Fake News Detection in Social Networks Using Machine Learning Techniques

Ammar Saeed [1, *] and Eesa Al Solami[2]
ammarsaeed1997@gmail.com eaalsulami@uj.edu.sa
[1]Department of Computer Science, COMSATS University Islamabad, Wah Campus, Wah Cantt, Pakistan
[2]Department of Cybersecurity, College of Computer Science and Engineering, University of Jeddah, KSA

Summary
The usage of social media platforms has excessively grown over the past decade due to their ease of excess, evolution, and accessibility. This increased usage has proved to be quite advantageous in terms of the benefits of staying connected, sharing posts and ideas, and exchanging thoughts, but it also has its fair share of drawbacks. These drawbacks arise mainly due to a lack of social media usage knowledge among consumers, their problem in understanding and comprehending non-native languages, and false efforts to obtain a response from other people within their community. As a result, of which, fake news, which has no real validity, starts accumulating over time and begins to appear in the feed of every social media consumer causing ambiguity and uncertainty. To sustain the integrity of social media platforms, such news and content must be distinguished from the real one. Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have accelerated the creation of autonomous systems capable of achieving any desired result in a minimal period. This research proposes a novel approach for detecting fake news. Fake news dataset acquired from online sources is first preprocessed, textual features are extracted based on N-gram methods such as Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BOW). Latent Dirichlet Allocation (LDA) based topic modeling is also employed on data compilation to derive dominant topics from it which are scaled later on. Finally, the textual features and topic vectors are assessed on standalone ML classifiers Support Vector Machine (SVM), Logistic Regression (LR), and Naïve Bayes (NB) and ensemble ML classifiers Random Forest (RF) and Gradient Boost (GB) where results are evaluated based on several performance metrics.

Keywords:
Machine Learning; Fake News; Social Media; Fake News Classification

1. Introduction

With the ever-growing advancements and breakthroughs in technology, the Internet of things (IoT), and with the introduction of the latest social media platforms on daily basis, the world has turned into a global village which has led to effortless accessibility of the internet to everyone. Furthermore, the widespread availability of low-cost, internet-enabled smartphones and other electronic devices has enabled people from varied educational, regional, and cultural backgrounds to share their expressions on social media platforms. Where this ease of access to community platforms has several advantages including the ability to investigate relevant news from across the world, stay updated with events occurring across the globe, and exchange thoughts about them, but it also has its fair share of drawbacks. The news circulating on social platforms is more likely to reach different corners of the world because practically everyone carries internet-enabled devices with them and prefers to read posts and news circulating on online forums rather than reading a newspaper, watching television, listening to the radio, or using any other conventional source. Aside from that, sharing news, facts, and ideas on community platforms is simple, painless, inexpensive, and quick, and therefore social media consumers feel free to do it [1]. The issue arises when certain sources propagate incorrect or misleading news on social media platforms owing to a lack of understanding required for ethical and proper usage of community platforms, mental immaturity, and public attention-seeking attitude. By tweaking the original content and reshaping the concerning news article as per their liking, fake news is generated by an individual or a whole community indulged in such purposeful activities. As previously mentioned, such news when published on online forums, reach different corners of the world where they get propagated and the residents of that region come across them. As a result, they continue to spread and accumulate and go trending on social platforms depending upon the location. Another reason behind the trend of such fake news is the unavailability of any automated review systems and processes on such sites that can analyze news and distinguish between true and fake ones among them [2].

The overspread of such content on social media has a deplorable influence on individuals, societies, political and economic activities of a country or region. It has the potential to influence an individual's strategy and decision-making abilities, as well as impact the perception of someone toward someone else. Certain groups, such as FlackCheck, OpenSecrets, Fact Check, and Snopes have tried to cope with this ever-increasing volume of fake news but there still is massive room for improvement in fake news detection models and methods in terms of accuracy and precision. The reason for such non-reliability and accuracy is that the aforementioned sites utilize some real-

Manuscript received April 5, 2022
Manuscript revised April 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.4.92

world reporters to check the veracity of multiple news headlines, which is a reasonable but time-consuming and tiresome method. This promotes an urge to develop an automated system that can verify the legitimacy of news articles on social media platforms and differentiate false news from true news. The recent advancements and breakthroughs in AI and ML have enabled researchers to employ their applications in challenging detection, analysis, and classification problems [3]. ML-based models these days are used to create small to large-scale automated systems that can be trained on a large amount of data and then execute real-time classification tasks with extreme accuracy and precision on their own without needing any manual intervention.

A hybrid approach is proposed in this paper where a dataset comprising fake and real news is acquired from Kaggle. Certain preprocessing steps that include stop words removal, punctuation removal, lemmatization, and tokenization are then employed on the concerned data to clean it and make it ready for the upcoming procedures. The textual features are then extracted from the cleansed dataset using N-gram methods TF-IDF and BOW while at the same time, key topics are derived from the dataset using LDA. The topics are then scaled using a standard scaler and they along with extracted textual features are provided to several standalone ML algorithms SVM, LR, NB, and ensemble ML algorithms for results derivation. Finally, the derived results are evaluated based on certain performance measures.

3. Related Work

Wani et al. [4] proposed a solution to the problem of the spread of fake news during pandemic covid19 by designing an automated solution for the classification of fake news using deep learning (DL) algorithms CNN and LSTM and a transformer-based algorithm Bidirectional Encoders Transformer Networks (BERT) on the covid 19 fake news dataset. They also evaluated the importance of unsupervised algorithms on the unlabeled covid 19. The best accuracy of about 94% was achieved by them while the classification of fake news. Alonso et al. [5] take fake news detection into account by conducting sentiment analysis on the fake news dataset taken from social media. They mentioned the fake news datasets with their complete description and source to date. They discussed multiple aspects of fake news detection using sentiment analysis as well as traditional ML and DL techniques. Jain et al. [6] proposed a fake news detection system using supervised ML algorithms SVM and NB. They applied a feature extraction method named Counter Vectorizer. They showed a concept and process for detecting bogus news. They attempted to aggregate the news using ML and natural language processing (NLP) techniques and then used an SVM classifier to assess if the news was authentic or fraudulent. The suggested model's findings were compared to current models and found to be up to 93.6 percent accurate. Jiang et al. [7] used to hold out cross-validation to compare the performance of five ML models and three DL models on two fake and real news datasets of various sizes. For ML and DL models, they employed TF, TF-IDF, and embedding approaches to acquire text representation. They utilized conventional assessment measures and a revised version of McNemar's test to assess the models' performance. On the ISOT and KDnugget datasets, the suggested stacking model obtained testing accuracy of 99.94 percent and 96.05 percent, respectively. Furthermore, when compared to baseline approaches, the performance of our suggested method is excellent, and it is thus recommended. Hakak et al. [8] introduced an ensemble classification model for detecting fake news that extracts essential characteristics from the used datasets and then classifies them using the proposed model that consists of three commonly used ML models DT, RF, ETC. On the Liar dataset, they obtained 99.8% training accuracy and 44.15 percent testing accuracy, respectively. They got 100% training and testing accuracy using the ISOT dataset.

To computerize faked news detection in Twitter datasets, Mahir et al. [9] created a strategy for spotting fabricated news messages from tweets by working out how to predict precision assessments. They then compared the classification performance of five well-known ML methods, including SVM, NB, LR, and RNN models, on the dataset. The SVM and NB classifiers outperformed the other methods, according to their findings.

Experiments for false news identification [10] were carried out using a tree-based Ensemble ML framework with optimum parameters and context level information. Gradient descent algorithms were used to generate to maximize a single common objective function, and this formulation justifies important elements and parameters in the methods. Experiments were carried out with a multi-class dataset (FNC) and a variety of ML models for classification. When compared to current benchmark results, the experimental findings showed that the ensemble framework was successful. We achieved an accuracy of 86% for multi-class classification of false news using the GB-based framework. By identifying news transmission channels, Liu et al. [11] established a novel methodology for the early detection of bogus news on social media. They initially characterized each news story's propagation path as a multivariate time series, with each tuple representing a numerical vector indicating the attributes of a person who propagated the news. Then, to detect bogus news, they created a time series classifier that included both recurrent and convolutional networks that collected both global and local alterations in user attributes along the propagation path. Experiments on three real-world datasets revealed that the suggested model could detect bogus news with an accuracy of 85 percent and 92 percent on Twitter and Sina Weibo, respectively, in just 5 minutes after the news began to propagate, which is substantially quicker than current benchmarks.

Shahbazi et al. [12] suggested an integrated system for different elements of blockchain and NLP and used ML approaches to detect false news and better anticipate bogus user accounts and postings. This procedure was carried out using the Reinforcement Learning approach. The decentralized blockchain infrastructure, which offered the outline of digital content authority proof, was used to increase the platform's security. The goal of this technology, in particular, was to provide a safe platform for predicting and identifying bogus news on social media networks. The authors of the study [13] used the Urdu language to detect bogus news. For the false news identification tasks, they created an annotated corpus of Urdu news items. Second, they investigated three different ML algorithms for detecting bogus news. To calculate the base predictor's predictions and improve the overall performance of the false news detection system, five ensemble learning approaches were employed to build an ensemble system. The results of the experiment on two Urdu news corpora indicated that ensemble models outperformed individual ML models. Three performance criteria were used to determine that Ensemble Selection and Vote models outperformed the other ML and ensemble learning models: accuracy, area under the curve, and mean absolute error. Adiba et al. [14] used ML based NB classifier to identify false news from an open-source dataset. On the dataset, tokenization, stemming, and lemmatization were used as preprocessing processes. TF-IDF vectorization and Count vectorization are used to process the data. TF-IDF and BOW are used to extract features, while the NB classifier is used to classify the data. Initially, a classification accuracy of 87 percent was attained, which was greater than previously reported accuracy, and later the same NB method with enhanced corpora achieved 92 percent accuracy.

4. Proposed Methodology

The proposed fake news detection models are discussed in detail in this section. The proposed model is set up in the form of several blocks. A dataset comprising fake and real news is retrieved from online sources and passed through the preprocessing block where it is cleansed and prepared through the methods of useless information removal, lemmatization, and tokenization. In the next block, textual features are extracted from data using N-gram methods TF-IDF and BoW. Apart from this, the prominent topics are also derived based on topic modeling methods LDA. The derived features and topics are then provided to standalone ML algorithms SVM, LR, NB, and ensemble ML models for classification. Furthermore, several performance evaluation measures are also implemented for model evaluation. The proposed model architecture is shown in Fig. 1 while all steps are discussed in detail in the coming sections.

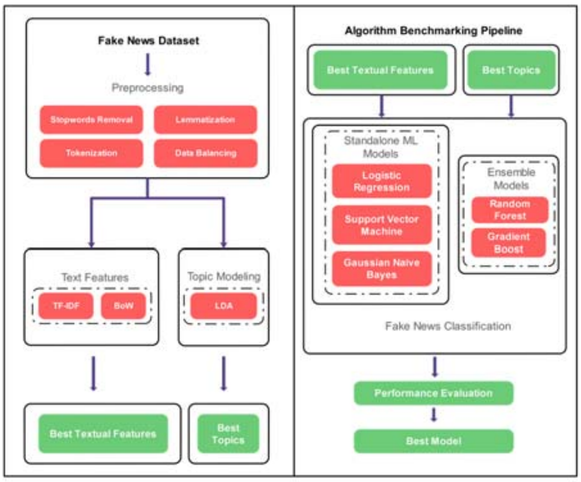

Fig. 1. Proposed fake news model framework

4.1 Data Acquisition

The Real and Fake news[1] dataset is utilized in this work that is prepared by nop.ai and it comprises a combination of fake and truth-based news gathered from the real-world based on real-world events. This dataset contains a total of 6335 instances divided into four columns or attributes such as index, title, text, and label. Among the dataset compilation, the fake news count is 3164 while the real news count is 3172 records along with other data as represented in Tab. 1. Each news record contains a label that indicates whether it is fake or real.

Tab. 1: Real and Fake news dataset statistics

Attribute	Value
Total News	6335
Real News	3171
Fake News	3164
Words Count	5365684
Characters Count	23842804

[1] https://www.kaggle.com/datasets/nopdev/real-and-fake-news-dataset

Sentence Count	6341

4.2 Preprocessing

The state and condition of the dataset are critical in an automated detection and classification methodology. Since the obtained dataset is from various online sources and social media platforms, it contains unnecessary words, punctuations, and other ambiguous information that may cause the proposed model that uses ML backend to slow down and not perform up to its potential. Therefore, the raw dataset needs to be cleansed and prepared before proceeding with it any further. Preprocessing and text mining approaches have become increasingly popular that can take up the raw data, process it, and make it ready for the model. In the proposed work, letter conversion to lowercase, stop words removal, hyperlinks removal, inappropriate full stop and half sentence removal, lemmatization, and tokenization are utilized as preprocessing methods.

4.2 Feature Extraction and Topic Modeling

After data preprocessing, the model still cannot classify the data in its ordinary textual form, thus it must be transformed into mathematical and vector format so that the ML algorithms can understand and classify it. Once the data is converted into vector format, it can easily be provided to classifiers for classification as per labels. In the proposed work, frequent textual features are extracted using N-gram approaches TF-IDF and BOW. The LDA-based topic modeling is also adopted to retrieve key topics from the preprocessed dataset. Both these procedures are discussed in detail for the proposed work in sections 4.2.1 and 4.2.2.

4.2.1. Textual Features Extraction

An N-gram is a collection of word tokens in a data set that is based on unigrams, bigrams, trigrams, and so on. In a data corpus, an n-gram model may compute and forecast the likelihood of certain word sequences. These models are useful for text classification problems in which the number of specific terms in the corpus vocabulary must be counted. The TF-IDF is a statistic for determining how well a word in a catalog matches its meaning or mood [15]. It works by increasing the frequency of keywords in a document by the inverse frequency of phrases that appear often in several texts. The mathematical formulation of TF-IDF is represented in Eq. (1).

$$wt_a, wt_b = f_{a,b}^{t} \, x \log\left(\frac{K}{f_a}\right) \quad (1)$$

Where, wt_a, wt_b represents the weights for any two data points a and b. $f_{a,b}^{t}$ represents the frequency of points concerning time t and K represents total data values [16]. BOW is also used to extract useful features from text data that can be classified later on. It works with a preset vocabulary and uses it to look for the frequency of particular phrases in the input data corpus [17]. The model is just concerned with whether or not known phrases exist in the document, not with where they appear, and it provides a histogram of such words within the data that can be readily given to classifiers. Eq. (2) shows the mathematical computation used by BOW to build word bags.

$$Dt_{a,b} = \sum_{xa,b=1}^{K} wt_a^b \; x \; wt_a \quad (2)$$

Where, $Dt_{x,y}$ indicates dataset containing points a,b. wt_a^b shows the weight of the frequently occurring word a concerning reference point b. wt_a represents the weight of the point of interest that appears most frequently [18]. In the presented work, the preprocessed dataset is provided to both TF-IDF and BOW to derive textual features.

4.2.2. Latent Dirichlet Allocation (LDA)

Topic modeling is an unsupervised document classification approach, that discovers natural groups of topics in a random document pool. It helps in discovering the hidden meanings behind the collection of words and classifies certain pairs of words based on their close meaning match with each other. LDA is one of the most often used topic modeling techniques. Each document is made up of different terms, and each topic has its own set of words. The goal of LDA is to determine which theme a document belongs to depending upon the words it contains. LDA works based on two things: the words present in a document that belong there and are known to us and the words present in a document that we need to calculate whether they belong there or not [19]. The generic approach through which LDA achieves this functionality is represented in Eq. (3).

$$p(wd.tc) = \sum_{doc=1}^{K} p\left(\frac{tc}{doc}\right) * p\left(\frac{wd}{tc}\right) \quad (3)$$

Where $p(wd.tc)$ represents the probability that word wd belongs to topic etc. The model goes through all the documents included in the data corpus starting from 1 to K and assigns each word to their corresponding most suitable topics [20]. In the proposed work, the dataset after preprocessing is given to LDA for most relative topics derivation where 15 topics are selected each having 30 frequent terms which are then scaled using a standard scalar before giving them to ML classifiers. Tab. 2 shows some frequent terms from 15 selected topics which are selected based on their gram weightage.

Tab. 2: Most frequent terms from LDA based selected topics

Unigram Terms	***Bigram Terms***	***Trigram Terms***
Campaign	Law Enforcement	Next-generation Neural
Vote	Illegal immigrant	Infowar life brain

Win	Climate change	American foreign policy
Government	Global Warming	
American	Health wellness	
Republican	Foreign Policy	
Conservative	Infowar placement	
Presidential	Marriage License	

4.3. Classification

After the steps of data cleansing, preprocessing, textual feature extraction, and most frequent topics derivation through LDA, the output is scaled using a standard scaler. Afterward, these textual features and topics are then given to three standalone ML classifiers SVM, NB, and LR as well as two ensemble-based ML models RF and GB for fake news classification. The results are analyzed and compared using several performance deduction measures. The experiments along with their results are discussed in detail in section 5.

5. Experimentation and Results

The proposed framework uses a data set comprising both fake and real news as input, applies certain preprocessing steps including stopwords removal, punctuation removal, lemmatization, and tokenization on it, extracts feature by first extricating unigrams, bigrams, and trigrams from the data, and then implementing N-gram methods TF-IDF and BOW on them. Ranked topics are also derived in parallel from the prepared data corpus. The output is then scaled and given to three standard ML models SVM, NB, LR, and two ensemble-based approaches RF and GB. The results are evaluated using performance measures accuracy, recall, precision, and f1-score.

Some experiments are performed using the proposed model for better evaluation and comparison. In the first experiment, the textual features obtained from TF-IDF, and BOW are given to standalone ML models SVM, NB, and LR, and the results are shown in Tab. 3 where the aforementioned performance standards are maintained. All the ML models are trained and tested against 90% and 10% of the dataset respectively. The experiments are carried out in Python, and the package used to integrate the model into our space is called sklearn ensemble.

Tab. 3: Classification results of TF-IDF and BOW with standard ML models

PEM	SVM-TFIDF (%)	SVM-BOW (%)	NB-TFIDF (%)	NB-BOW (%)	LR-TFIDF (%)	LR-BOW (%)
Acc.	95.5	87.40	88.97	77.16	94.96	92.12
Prec.	96	88	90	78	95	92
F1-Sc.	96	87	89	77	95	92
Rec.	96	87	89	77	95	92

Among the standalone models, SVM seems to be performing better than others in terms of accuracy as well as other performance measuring standards. In the next experiment, the same textual features are given to ensemble ML classifiers RF and GB where the same experiment settings are maintained, and results are evaluated. Tab. 4 shows the results of ensemble ML classifiers with textual features.

Tab. 4: Classification results of TF-IDF and BOW with ensemble ML models

PEM	RF-TFIDF (%)	RF-BOW (%)	GB-TFIDF (%)	GB-BOW (%)
Acc.	85.8	86.29	89.5	86.25
Prec.	86	86	89.57	86
F1-Sc.	86	86	89.52	86
Rec.	86	86	89.52	86

The results mentioned in Tab. 3 and Tab. 4 indicate that SVM stands out in terms of accuracy among other ML models by achieving a score of 95.5% on TF-IDF-based features while LR performs better on BOW based features with a 92.12% accuracy rate. Nearly all the ML models provide decent performance scores as shown in Fig. 2.

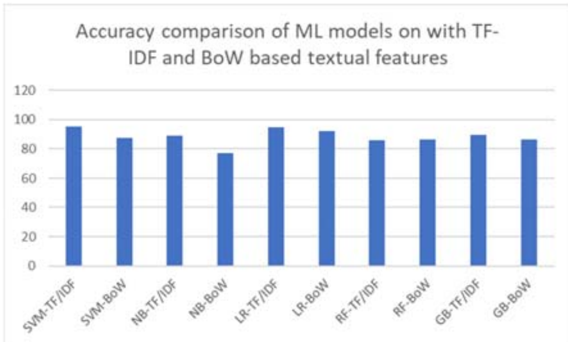

Fig. 2. Accuracy comparison of ML models with TF-IDF and BOW features

In the next experiment, unigrams, bigrams, and trigrams are extracted from the preprocessed dataset and are given as input to LDA for topics derivation. The extracted topics are then scaled using a standard scaler and are given to the same set of standard and ensemble ML models. All the ML models are trained and tested against 90% and 10% of the dataset respectively. The experiments are carried out in

Python, and the package used to integrate the model into our space is called sklearn ensemble. The results are shown in Tab. 5.

Tab. 5: Classification results of LDA topics with ML models

PEM	Standard			Ensemble	
	LDA-SVM (%)	LDA-NB (%)	LDA-LR (%)	LDA-RF (%)	LDA-GB
Acc.	80.08	69.2	79.08	79.82	68.7
Prec.	80	71	79	80	69
F1-Sc.	80	69	79	80	69
Rec.	80	68	79	80	69

As it is quite evident from Tab. 2 that among all ML models implemented on scaled features including both standard and ensemble methods, SVM provides the highest accuracy rate of 80.08% which is by far the best as compared to the rest as depicted in Fig. 3.

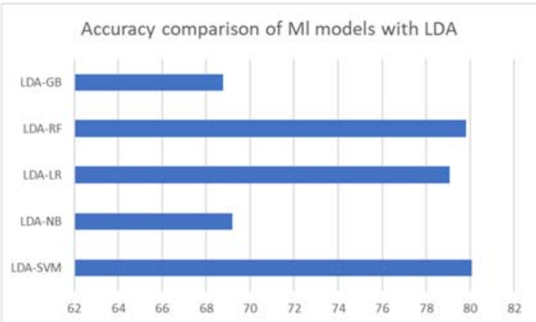

Fig. 3. Accuracy comparison of ML models with LDA

5. Discussion

In the proposed work, fake and real news dataset is obtained from online sources which are passed through certain preprocessing steps to cleanse and prepare them. The data is then given to N-gram methods TF-IDF and BOW for the extraction of textual features. At the same time data is also given to a topic modeling algorithm LDA for the derivation of key topics. The textual features are provided to a set of standard ML algorithms SVM, LR, NB and ensemble ML models RF, GB for classification. All the experiments are their results have been mentioned in section 4. It is noticed that nearly all ML models prove to be performing better on textual features in terms of accuracy. SVM provides the best accuracy of 95.5% on TF-IDF features while LR provides 92.12% accuracy on BOW features which is better than the rest. When the same dataset is passed through LDA-based topic modeling, key topics are retrieved which are scaled and given to the same set of ML classifiers. In this case, after looking at the results it can be deduced that SVM turns out to be the best accuracy providing model upon LDA output and achieves 80.08% accuracy. But when the results are of textual features classification and LDA classification are compared, it is noted that ML performs better on textual features as compared to LDA as the outcomes of ML models concerning all the performance measures are far better than any LDA based outcome. Fig. 4 further elaborates this observation.

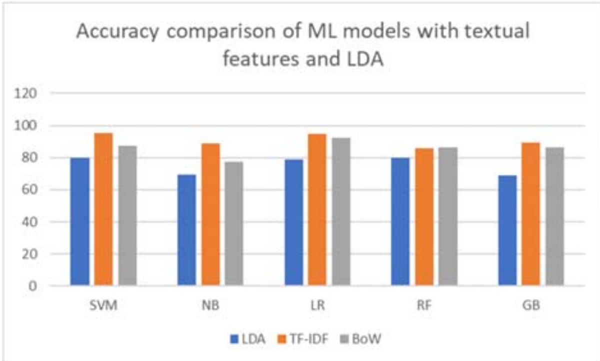

Fig. 4. Accuracy comparison of ML models on textual features and LDA topics

Conclusion

The proposed work focuses on retrieving fake and real news dataset from online sources, passing it through certain preprocessing steps and extracting textual features from them based on N-gram methods TF-IDF and BOW. Moreover, it provides a method to extract key topics from the dataset through the implication of a topic modeling-based LDA algorithm. These topics are then scaled using standard scalar and finally classification of fake news is done using both textual features and post-scaled topics provided by LDA through standard ML models SVM, NB, LR, and ensemble ML models RF and GB. The results indicate that ML models perform better on textual features as compared to extracted LDA-based topics.

In future, we can utilize transformer-based techniques such as BERT and Generative Pre-Trainings (GPT) along with word embeddings Word2Vec, Global Vectors for word representation (GloVe) and Deep Learning traditional algorithms Convolutional Neural Networks (CNN) and Long Short-Term Memory Network (LSTM).

References

[1] Di Domenico, G., Sit, J., Ishizaka, A. and Nunan, D., 2021. Fake news, social media, and marketing: A systematic review. Journal of Business Research, 124, pp.329-341.

[2] Apuke, O.D. and Omar, B., 2021. Fake news and COVID-19: modeling the predictors of fake news sharing among social media users. Telematics and Informatics, 56, p.101475.
[3] Mahir, E.M., Akhter, S. and Huq, M.R., 2019, June. Detecting fake news using machine learning and deep learning algorithms. In 2019 7th International Conference on Smart Computing & Communications (ICSCC) (pp. 1-5). IEEE.
[4] Wani, A., Joshi, I., Khandve, S., Wagh, V. and Joshi, R., 2021, February. Evaluating deep learning approaches for covid19 fake news detection. In International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation (pp. 153-163). Springer, Cham.
[5] Alonso, M.A., Vilares, D., Gómez-Rodríguez, C. and Vilares, J., 2021. Sentiment analysis for fake news detection. Electronics, 10(11), p.1348.
[6] Jain, A., Shakya, A., Khatter, H. and Gupta, A.K., 2019, September. A smart system for fake news detection using machine learning. In 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)(Vol. 1, pp. 1-4). IEEE.
[7] Jiang, T., Li, J.P., Haq, A.U., Saboor, A. and Ali, A., 2021. A novel stacking approach for accurate detection of fake news. IEEE Access, 9, pp.22626-22639.
[8] Hakak, S., Alazab, M., Khan, S., Gadekallu, T.R., Maddikunta, P.K.R. and Khan, W.Z., 2021. An ensemble machine learning approach through effective feature extraction to classify fake news. Future Generation Computer Systems, 117, pp.47-58.
[9] Mahir, E.M., Akhter, S. and Huq, M.R., 2019, June. Detecting fake news using machine learning and deep learning algorithms. In 2019 7th International Conference on Smart Computing & Communications (ICSCC) (pp. 1-5). IEEE.
[10] Kaliyar, R.K., Goswami, A. and Narang, P., 2019, December. Multiclass fake news detection using ensemble machine learning. In 2019 IEEE 9th International Conference on Advanced Computing (IACC) (pp. 103-107). IEEE.
[11] Liu, Y. and Wu, Y.F., 2018, April. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
[12] Shahbazi, Z. and Byun, Y.C., 2021. Fake media detection based on natural language processing and blockchain approaches. IEEE Access, 9, pp.128442-128453.
[13] Akhter, M.P., Zheng, J., Afzal, F., Lin, H., Riaz, S. and Mehmood, A., 2021. Supervised ensemble learning methods towards automatically filtering Urdu fake news within social media. PeerJ Computer Science, 7, p.e425.
[14] Adiba, F.I., Islam, T., Kaiser, M.S., Mahmud, M. and Rahman, M.A., 2020. Effect of corpora on classification of fake news using naive Bayes classifier. International Journal of Automation, Artificial Intelligence and Machine Learning, 1(1), pp.80-92.
[15] Al-Shalabi, R. and Obeidat, R., 2008, March. Improving KNN Arabic text classification with n-grams based document indexing. In Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt (pp. 108-112).
[16] Wei, Z., Miao, D., Chauchat, J.H., Zhao, R. and Li, W., 2009. N-grams based feature selection and text representation for Chinese Text Classification. International Journal of Computational Intelligence Systems, 2(4), pp.365-374.
[17] Soumya George, K. and Joseph, S., 2014. Text classification by augmenting bag of words (BOW) representation with co-occurrence feature. IOSR Journal of Computer Engineering, 16(1), pp.34-38.
[18] Zhang, Y., Jin, R. and Zhou, Z.H., 2010. Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics, 1(1), pp.43-52.
[19] Zrigui, M., Ayadi, R., Mars, M. and Maraoui, M., 2012. Arabic text classification framework based on latent dirichlet allocation. Journal of computing and information technology, 20(2), pp.125-140.
[20] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.

Ammar Saeed did his bachelor's in computer science from COMSATS University Islamabad, Pakistan in 2019. Currently, he is pursuing master's in computer science from COMSATS University Islamabad, Pakistan. His major areas of research interest are Machine Learning and Data Analytics. His research projects involve Machine Learning techniques for vehicle classification and database watermarking using Block Chain.

Eesa AL Solami is Associate professor of computer science and engineering at university of Jeddah Currently, he is a dean of Admission and registration of university of Jeddah. He got his PhD in 2012 from Queensland university of technology from Australia. His research projects involve feature selection techniques for continuous biometric authentication. Eesa graduated in computer science in 2002 from King Abdul-Aziz University, and then in 2008 he received MSc in IT from Queensland university of technology.