Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 05-2022
DOI: 10.17605/OSF.IO/5RUH2

PERFORMANCE ANALYSIS OF FAKE SOCIAL MEDIACONTENT BASED ON DEEP LEARNING METHODS

[1]FAYYAZ ALI, [2] IRFAN ALI KANDHRO, [3] SYED ADNAN ALI ZAIDI, [4] ANWAR ALI SANJRANI, [5] AZAM KHAN and [6] SYEDA NAZIA ASHRAF

[1]Department of Software Engineering, Sir Syed University of Engineering and Technology, Karachi Sindh, Pakistan

[2,6]Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan

[3]Department of Computer Science, Muhammad Ali Jinnah University, Karachi, Pakistan

[4,5]Department Computer Science &Information Technology, University of Balochistan, Pakistan

ABSTRACT - Nowadays, social media platform has enhanced tremendously in the last few years and considered one of major source of for information seeking and sharing in low cost. Due to the growth of data, The Fake news are quickly dominating and spreading the information, and distorting the community for sharing their own thoughts, knowledge and point of view regarding towards to any topic. In this paper, structural features with updated RNN and LSTM methods are proposed to improve the performance of system on fake news data. The system uses attention layer with RNN and LSTM to update the weights and values of different features. The performance of model also compared with various hyper parameters such activation, optimization, and dropout. The Proposed Model based with Long Short Term Memory (LSTM) categorize the features closer on original and fake news with customized hyper parameters and random search. The experimental results also depicted that deep learning methods outperformed when size of data samples is high. Furthermore, we showed that combining strong feature engineering with deep learning models, we can more concisely identify the fake news with state-of-art results.

Keywords: Deep Learning; Fake News; Classification, performance analysis.

1. INTRODUCTION

In these days, there are many news organizations and agencies and different other social media applications that spread the news and daily updates globally. They play a virtual role to publish news globally. There are many people using social media through the internet. With interest being more accessible to people around the globe these platforms saw a rapid increase in users. There are several issues that can use different social media apps like Twitter, Facebook. Facebook is one of the best platforms to spread the news after the success of Facebook, many new social media platforms came into existence to publish and spread the news globally or publicly. There are many recognized agencies and that publish and spread the news, but they are not true or reliable. There is several false news that create a negative effect on our society. Few media and other different platforms may deliberately spread false news to gain the popularity or earn money. The uses of social media that handle the different news portal, post these recent articles. And moreover, this false news that can cause the attention of researchers. It is very difficult to handle the large amount of data and unauthorized posts that create a negative impact on our society. Every day these social media sites handle the data and grade every single post. This fake news can have many names like rumor, misinformation, or hoax. Fake news is so massive, it creates the bad and negative impact on our society [1-3]. In election to attack the voter there are number of fake news spread against the opposition party. Many different fake news that defames the celebrities. Nowadays there is several fake news and rumors about covid-19(Corona Pandemic) that badly affects our society, much fake news has been spread to misguide the public and increase the spread of the virus by making the public more careless. Moreover, the different media platforms spread the fake news about the weather forecasting like cyclone, storms this fake news create the fear in people's heart. Moreover, fake news was spread across the globe with different intentions. There is different name of fake news all these sounds different but are different forms of counterfeit information. There are many platforms that spread and publish 1000's of fake news and it is very difficult for media sites to handle the large amount of data and counterfeit and fraudulent information. The solution to handle the wrong or unauthorized information fraudulent information or counterfeit information. Most of the people interact with social media daily and share the post and news activities that are not constitutionality and authorized which amplifies the effect of the fake news. The world moved through the internet or social media. There are several news articles being published and classifying fake news is becoming increasingly difficult to check constitutionality of the news manually and it is more costly day by day. Therefore, the use of deep fake checking information is essential [10]. Therefore, the computer researchers automate the process in which propose a model to automate the process to identify the fake news. There are number of algorithms that we used for our model to automate our process like SVM, Naive Bayes, KNN, Decision Trees, Recurrent Neural Network RNNs and LSTM SVM and Naive Bayes are classifiers that both are baseline methods. KNN and Decision Trees are classifying algorithm in which we divide our data into groups and select nearest points. RNNs usually face the problem of vanishing gradients and their capability of learning long data sequences which are solved by LSTM and now solving the NLP task problem we use word embedding. It is very essential factor to have considered while designing a model for NLP task problems. It is used in the proposed research for contextual and the using the model BERT [15-18]. BERT is the very capable for the solving the NLP task, BERT can contextualize word representation by executing the large amount of untitled text corpora. BERT has a wonderful nonlinear representation learning capability.[3-9, 11] The long, short-term Memory (LSTM) adequately boost performance by learning, and d find out the key

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 05-2022
DOI: 10.17605/OSF.IO/5RUH2

information. BERT and LSTM is used to rise or improve the fake news classification because they both are very powerful for using contextual words representation and have ability to capture meanings and long-distance sequence in news title. BERT and LSTM is combining for improve the fake news classification either the fake or rumor, unauthorized, hoax etc. The research steps have been summarized as follows: The planned methodology define the lexical morphological planning, such as syntactical, grammatical, and meaning of words or semantical aspects of the news reports and articles. The approach for fake news classification is by combining BERT with an LSTM which classifies news articles. Accuracy, precision, Recall, and F1 score have been used as the interpretation criteria for testing the working of the proposed or planned methodology. In the proposed methodology empirical evaluation has been conducted with the state-of-the-art methodologies. There are many training or testing phases that we are used. In the proposed or planned methodology to checking the robustness of the planned methodology, such as TCNNURG, LIWC, CSI, HAN, SAFE etc [12-14].

The research paper is organized as follows: Section 2; Explain the literature Section 3; In section 3 briefly explain dataset information and architecture or framework of BERT and LSTM which is followed by the planned or proposed model. Section 4; In this characterize the detailed Result and Analysis. In the proposed methodology performance has been compared and analyzed the state-of-the-art methods and these methods implement or followed in the conclusion section

2. LITERATURE REVIEW

Social media has a huge number of active users and is utilized for news sharing. The news propagating on social media contains more disinformation than traditional news organizations because there is no authority to control the material. Many studies have lately been undertaken to identify bogus news on social media to curb the spread of disinformation. The following section reviews some of the most prominent studies on misinformation on social media.

To comprehend and detect false news via social media, Shu et al. [19] presented the FakeNewsTracker system. To investigate false news, news material and social media data were collected. In this study, the auto encode approach was utilized to learn the characteristics of news material. The feature learning approach was used to identify false news with the LSTM-RNN. The created model's performance was evaluated using the PolitiFact and BuzzFeed datasets. The experimental results suggest that the created model performs well in detecting false news. It also suggests that the feature selection strategy improves detection performance.

Boididou et al. [20] created a model that analyses tweet textual information to evaluate if the tweet is authentic or fraudulent. The semi-supervised learning strategies based on two independent credibility classifiers are used in this method. The method's performance was validated using the MediaEval benchmark datasets from 2015 and

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 05-2022
DOI: 10.17605/OSF.IO/5RUH2

2016. The new approach was compared to existing methods, and the findings demonstrated that the developed method performed better. This approach predicts false news through social media using random forests. According to the experimental results, the created technique performed admirably in detecting false news. The LSTM classifier may be used to boost efficiency because of its ability to maintain memory for an extended period.

The advantages of using transformer-based approaches rather than classification algorithms or machine methods were highlighted by using a bidirectional(two-way) transformer approach with a feed-forward (predictive) learning algorithm. In comparison to XGBoost, (Aggarwal et al. [21]) the authors stated that the detection performance of fine-tuned BERT was 97.02 percent and the accuracy of the NewsFN dataset is 89.37 percent. In contrast investigation of deep learning algorithms for detecting false information using the COVID-19 false news dataset, researchers discovered that a fine-tuned BERT outperformed different models in which BiLSTM is also included. The study also demonstrates the need for pre-training on domains of interest like corpus. It was determined that the transformer models outperformed the word-based models and non-transformers (Wani et al.[22]).

3. PROPOSED METHODOLOGY

This section shows the detail of the proposed model. And point out the details of RNN, LSTM and other processing with structural features.

The steps or background of RNN, LSTM, dataset and structural features and preprocessing methods have been explained in section. Fig 1 showed the step-by-step process of proposed framework for content based classification.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 05-2022
DOI: 10.17605/OSF.IO/5RUH2

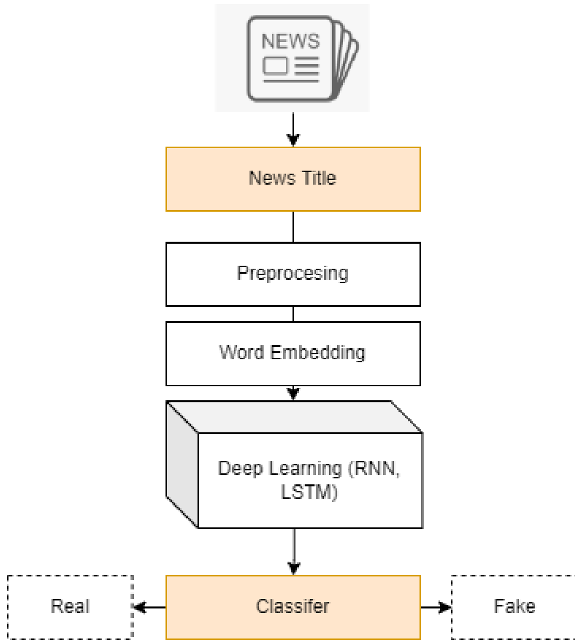

Figure 01: System Architecture

Data preparation is the first step of text preprocessing, which is used for most natural language processing tasks, where its reform text data in a sequence of manner. In this complete process of execution, it performed the following preprocessing steps in the preprocessor function defined below: It removes unwanted characters/symbols such as punctuation, HTML tags, and emoticons using automata-based regular expressions.

Further, it discards stops words (by means of the words that are most common in the English language). The concept of lemmatization is the process of reducing a word to its lemma (text-enabled) or dictionary form. For instance, the word run is the lemma for the words analysis, such as run, ran, and running.

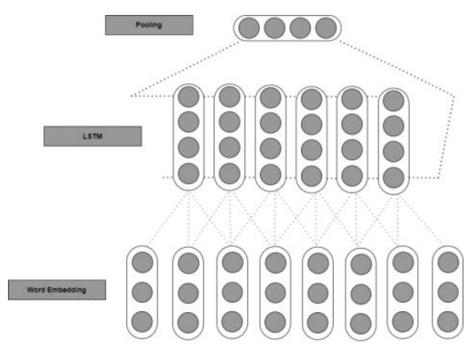

Figure 02: Steps of LSTM

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 05-2022
DOI: 10.17605/OSF.IO/5RUH2

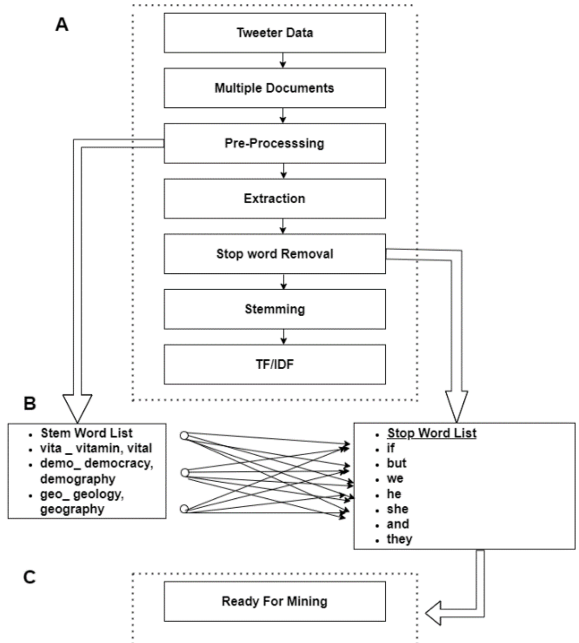

Figure 03: Preprocessing Steps

The Long Short-Term Memory is also known as LSTM neural network is units form the recurrent part of the recurrent convolutional neural network. LSTM is often used for tasks involving sequence data such as time series forecasting and text classification. It would not dive deeply into the mathematical background behind LSTMs because that topic is out of the scope of this article, but essentially an LSTM is a unit in a neural network capable of remembering essential information for long periods of time and forgetting information when it is no longer relevant (hence the name, long short-term memory). An LSTM unit consists of three gates:

LSTM: Our RNN layer will immediately follow the input and must occur before any dense or pooling layers. This is where the magic happens for a recurrent neural network (RNN). Traditional RNNs have lost popularity in favor of more robust versions like LSTM (Long Short-term Memory) and GRU (Gated Recurrent Unit) layers. However, what makes them special remains the same. An RNN has a memory and the ability to remember and/or forget. RNNs can temporarily store word weights based on their importance and relative location to other words in the text. This also makes them great for NLP as it lets your model pick up on nuances of structured language like proximity and context using relative positions within the text, and by remembering important words for later in the text. We follow our layer with a 1D pooling layer to simplify the output. Dense and Dropout Layers: We add some standard dense layers to our network. We also add dropout layers with the intention of improving generalization. The

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 05-2022
DOI: 10.17605/OSF.IO/5RUH2

number of layers and nodes you use are adjustable if you want to tune your model further as shown fig 2.

3 RESULT AND DISCUSSION

The experimental results shows that fake news classification is a challenging due to limitation of data such as (Nosy and unstructured format). This research study focuses on structural features with RNN and LSTM to detect fake news. it involves utilizing the preprocessing method to provide the proper structure of data. Moreover, structural features are extracted from social media such as twitter and Facebook.

In this experiment, we investigated impact of regularization and batch size on training and testing dynamics. Further, we used accuracy metrics to identify the generalization gap between train and test time and investigate epochs and batch size in context for fake and real classification. These experiments we meant to provide intuition on the effects batch size and epochs as shown in Table 1.

Table 1. RNN and LSTM Results with Epochs and Batch Size

Table 2. Testing and Training Accuracy RNN and LSTM

Type of News		RNN(Training)	RNN(Testing)	LSTM(Training)	LSTM(Testing)
Fake News	Accuracy	72.0	72.0	75.0	72.0
	Precision	72.0	72.0	72.0	72.0
	Recall	74.0	72.0	72.0	74.0
	F1 Score	73.0	76.0	72.0	72.0
Real News	Accuracy	72.0	71.0	72.0	74.0
	Precision	72.0	72.0	72.0	72.0
	Recall	70.0	73.0	74.0	73.0
	F1 Score	72.0	74.0	71.0	72.0

Table 01 shows the training and testing accuracy of model with of respect epochs and batch size (5,10,15,20,25) to analyze the fake news with structural features. The LSTM

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 05-2022
DOI: 10.17605/OSF.IO/5RUH2

network showed state of art results on (epochs 30 and batch size 256) as compared to RNN. The RNN and LSTM consider the irrelevant features for the classification. The LSTM method analysis the core information of twitter and applies. The fake news classification based on structural features. The structural features in dataset such as URL, hashtags, retweet, and text length incases the performance of fake news classification.The model-based LSTM archives an accuracy 89% and RNN method shows 87% of accuracy as shown fig 4 and 5.

For effortlessness, we will generally talk about things as far as a parallel grouping issue where suppose we'll need to find assuming a picture is of a feline or a canine. Or then a patient is having the disease (positive) or is seen as sound (negative). Typical terms to be clear with are:

There are four (4) main parts that include the numerical recipe for ascertaining Exactness, viz. TP, TN, FP, FN, and these parts award us the capacity to investigate other ML Model Assessment Measurements. The recipe for computing precision is as per the following:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

Where:

TP addresses the number of genuine upsides. This alludes to all out number of perceptions that have a place with the positive class and have been anticipated accurately. TN addresses the number of Genuine Negatives. This is the all-out number of perceptions that have a place with the negative class and have been anticipated accurately. FP is the quantity of Misleading Up-sides. It is otherwise called a Sort 1 Mistake. This is the absolute number of perceptions that have been anticipated to have a place with the positive class; however, rather has a place with the negative class. FN is the number of Bogus Negatives. It very well might be alluded to as a Sort 2 Blunder. This is the absolute number of perceptions that have been anticipated to be a piece of the negative class yet rather have a place with the positive class.The reason for people to use the Precision Assessment Metric is for usability. This Assessment Metric has a basic methodology and clarification. It is, as examined previously, just the all-out extent (absolute number) of perceptions that have been anticipated accurately. Precision, in any case, is an Assessment Metric that doesn't perform well when the presence of imbalanced classes-when within the sight of imbalanced classes, Exactness experiences an oddity, i.e., where the Precision esteem is high however the model needs prescient power and the vast majority, expectations will be mistaken. For the above reason, when we can't utilize the Exactness Assessment Metric, we are constrained to go to other assessment measurements in the scikit-learn stockpile. These incorporate, however, are not restricted to, the accompanying Assessment Measurements:

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 05-2022
DOI: 10.17605/OSF.IO/5RUH2

This alludes to the extent (complete number) of all perceptions that have been anticipated to have a place with the positive class and is positive. The recipe for Accuracy Assessment Metric is as per the following:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

This is the extent of perception anticipated to have a place with the positive class that genuinely has a place with the positive class. It is a roundabout way that lets us know the model's capacity to haphazardly recognize a perception that has a place with the positive class. The equation for Review is as per the following:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

This is an averaging Assessment Metric that is utilized to produce a proportion. The F1 Score is otherwise called the Consonant Mean of the accuracy and review Assessment Measurements. This Assessment Metric is a proportion of in general rightness that our model has accomplished in a positive expectation of all perceptions that our model has named as sure, the number of these perceptions are positive. The recipe for the F1 Score Assessment Metric is as per the following:

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$TruePositiveRate(TPR) = RECALL = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$FalsePositiveRate(FPR) = 1 - Specificity = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$$

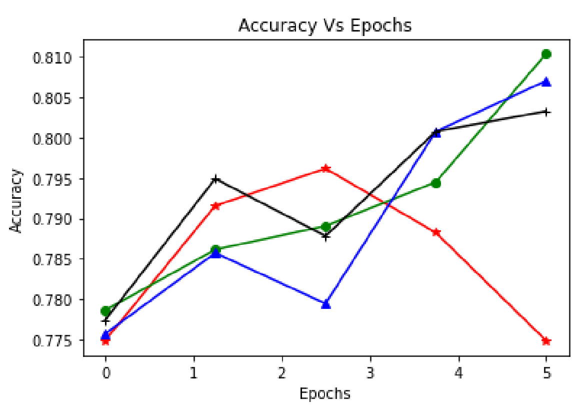

Figure 04: Steps of LSTM

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 05-2022
DOI: 10.17605/OSF.IO/5RUH2

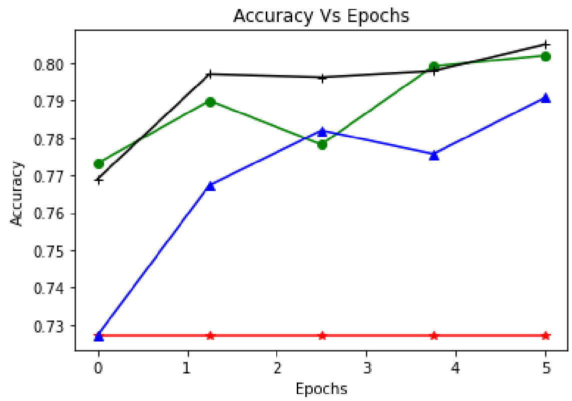

Figure 05: Steps of LSTM

4 CONCLUSION

In this paper, we proposed system based on Deep learning to classify the real and fake news with respect of features. Recently, with rapid growth social platform and used of internet, the users are unaware to verify the news authenticity. The news printed on social platforms leads to deformity due to the absences of filters and verification tools to check the originality of news. Many of existing studies and methods are come out to analyze the news originality on social media. The analysis shows the model with LSTM delivers higher performance as compared to other DL methods, the news can be detected accurately on fake and real news data samples with combination of feature engineering with LSTM method. Moreover, many of researchers has developed systems for identify the fake news on social world to handle fake news in its early stages. In future, we will extend our model with more linguistic features that enable to identify newly generated fake ns effectively and determine authoritative information and sources automatically.

REFERENCES

1. Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31(2): 211–236
2. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Eugene Stanley H, Quattrociocchi W (2016) The spreading of misinformation online. Proc Natl Acad Sci 113(3):554–559
3. Qawasmeh E, Tawalbeh M, Abdullah M (2019) Automatic Identification of Fake News Using Deep Learning. 2019 Sixth international conference on social networks analysis, Management and Security (SNAMS), pp. 383–388,https://doi.org/10.1109/SNAMS.2019.8931873
4. Castillo C, Mendoza M, Poblete B (2011) Information credibility on twitter. In: Proceedings of the 20th international conference on world wide web, pp 675–684
5. Berinsky A (2017) Rumours and health care reform: experiments in political misinformation. Br J Polit Sci 47(2):241–262

ISSN：1671-5497
E-Publication：Online Open Access
Vol: 41 Issue: 05-2022
DOI: 10.17605/OSF.IO/5RUH2

6. Vosoughi S, Neo Mohsenvand M, Roy D (2017) Rumour gauge: predicting the veracity of rumours on twitter. ACM Trans KnowlDiscov Data 11(4):1–36
7. Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31(2): 211–236
8. Lazer D, Baum M, Benkler Y, Berinsky A, Greenhill K, Menczer F, Metzger M, Nyhan B, Pennycook G, Rothschild D, Schudson M, Sloman S, Sunstein C, Thorson E, Watts D, Zittrain J (2018) The science of fake news. Science." Published to Science 09 Mar 2018 359(6380):1094–1096
9. Mo Jang S, Kim JK (2018) Third person effects of fake news: Fake news regulation and media literacy interventions. Comput Hum Behav 80:295–302. https://doi.org/10.1016/j.chb.2017.11.034
10. Prasetijo, Agung B., et al. "Hoax detection system on Indonesian news sites based on text classification using SVM and SGD." 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE). IEEE, 2017.
11. Granik, Mykhailo, and Volodymyr Mesyura. "Fake news detection using naive Bayes classifier." 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON). IEEE, 2017.
12. Goyal, Lalit Mohan, Mamta Mittal, and Jasleen Kaur Sethi. "Fuzzy model generation using Subtractive and Fuzzy C-Means clustering." CSI transactions on ICT 4.2 (2016): 129-133.
13. Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney. "LSTM neural networks for language modeling." Thirteenth annual conference of the international speech communication association. 2012.
14. Mikolov, Tomas, et al. "Advances in pre-training distributed word representations." arXiv preprint arXiv:1712.09405 (2017).
15. Kandhro, Irfan Ali, et al. "Classification of Sindhi Headline News Documents based on TF-IDF Text Analysis Scheme." Indian Journal of Science and Technology 12.33 (2019): 1-10.
16. Kandhro, Irfan Ali, et al. "Roman Urdu headline news text classification using RNN, LSTM and CNN." Advances in Data Science and Adaptive Analysis 12.02 (2020): 2050008.
17. Kandhro, irfanali, et al. "psat-based sentiment analysis: for text and data mining."
18. Kandhro, Irfan Ali, et al. "Student feedback sentiment analysis model using various machine learning schemes: a review." Indian Journal of Science and Technology 12.14 (2019): 1-9.
19. Shu K, Mahudeswaran D, Liu H (2019) FakeNewsTracker: a tool for fake news collection, detection, and visualization. Comput Math Organ Theory 25:60–71. https://doi.org/10.1007/s10588-018-09280-3
20. Boididou C, Middleton SE, Jin Z, Papadopoulos S, Dang-Nguyen DT, Boato G, Kompatsiaris Y (2018) Verifying information with multimedia content on twitter. Multimed Tools Appl 77:15545–15571. https:// doi.org/10.1007/s11042-017-5132-9
21. Wani, Apurva, et al. "Evaluating deep learning approaches for covid19 fake news detection." International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation. Springer, Cham, 2021.
22. Aggarwal, Akshay, et al. "Classification of fake news by fine-tuning deep bidirectional transformers based language model." EAI Endorsed Transactions on Scalable Information Systems 7.27 (2020).