal states (memory). It equips RNN with the ability to do NLP, speech recognition, and handwriting recognition tasks.

DATA SET: AG News is a four-category subject classification system for online news stories that includes names and descriptions.

classes: business, sports, entertainment, and world.

The dataset is already divided into the train (120000 text examples) and test (7600 text examples) sets.



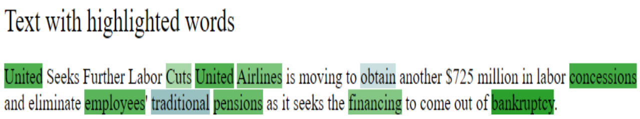


FIG: Research Methodology.

CONFUSION MATRIX:

- These are the evaluation measures to evaluate

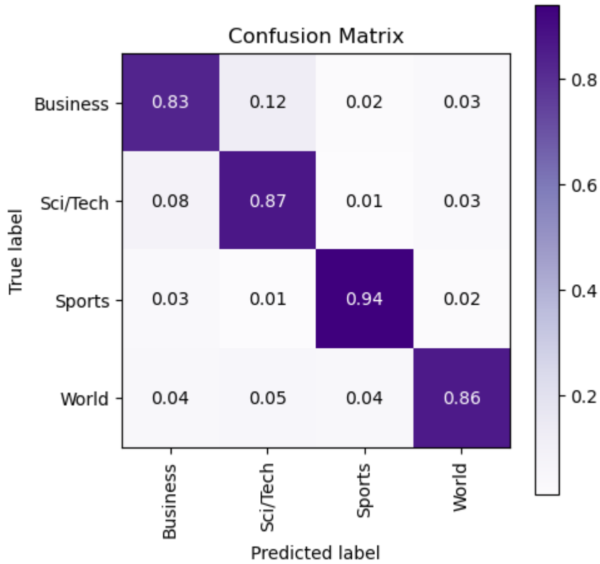
Classification Report :

	precision	recall	f1-score	support
World	0.90	0.86	0.88	1900
Sports	0.93	0.94	0.93	1900
Business	0.84	0.83	0.84	1900
Sci/Tech	0.83	0.87	0.85	1900
accuracy			0.88	7600
macro avg	0.88	0.88	0.88	7600
weighted avg	0.88	0.88	0.88	7600

the performance of the model.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

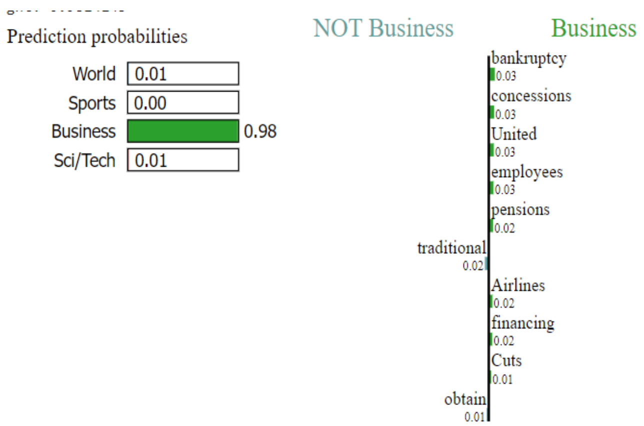
- Dark blue boxes are the correct prediction with trained model and skyblue boxes shows the wrong predictions.



Y-axis: Ture Label

X-axis: Predicted Label

RESULTS:



Performance Metrics	Results (%)
Accuracy	76.57
Precision	87.78
F1 Score	81.75

Table: Performance evaluation measures of SVM.

FIG: Performance evaluation measures of RNN.

CONCLUSION:

According to the experiments described above, the proposed RNN-LSTM with sequential models has been successful in classifying the text. The greatest accuracy, recall, and f1-score scores on the Adam optimizer are 97 after training 4 models of 1-Layer LSTM with various hyper-parameters. whereas 96.53% is the greatest accuracy. In this investigation, a comparison of recurrent

For customer review on text categorization, neural networks, and support vector machines a benchmark dataset for online shopping. On the same dataset, the support vector machine technique and the recurrent neural network were both implemented. F-score, accuracy, specificity, sensitivity, false positive rate, false negative rate, and precision were used to obtain the outcome.

Using customer reviews, RNN will help businesses understand what customers think of a product and influence others' purchasing decisions. Additionally, it should be highlighted that the two algorithms developed in this study outperformed the already available methods by a wide margin.

REFERENCE:

[1] Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, & Brown. (2019). Text Classification Algorithms: A Survey. Information, 10(4), 150. https://doi.org/10.3390/info10040150
[2] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2020). Deep Learning-Based Text Classification: A Comprehensive Review. http://arxiv.org/abs/2004.03705
[3] Chaya Liebeskind, S. L. (2020). Deep Learning for Period Classification of Historical Hebrew Texts. Journal of Data Mining and Digital Humanities ISSN 2416-5999, an Open-Access Journal.

[4] Yi, S., & Liu, X. (2020). Machine learningbased customer sentiment analysis for recommending shoppers, shops based on customers' review. Complex & Intelligent Systems, 6(3), 621–634. https://doi.org/10.1007/s40747-020-00155-2
[5] Thangaraj, M., & Sivakami, M. (2018). Text Classification Techniques: A Literature Review. Interdisciplinary Journal of Information, Knowledge, and Management, 13, 117–135. https://doi.org/10.28945/4066
[6] Wahdan, A., AL Hantoobi, S., Salloum, S. A., & Shaalan, K. (2020). A systematic review of text classification research based on deep learning models in the Arabic language. International Journal of Electrical and Computer Engineering (IJECE), 10(6), 6629.https://doi.org/10.11591/ijece.v10i6.pp6629-6643
[7] C A Padmanabha Reddy, Y., Viswanath, P., & Eswara Reddy, B. (2018). Semi-supervised learning: a brief review. International Journal of Engineering & Technology, 7(1.8), 81. https://doi.org/10.14419/ijet.v7i1.8.9977
[8] Kazeem Moses, A., Joseph Bamidele, A., Roseline Oluwaseun, O., Misra, S., & Abidemi Emmanuel, A. (2021). Applicability of MMRR load balancing algorithm in cloud computing. International Journal of Computer Mathematics: Computer Systems Theory, 6(1), 7-20.
[9] Brownlee, J. (2020). Supervised and unsupervised machine learning algorithms. Machine Learning Mastery. https://machinelearningmastery.com/supervised-and-unsupervised-machine-learningalgorithms/
[10] Abiodun, M. K., Awotunde, J. B., Ogundokun,R. O., Misra, S., Adeniyi, E. A., Arowolo, M. O., & Jaglan, V. (2021). Cloud and Big Data: A Mutual Benefit for Organization Development.
In Journal of Physics: Conference Series, 1767(1) IOP Publishing.
[11] Jin Huang, and Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms.
IEEE Transactions on Knowledge and Data Engineering, 17(3), 299–310. https://doi.org/10.1109/TKDE.2005.50
[12] Ramola, R., Jain, S., and Radivojac, P. (2019). Estimating classification accuracy in positiveunlabeled
learning: characterization and
correction strategies. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 24, 124–135. http://www.ncbi.nlm.nih.gov/pubmed/30864316
[13] Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., and Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802
[14] Hartmann, J., Huppertz, J., Schamp, C., and Heitmann, M. (2019). Comparing automated text classification methods. International Journal of Research in Marketing, 36(1), 20–38. https://doi.org/10.1016/j.ijresmar.2018.09.009
[15] Soria, D., Garibaldi, J. M., Ambrogi, F., Biganzoli, E. M., and Ellis, I. O. (2011). A 'non-parametric' version of the naive Bayes classifier. Knowledge-Based Systems, 24(6), 775–784. https://doi.org/10.1016/j.knosys.2011.02.0
[16] Tsangaratos, P., and Ilia, I. (2016). Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. CATENA, 145, 164–179. https://doi.org/10.1016/j.catena.2016.06.004
[17] Burba, F., Ferraty, F., and Vieu, P. (2009). k - Nearest Neighbour method in functional nonparametric regression. Journal of Nonparametric Statistics, 21(4), 453–469. https://doi.org/10.1080/10485250802668909

[18] Thanh Noi, P., and Kappas, M. (2017).Comparison of Random Forest, k-NearestNeighbor, and Support Vector MachineClassifiers for Land Cover Classification Using Sentinel-2 Imagery. Sensors, 18(2), 18.https://doi.org/10.3390/s18010018.

[19] Apaydin, H., Feizi, H., Sattari, M. T., Colak, M. S., Shamshirband, S., & Chau, K.-W. (2020). Comparative Analysis of Recurrent Neural

Network Architectures for Reservoir Inflow Forecasting. Water, 12(5), 1500. https://doi.org/10.3390/w12051500

[20] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

[21] Shan Lee, V. L., Gan, K. H., Tan, T. P., & Abdullah, R. (2019). Semi-supervised Learning for Sentiment Classification using a Small Number of Labeled Data. Procedia Computer Science, 161, 577–584. https://doi.org/10.1016/j.procs.2019.11.159

[22] Boiy, E., & Moens, M.-F. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. Information Retrieval, 12(5), 526–558. https://doi.org/10.1007/s10791-008-9070-z

[23] Chauhan, P. (2017). Sentiment Analysis: A Comparative Study of Supervised Machine Learning Algorithms Using Rapid miner. International Journal for Research in Applied Science and Engineering Technology, V(XI), 80–89. https://doi.org/10.22214/ijraset.2017.11011

[24] Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. Journal of Big Data, 6(1), 62. https://doi.org/10.1186/s40537-019-0224-1

[25] Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, M. R. and C.-J. L. (2010). Training and Testing Low-degree Polynomial Data Mappings via Linear SVM. Journal of Machine Learning Research.