e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Conference:01/December-2022 Impact Factor- 6.752 www.irjmets.com
1st National Conference on Applications of soft Computing Techniques in Engineering NCASCTE-2022
Organized by Department of Electrical & Electronics Engineering, ACE Engineering College, Hyderabad

COMPARATIVE STUDY OF FAKE NEWS DETECTION BETWEEN MACHINE LEARNING AND DEEP LEARNING APPROACHES

Dr. Khaleel Ur Rahman Khan*1, Gunti Sai Nikhil*2, Mekalateja Sree*3, Merugu Akhil*4

*1Dean & Professor, Department Of Computer Science And Engineering, ACE Engineering College, Hyderabad, Telangana, India.

*2,3,4Student, Department Of Computer Science And Engineering, ACE Engineering College, Hyderabad, Telangana, India.

DOI: https://www.doi.org/10.56726/IRJMETS-NCASCTE202230

ABSTRACT

Most of the people now-a-days prefer to read the news via social media over internet. Many websites are publishing the news and provide the source of authentication. The question is how to authenticate the news and articles which are circulated among social media like WhatsApp groups, Facebook Pages, Twitter, and other micro blogs & social networking sites. It is harmful for the society to believe on the rumors and pretend to be a news. The need of an hour is to stop the rumors and focus on the correct, authenticated news articles. Aim of this project is to develop two models for detecting fake news by using Machine Learning algorithms (usingSVM) and Deep Learning algorithms (using LST) respectively.

With the help of Machine learning and Deep Learning, it is tried to aggregate the news and later determine whether the news is real or fake using Support Vector Machine and Long Short-Term model. For the model, the dataset is cleaned and the data is pre-processed. Then on the pre-processed data, feature extraction techniques are used and the model is trained using both the algorithms separately to get two different models using SVM algorithm and LSTM algorithm respectively. Confusion Matrix, Classification reports and accuracies of both the models are calculated and compared to find the best model for Fake News Detection.

SVM Classifier gave 99.29% accuracy and LSTM classifier gave 99.54% accuracy. Both the algorithms gave good accuracies but LSTM gave better results compared to SVM in classifying the news articles. The major contribution of this project is to find the better fit algorithm and techniques for Fake News Detection between SVM and LSTM algorithms by comparing their accuracies.

Keywords: SVM, LSTM, Machine Learning, Deep Learning, Fake News.

I. INTRODUCTION

In modern times, people utilize social media for knowledge, entertainment, education and business purposes. With a lot of information or news, the one question occurred whether the given news or information is True or Fake. Fake news is commonly distributed with an intent to mislead or make an inclination to get political or monetary benefits.

A survey found that globally percentage of internet users in July 2020 stood at 59% of the global population. Fake news aims to mislead the news users' opinion. For a conspicuous example, COVID-19 pandemic quickly spreading around the world in this period and fake information related to this disease gets created in our society. Such misinformation caused anxiety among people and even deadly consequences in health problem.

Due to the exponential growth of information online, it is becoming impossible to decipher the true from the false. Thus, this leads to the problem of fake news. Detection of fake news online is important in today's society as fresh news content is rapidly being produced as a result of the abundance of available technology. That makes detecting fake news be a necessary task. To show what is true and what is false news.

II. METHODOLOGY

OVERVIEW OF A DATASET:

The data available for training the models consists of two CSV file datasets from Kaggle which are distinguished into real and fake news files. Dataset has 4 attributes which tells about the title, text, subject & date of news articles. Fake CSV file has fake news data and True CSV file has real news data.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Conference:01/December-2022 Impact Factor- 6.752 www.irjmets.com
1st National Conference on Applications of soft Computing Techniques in Engineering NCASCTE-2022
Organized by Department of Electrical & Electronics Engineering, ACE Engineering College, Hyderabad

Type of Articles	Total Count	Total Attributes
Real	21417	4
Fake	23481	4

DATA CLEANING & PRE-PROCESSING:

Data Cleaning is removing Noisy data with null values or missing values in a file. It is to be done before pre-processing the data. Data pre-processing phase is very important step in any data science lifecycle, as pre-processed data will make that the data is reliable and consistent in building machine learning models. Social media news is unstructured, big and noisy. Thus, these data need to be pre-processed. In data pre-processing step, we have first dropped duplicate rows and missing values from the dataset. After that, we have removed punctuations, numbers, alphanumeric texts, non-English words and stop words from the news articles. These values can cause the model over fitting the data and these are not added any real values to the actual news content.

FEATURE EXTRACTION:

Machine learning model can deal with numeric values only. Therefore, we need to transform 8 the text input data into numeric vector form for applying classification methods. The process of converting text data into numerical representation is called vectorization. There are many word vectorization techniques in text processing. Among them, our model uses frequencybased method (TF-IDF) and pre-trained word embedding methods.

TF-IDF:

TF-IDF is sparse vector representation and it evaluates how relevant a word in a collection of documents or corpus. TF-IDF techniques use all the tokens in the sentence as vocabulary. TF means term frequency which can use to measure how frequently a term in one document. IDF measures the important of a term in document. The value of TF-IDF weight can be obtained as a way of normalization in Equation (1), where TF is term frequency, d is a document, IDF is inverse document frequency and D is corpus or documents in the dataset.

WORD EMBEDDING:

Word embedding is a learned representation for text where words that have the same meaning have a similar representation. It helps transform raw data (characters in text documents) to a meaningful alignment of word vectors in the embedding space that the model can work with more effectively. It is this approach to representing words and documents that may be considered one of the key breakthroughs of deep learning on challenging natural language processing problems.

III. MODELING AND ANALYSIS

SVM MODEL:A support vector machine (SVM) is a supervised learning algorithm. SVMs work by being trained with specific data already organized into two different categories. Hence, the model is constructed after it has already been trained. Furthermore, the goal 9 of the SVM method is to distinguish which category any new data falls under, in addition, it must also maximize the margin between the two classes. The optimal goal is that the SVM will find a hyperplane that divides the dataset into two groups.

The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is.

C is a hypermeter which is set before the training model and used to control error and Gamma is also a hypermeter which is set before the training model and used to give curvature weight of the decision boundary. Intuitively, the gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. The gamma parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors.

Kernel Function is a method used to take data as input and transform it into the required form of processing data. The kernel used in this application is Linear Kernel which is used when the data is Linearly separable, that

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Conference:01/December-2022 Impact Factor- 6.752 www.irjmets.com
1st National Conference on Applications of soft Computing Techniques in Engineering NCASCTE-2022
Organized by Department of Electrical & Electronics Engineering, ACE Engineering College, Hyderabad

is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a large number of Features in a particular Data Set.

LSTM MODEL: It is a special type of recurrent neural network that process the sequential data and time series data. It is more suitable for text classification rather than machine learning algorithms because LSTM can capture long-term dependencies between word sequences and it composed of three gates. A cell state helps the data to flow through the cell units without being changed by allowing only a few linear interactions. The input gate in LSTM controls the new values flow into the cells, the forget gate control the extent to which a value remains in the cell and the output gate use to compute the activation output of the LSTM cell. In the proposed model, we applied sigmoid activation function and Adam optimization function to classify news.

TESTING THE CLASSIFICATION RESULTS: Trained models are tested and validated using different approaches. The results are predicted and accuracy of the models are calculated for the classification of news given to the system.

Model and Material which are used is presented in this section. Table and model should be in prescribed format.

IV. RESULTS AND DISCUSSION

CONFUSION MATRIX:

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Fig: Confusion Matrix

CLASSIFICATION REPORT:

It is one of the performance evaluation metrics of a classification-based machine learning model. It displays your model's precision, recall, F1 score and support. It provides a better understanding of the overall performance of our trained model.

PRECISION:

Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class, it is defined as the ratio of true positives to the sum of a true positive and false positive.

Precision: - Accuracy of positive predictions.

Precision = TP/ (TP + FP)

RECALL:

Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

Recall: - Fraction of positives that were correctly identified.

Recall = TP/ (TP + FN)

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Conference:01/December-2022 Impact Factor- 6.752 www.irjmets.com
1st National Conference on Applications of soft Computing Techniques in Engineering NCASCTE-2022
Organized by Department of Electrical & Electronics Engineering, ACE Engineering College, Hyderabad

F1 SCORE:

The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

F1 Score = 2*(Recall * Precision)/ (Recall + Precision)

SUPPORT:

Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

PERFORMANCE ANALYSIS:

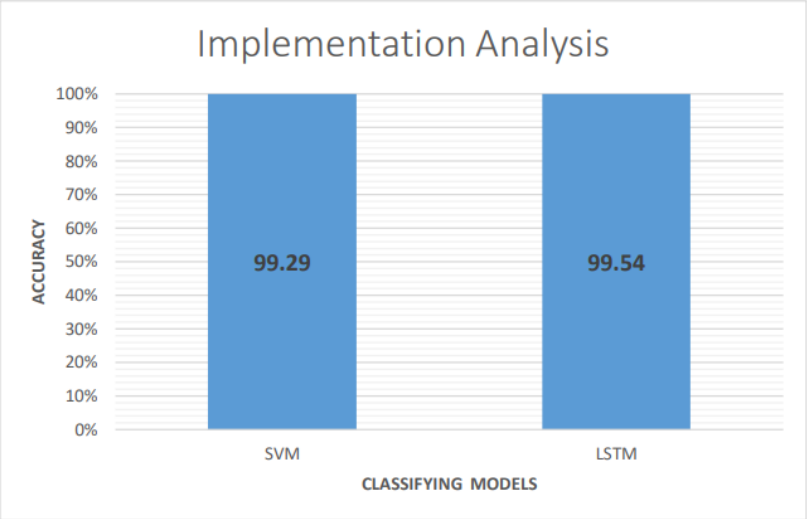

Fig: Implementation Analysis

RESULTS:

Table: Confusion Matrix for SVM Classifier

Test Cases	Actual Class	Predicted Class	
		Positive	**Negative**
Train = 80 Test = 20	**Positive**	4640	42
	Negative	21	4277
Train = 70 Test = 30	**Positive**	6998	55
	Negative	28	6389
Train = 60 Test = 40	**Positive**	9365	73
	Negative	45	8477
Train = 50 Test = 50	**Positive**	11678	101
	Negative	65	10605

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Conference:01/December-2022 Impact Factor- 6.752 www.irjmets.com
1st National Conference on Applications of soft Computing Techniques in Engineering NCASCTE-2022
Organized by Department of Electrical & Electronics Engineering, ACE Engineering College, Hyderabad

Table: Confusio Matrix for LSTM Classifier

Test Cases	Actual Class	Predicted Class	
		Positive	**Negative**
Train = 80 Test = 20	**Positive**	5912	37
	Negative	51	5225
Train = 70 Test = 30	**Positive**	5918	31
	Negative	23	5253
Train = 60 Test = 40	**Positive**	5915	34
	Negative	33	5243
Train = 50 Test = 50	**Positive**	5919	30
	Negative	26	5250

Table: Accuracy & Classification Report of SVM Algorithm

Test Cases	SVM Model				
Train = 80 Test= 20	**Accuracy:** 99.2984409799556				
		Precision	Recall	f1-Score	support
	Fake	1.00	0.99	0.99	4682
	Real	0.99	1.00	0.99	4298
	Avg	0.99	0.99	0.99	8980
Train = 70 Test= 30	**Accuracy:** 99.38381588715664				
		Precision	Recall	f1-Score	support
	Fake	1.00	0.99	0.99	7053
	Real	0.99	1.00	0.99	6417
	Avg	0.99	0.99	0.99	13470
Train = 60 Test= 40	**Accuracy:** 99.34298440979956				
		Precision	Recall	f1-Score	support
	Fake	1.00	0.99	0.99	9438
	Real	0.99	0.99	0.99	8522
	Avg	0.99	0.99	0.99	17960
Train = 50 Test= 50	**Accuracy:** 99.26054612677625				
		Precision	Recall	f1-Score	support
	Fake	0.99	0.99	0.99	11779
	Real	0.99	0.99	0.99	10670
	Avg	0.99	0.99	0.99	22449

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Conference:01/December-2022 Impact Factor- 6.752 www.irjmets.com
1st National Conference on Applications of soft Computing Techniques in Engineering NCASCTE-2022
Organized by Department of Electrical & Electronics Engineering, ACE Engineering College, Hyderabad

Table: Accuracy & Classification Report of LSTM algorithm

Test Cases	LSTM Model				
Train = 80 Test= 20	**Accuracy:** 99.3103563474387				
		Precision	Recall	f1-Score	support
	Fake	0.99	0.99	0.99	5949
	Real	0.99	0.99	0.99	5276
	Avg	0.99	0.99	0.99	11225
Train = 70 Test= 30	**Accuracy :** 99.51893095768374				
		Precision	Recall	f1-Score	support
	Fake	1.00	0.99	1.00	5949
	Real	0.99	1.00	0.99	5276
	Avg	1.00	1.00	1.00	11225
Train = 60 Test = 40	**Accuracy:** 99.50311804008909				
		Precision	Recall	f1-Score	support
	Fake	0.99	0.99	0.99	5949
	Real	0.99	0.99	0.99	5276
	Avg	0.99	0.99	0.99	11225
Train = 50 Test = 50	**Accuracy :** 99.5011135857461				
		Precision	Recall	f1-Score	support
	Fake	1.00	0.99	1.00	5949
	Real	0.99	1.00	0.99	5276
	Avg	1.00	1.00	1.00	11225

V. CONCLUSION

Our proposed system classifies the given article into real or fake news based upon the information given in that article using SVM approach (ML) and LSTM approach (DL). This helps people to know if the information present in various social media sites and on internet is real or misleading news. Accuracy of SVM classifier is 99.29 % and accuracy of LSTM classifier is 99.54%. Both the algorithms give good accuracy but LSTM gives better results compared to SVM in classifying the news articles. From our comparative study, we conclude that the LSTM model gave more accuracy than the SVM model for different test cases. The classification reports of LSTM model gave better precision values than SVM model. Classifying the given articles into fake or real news by LSTM model gave accurate results than SVM model. Thus, LSTM model is better for Fake news detection than SVM model.

ACKNOWLEDGEMENT

We would like to thank our guide **Dr. Khaleel Ur Rahman Khan** for his continuous support and guidance. Also, we are thankful to our project coordinator **Mr. K. Prem Kumar** &**Mr. K. Lal Bhahudhur**and we are extremely grateful to **Dr. M. V. Vijaya Saradhi**, Head of the Department of Computer Science and Engineering, ACE Engineering College for his support and invaluable time.

VI. REFERENCES

[1] Sobhani, Parinaz, Saif Mohammad, & Svetlana Kiritchenko. (2016). "Detecting stance in tweets and analyzing its interaction with sentiment." Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics.

[2] Parinaz, &Sobhani. (2017). “Stance Detection and Analysis in social media” Ph.D. Dissertation. University d’ Ottawa.

[3] Davis, Richard, & Chris Proctor.ET.AL. "Fake News, Real Consequences: Recruiting Neural Networks for

the Fight Against Fake News."

[4] Alim Al Ahmed, Ayman Aljabouh, Praveen Kumar Donepudi, & Myung Suh Choi. (2021 February 8). "Fake News Detected through Machine Learning Classifiers by using KNN model", site as arXiv:2102.04458[cs.CY].

[5] K. Shu, A. Sliva, S., Wang, J., Tang & H., Liu. (2017 June). "Fake News Detection on Social Media: A Data Mining Perspective," Newsletter ACM New York. https://doi.org/10.1145/3137597.3137600.

[6] Rubin, V.L., N.J., Conroy, & Y., Chen. (2016 June). "Fake news or truth Using satirical cues to detect potentially misleading news in proceedings of the second workshop on computational approaches to deception detection." proceeding of NAACL-HLT.

[7] Krejzl, & Steinberger. (2016). "Focused on using domain knowledge related features", a stance detection task on tweets. In SemEval- Task 6.

[8] Fatima Nada, Bariya Firdous Khan, Aroofa Maryam, Nooruz-Zuha&Zameer Ahmed. (2019 May 5). "Fake news detection using logistic Regression" International research journal of engineering and technology (IRJET) volume:06 ISSUE.

[9] Brambrick, Aylien(April 20, 2019). "Support Vector Machines: A Simple Explanation" https://www.kdnuggets.com /2016/07/support-vector-machines-simple-explanation.html.

[10] Hochreiter, S., Jrgen, S (October 1997), "Long short-term memory", Institute of Bioinformatics Publications Newsletter.

[11] Vasu Agarwal, H., Parveen sultana, Srijan Malhotra &Amitrajit Sarkar. (2019) "Analysis of classifiers for fake news Detection", "International conference on report trends in advanced computing" (ICRTAC 2019):165(2019)377-383.

[12] Ben Roshan-Kaggle expert ranker. (December 2020) "Fake news classifier on US Election news| LSTM", Analytics Vidhya -guest_blog_2020.

[13] Ajay Kulkarni, Feras A., Batarseh. (2020) " Confusion Matrix on Binary Classification" 27 in Data Democracy,.2020.