Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021)
IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7

2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395896

A Novel Score-Based Multi-Source Fake News Detection using Gradient Boosting Algorithm

S. Selva Birunda
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
India
sbirunda@gmail.com

Dr. R. Kanniga Devi
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
India
rkannigadevi@gmail.com

Abstract - **The dissemination of fake news in an online social network platform has been a real concern nowadays. Through social media, news articles are posted by many sources like news channels, websites, or even newspaper websites. There is a need to be sure that the information posted is only from credible sources and these posts have to authenticate. The intensity of genuineness of the news posted online cannot be measured absolutely and is still challenging. A novel Score-based Multi-Source Fake News Detection framework is proposed in this work to automate the detection of fake news from multiple news sources. This framework extracts the text-based features from genuine and fake news articles using Term Frequency - Inverted Document Frequency. Then the credibility score of the sources is calculated based on the site_url features and Top Level Domain. By assimilating the text-based features with the credibility score of multi-source, the credibility of the news is estimated. The proposed framework is applied to the Machine Learning (ML) classifiers to examine their performance in the detection of fake news. The experimental results determine the efficacy of the proposed framework with the Gradient Boosting algorithm of about 99.5% to the utmost level.**

Keywords—multi-source; fake news; top-level domain; natural language processing; score; gradient boosting; credible

I. INTRODUCTION

In the flash of the current attention to the demand of Social Networking sites has expeditiously increased in recent years. Many people tend to consume news information from social networking sites rather than the traditional news media [1]. The reason is that social networks help the users to connect with link-minded people, share their views by interacting with them. However, the quality of news is a query. News information with deceitful information is shared through this media [2]. This extensive spread can have a pessimistic impact on society and individuals. Fake news is a complete fabrication of information that is not able to confirm the facts. The degree of reliability of the information posted on social media cannot be measured accurately. To tackle the above problem, a standard approach is needed.

Some of the earlier fake news detection methods follow almost the same detection procedure. Every dataset undergoes preprocessing to remove noise information from the dataset. Even though, by using a large and noisy dataset fake news prevailing in Twitter is detected with pretty good performance in [5]. Various features extracted from the news information include linguistic features, linguistic cue methods, tweet level features, and Natural Language Processing (NLP) feature. Some of the feature extraction methods are Linguistic Analysis and Word Count (LIWC) [2], Semantic Analysis [3], Probabilistic Context-Free Grammar, (TF-IDF) Term Frequency-Inverted Document Frequency [6], bag-of-words, *n*-grams, and Doc2Vec [9]. Once the features are extracted, it is given to the Machine Learning classifiers for training and to test the effective prediction of new unlabeled news information. Once the features got extracted, it is given to the Machine Learning classifiers for learning. Most of the Machine Learning classifiers investigated in the previous fake news detection methods are Naïve Bayes Classifier [10], Neural networks, Support Vector Machine (SVM), Random Forest, XGBoost [5], Decision Tree [2], Linear Regression, Logistic Regression, K Nearest Neighbors (KNN) [6], AdaBoost Stochastic Gradient Descent (SGD), and LinearSVM [9].

Naïve Bayes classifier lowers in performance as it assumes the independent features or predictors. In the SVM classifier, support vectors are the coordinates of a particular sample. Kernel trick is the biggest strength of this classifier. Even though the SVM classifier can successfully classify the two classes, it does not fit the datasets that are large and that have noise. KNN requires more memory to store the training data and it is expensive. Ensembles provide high accuracy compared to other algorithms. Under Boosting and AdaBoost are sensitive to outliers and noisy data. Hence, Gradient Boosting is chosen in this proposed framework to overcome the issues mentioned above. Most of the existing datasets lack citing the source from where the news information originates. Besides, the previous fake news detection methods are single-sourced. Citing the source information is unexplored, but it is useful as this aspect has the insight to reduce the fake news mitigation to the maximum level. The source from which news information originates may be credible or non-credible. As far as the credibility of the sources is concerned, checking the sources for reliability is essential. Several factors needed to identify the reliable sources are timeliness (Date of publication), Authority (Author Credentials), and Purpose (Top Level Domain and Domain).

Outdated sources are often non-credible. If a news article is published 10 years back, then the reliability of the article is a query. The author credentials are used to evaluate the reliability of the author of a news event [8]. To identify the reliability of news articles it is needed to examine the purpose of the source learned from Top Level Domain (TLD). None of the existing works focused on the purpose of the source.

To overcome the aforementioned issues and to build an efficient classification model, A Novel Score-based Multi-Source Fake news Detection Framework has been devised. In this proposed framework, the purpose of the source is the primary concern to determine the credibility of the news. Traditionally, the Machine Learning (ML) classifiers mentioned above can predict the fake news prevailing only in a single source. In contrast, the proposed framework concentrates on multiple sources. Some of the existing methods for multi-news sources use approaches such as the integration of author-based and content-based features [8], publishers' site age [6] to establish the credibility of the source and news articles. The purpose of the source is unexplored.

The objectives and key results of the proposed work is a goal-setting framework for detecting the fake news prevailing in the multi-source platform and are listed below:

a) To extract the top-level text-based features from genuine and fake news articles using TF-IDF

b) To extract the site_url features from site_url (domain)

c) To estimate the credibility score of multi-source based on-site URL features

d) To estimate the credibility of the news by integrating the text-based features and the Credibility Scores of multi-source

The remaining section of this article is systematized as follows: Section 2 gives detailed explanations of the related works regarding fake news detection. Section 3 explains the methods and execution of the proposed framework. Section 4 describes the details of the classification model used for prediction. Finally, Section 5 depicts the experimental results of the proposed framework.

II. Related Work

This section characterizes the traditional approaches and methodologies used to detect fake news. Besides, the pros and cons of earlier approaches discussed as follows:

D.V.Singh et al. [2] obtained linguistic features from each article using the Linguistic Analysis and Word Count Package to detect fake news automatically. The extraction of Authentic, Tone, Word count, Clout, and Analytic features to identify fake news, rather than extracting a single feature. Besides, Z-score normalization preferred to normalize features. Comparison of Multiple classifiers like Support Vector Machine (SVM), Decision Tree, Logistic Regression, Random Forest, and K-neighbors used to analyze the performance of the classifier. Consequently, SVM could predict fake news automatically with better prediction results. Even this approach presented better detection of fake news; it focused only on the detection of the truthfulness of articles rather than aiming at the credibility of multi-source.

K. Stahl [3] proposed a method to detect fake news prevailing in social media by integrating SVM, Semantic Analysis, and Naïve Bayes classifier into one algorithm. SVM and Naïve Bayes classifiers were combined to classify the data efficiently than each classifier was individual. Semantic Analysis is one of the linguistic cue methods used to find the relationship among the words. Integrating the afore-mentioned techniques improved the performance of the classification model in detecting fake information. This suggested approach targeted only on single-source failed to target information coming from multi-source.

S. Helmstetter and H. Paulheim [5] detected fake news on Twitter by training a large and noisy dataset. Tweet level features, topic features, user-level features, sentiment features, and text features extracted from a tweet. The proposed approach aimed at identifying fake news tweets and sources. Here, the source indicated the user account from where the tweets originated. Multiple classifiers such as Neural networks, Random Forest, SVM, Naïve Bayes, and XGBoost classifiers deployed to classify the tweets. XGBoost algorithm classified the tweets with superior performance. This proposed approach has a drawback in that it targeted only a single source, i.e. Twitter. Besides, it failed to target news articles from multiple sources.

S. Gaonkar et al. [6] identified the news articles as genuine and fake based on scores by combining many NLP techniques with the Machine Learning algorithms. Features extracted from headlines and articles using Probabilistic Context-Free Grammar and TF-IDF. A score was generated to detect the author's stance approaching the articles. Assigning a score to the site based on the site availability time online. The site name got compared with a list of credible websites. If the site age is more, there is a chance that the site is genuine, fake otherwise. Machine Learning algorithms such as SVM, Linear Regression, Naïve Bayes, Logistic Regression, K Nearest Neighbor (KNN) and Linear SVM have been used. The scores have been generated for each of the classifiers. So far, the scores generated were aggregated, added, and then averaged to obtain an ending score that helped to decide the news articles as genuine or fake. This proposed approach targeted the parameter site age and failed to target the Top Level Domain and website verification. Fake news publishers used credible sites registered in the early 1990s and register them again in a short period [7].

H. Liu et al. [8] proposed a Fake News Detector Based on Multi-Source Scoring (FNDMS) framework to detect fake news from multiple news sources. Content-based and author-based features are employed to evaluate the trustworthiness of a single news source. Dempster-Shafer Theory (DST) model integrated the genuineness of multiple sources that end up giving a judgment on the legitimacy of a news event. The effectiveness of the framework was validated when comparing it with SVM, Logistic Regression, Random Forest, and AdaBoost. It would be better if this framework focuses particularly on the purpose of the source.

M. Z. Khan and O. H. Alhazmi [9] suggested a solution to the issue of unreliable news. Doc2Vec technique was used to extract the features from the news articles. The ensemble classification approach was developed using the techniques AdaBoost and Bagging. Random Forest, Decision Tree, Naïve Bayes, and SVM acted as base learner for ensemble techniques. AdaBoost combination with LinearSVM and Random Forest performed better with high accuracy. This suggested approach needed to focus on the purpose of the source.

M. Granik and V. Mesyura [10] detected fake news in the dataset of Facebook news posts using Naïve Bayes Classifier. A threshold-based detection method was proposed based on the probability of the article. If the probability of the article is greater than the threshold value, then the news was genuine otherwise fake. This approach classified the news articles with better accuracy even though accuracy could improve further. Besides, this method failed to focus on the news coming from multi-source.

H. Ahmed et al. [11] proposed an *n*-gram analysis model to detect fake news automatically. TF and TF-IDF feature selection techniques extracted the features from the text. Six different Machine Learning Classifiers, two methodologies for selecting features, were compared to determine the reliability of the news. SVM, KNN, SGD, LinearSVM, and Decision Tree were compared to six classifiers. The proposed model achieved the highest accuracy while using the Unigram, TF-IDF, and LinearSVM classifiers. This suggested model did not focus on the multi-source and credibility of the project.

V. Agarwal et al. [12] suggested a fake news detection approach by extracting the features from the text using a bag of words, *n*-grams, and TF-IDF techniques. The extracted features were given to the classifiers namely, Naïve Bayes classifier, LinearSVM, Logistic Regression, Random Forest, and SGD for classification. Logistic Regression and SVM outperformed all the classifiers with the best performance. This suggested model did not concentrate on the multi-source and credibility of the source.

S. S. Birunda and R. K. Devi [15] proposed a classification framework that is energy efficient that uses Voting Ensemble Machine Learning for higher accuracy but it fails to concentrate on Multi-source.

III. Score-Based Multi-Source Fake News Detection Framework

This section contributes to the detailed methodology of the proposed 'Score-Based Multi-Source Fake News Detection' framework. Initially, the news articles are collected, filtered, and preprocessed. Then, the text-based features are extracted that assist in evaluating the credibility of news articles. Furthermore, a score-based multi-source method is proposed for assimilating the credibility score for the source and the text-based features in choosing the legitimacy of a news article. The pipeline of the proposed framework is illustrated in Fig.1.

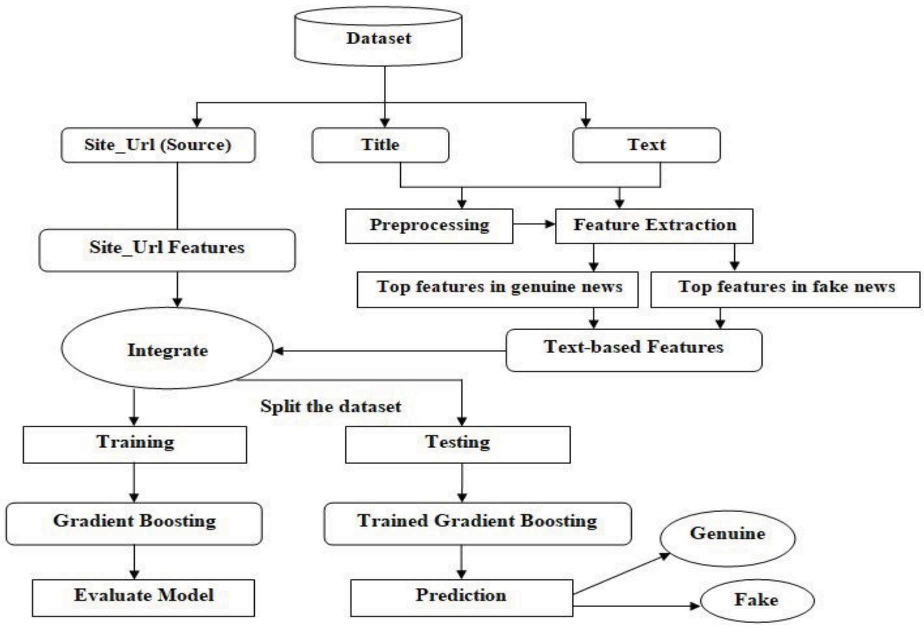

Fig. 1. Pipeline of the proposed Framework.

A. Problem Statement

Definition: A news article originating from a particular source is said to be fake if its content and its credibility score are both verified to be non-credible (false) and legitimate (true) otherwise.

Let $D = \{D_1, D_2 \ldots D_n\}$ be a dataset consisting of n news articles originating from multiple sources, say (s), it is denoted as $D = \{d_1^1, d_2^2, \ldots, d_n^s\}$. Furthermore, Y = {Real, Fake} be the class labels defined in the dataset. Given the dataset *D,* news articles from multiple sources (s), and with the class labels Y, are used to predict the degree of fakeness for the new unlabeled news article.

B. Dataset

The experimental dataset collected from Kaggle.com contains 2050 news articles labeled as fake or real. The size of the dataset is 10.4 MB. This dataset includes the metadata regarding the source of the news articles, date of publication, author, title, text, does it contains an image, type of the text, labels, etc. The dataset gets split up into 80% of training data and 20% of testing data. For training, out of 1574 news articles, 976 fake samples and 598 real samples are used. For testing, out of 393 news articles, 241 fake samples and 152 real samples are used. Fake news article detection is a problem of binary classification. Fig.2. depicts the distribution of labels with 1217 fake news articles and 750 real news articles.

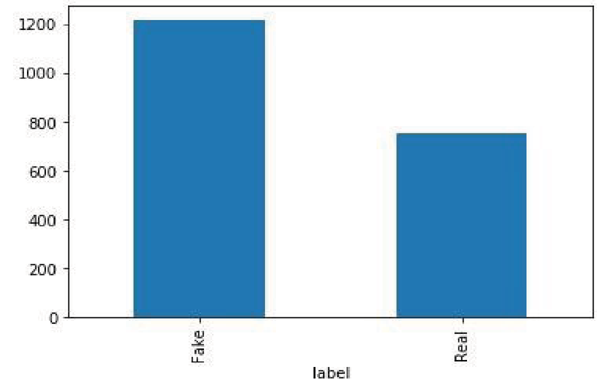

Fig. 2. Distribution of Labels.

From the dataset, there were only four fields utilized by our proposed work such as Site_Url, Text, Title, and Date of Publication. All the news articles published within 5 years are likely to be current data and have less chance of fake news propagation. Here all the news articles are published in the year 2016, which is current. But this aspect will not work out often. There is a need to review the content for credibility. The news articles in the English language were selected among English and German languages news articles provided in the dataset. The parameters that influence mostly the consideration of news as fake; are text features, Date of publication, Top Level Domain, text content having all uppercase letters, and site_url features.

C. Preprocessing

A news article endures preprocessing for cleaning the raw text data which is to be used by the Machine Learning classification model. Preprocessing encompasses tokenization, removal of stop words and punctuation, and lowercasing. Tokenization is a process of segregating the text of news articles into tokens which will be useful to remove stop words, punctuation easily. Stop words such as ‘it’, ‘as’, ‘and’ are removed as these words are having low importance in the news articles. Punctuation is removed, as it has no importance in finding fake or genuine news articles because both have punctuations. Lowercasing is done because articles having full uppercase letters got a chance of having fake content.

D. Feature Extraction

In this framework, text-based features are extracted using TF-IDF, which aids in detecting the news articles as fake or genuine. This technique is a weighting metric in NLP that converts text into features, as the Machine Learning classification model will discard processing the raw text. It calibrates the importance of a particular word in the news article [15]. The frequency of the extracted features from the genuine and fake news articles is shown in Fig.3 and 4.

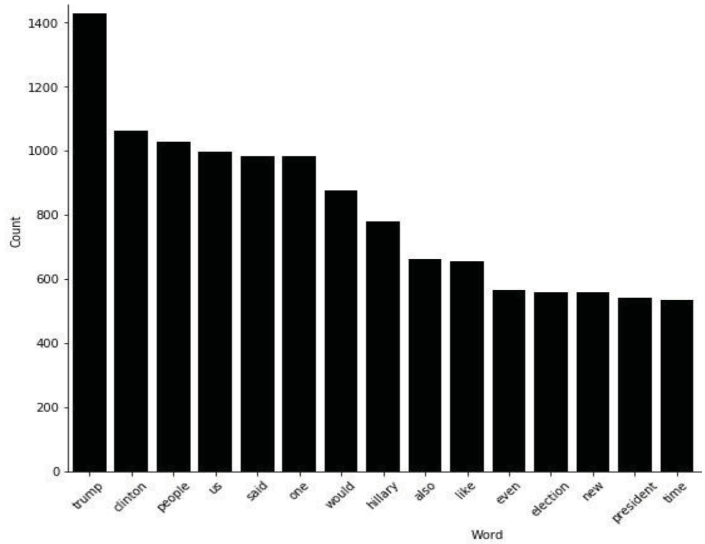

Fig. 3. Frequency of top 15 genuine features.

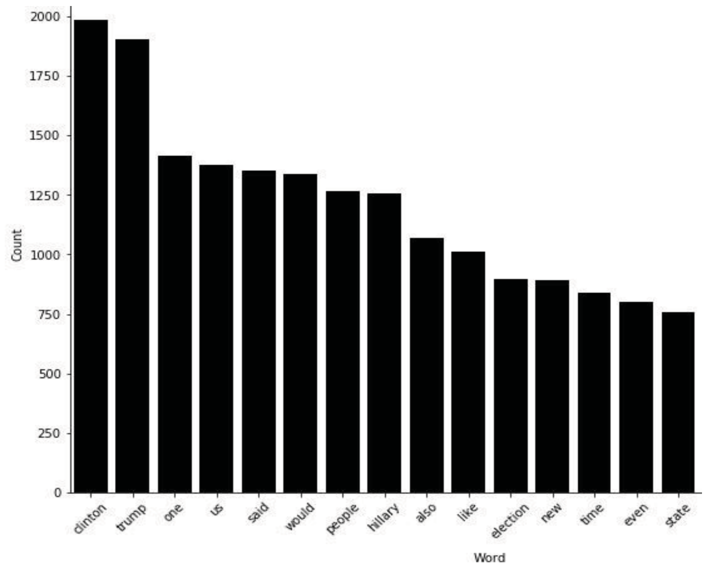

Fig. 4. Frequency of top 15 fake features.

Input: News articles
Output: Top features from the text

1. Extract top features from the text using TF-IDF
2. TF-IDF calculates the importance of a term in an article
3. Calculation of TF:
TF (t, n) = Occurrence of term t in news article, n / Total terms t_i in the news articles
4. Calculation of IDF:
IDF (t, n) = log [Total number of news articles / Number

of articles containing that term]
if IDF score closer to 0,
then the term is more common skip that term
else
include the term whose IDF score is closer to 1
5. Calculation of TF-IDF:
$$TF\text{-}IDF\ (t, n) = TF(t, n) * IDF\ (t, n)$$
6. Best features are selected based on TF-IDF score
7. Choose n terms, t_n with top feature values
8. Extract top features out of real articles and fake articles
end

Word Cloud presents the visualizations of frequently used features/words in word clusters. The more frequently a word is mentioned in the text, the larger the words in the word cloud. Word cloud visualization for fake and legitimate news is depicted in Fig.5 and 6.

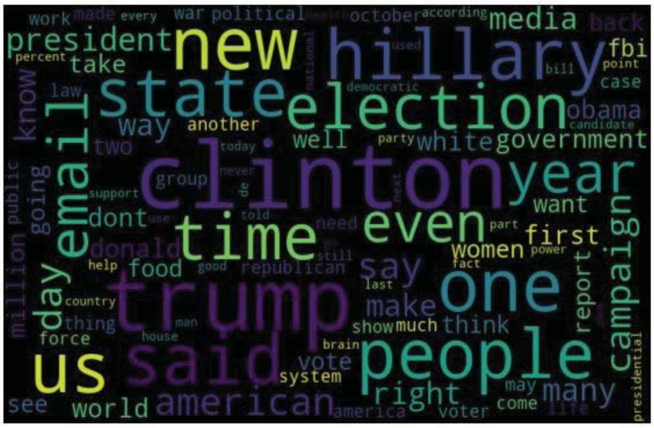

Fig. 5. Word Cloud for fake news.

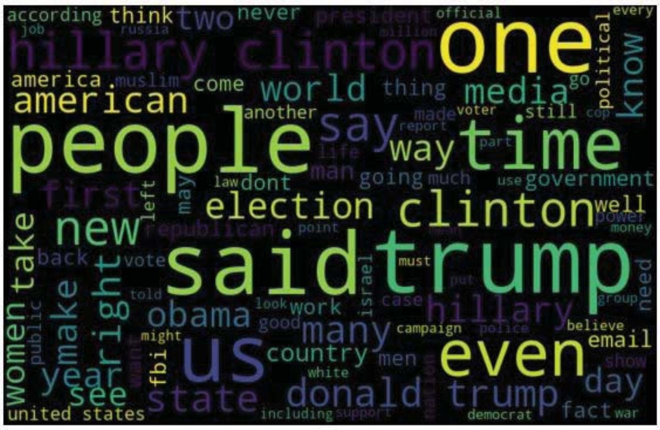

Fig. 6. Word Cloud for legitimate news.

E. Credibility Scoring of Multi-Source based on Site_Url Features

In a nutshell, domain name or site_url (source) incorporates Top Level Domain. TLDs are of two types Generic TLD and Country Code TLD. Some of the Generic TLD is '.edu' (educational), '.gov' (government websites), '.com' (commercial), '.org' (non profit), '.net' (network) etc. Some of the Country Code TLD is '.in' (India), '.ca' (Canada), etc. There are about 700 TLD available while those mentioned above are unique among them. Even, these domains are top-level domains; some sites use these TLD to mislead.

In this Score-Based framework, the site-url is assigned scores ranging from 0 to 10. Site_url features are extracted from the multi-source to calculate the credit score of site_url (source). Scoring is assigned to the source based on the TLD. A pipeline of the Credibility Score of Multi-Source based on Site_Url features is depicted below in Fig. 7.

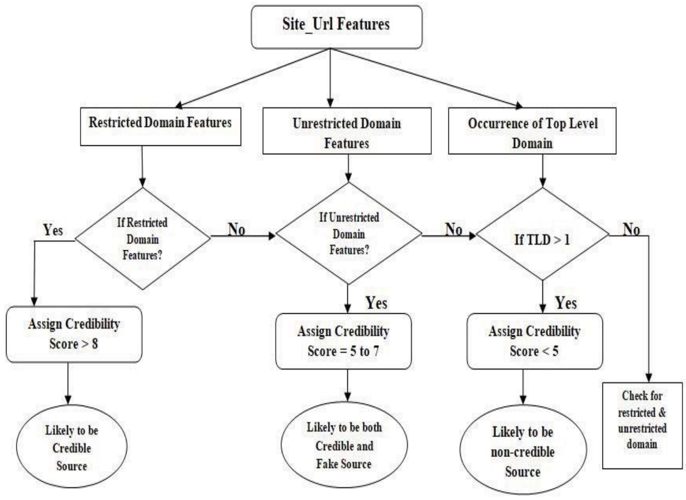

Fig. 7. Pipeline of Credibility Score of Multi-Source based on Site_Url features.

1) Site_Url Features

a) Restricted Domain Features: Generally, '.edu' '.gov' and '.mil' are sponsored Top Level Domain, hence considered the most trusted and credible sources. Besides, these TLD are restricted domains as they have to meet certain procedures to register domains [3]. If TLD contains restricted domain features such as '.edu' '.gov' and '.mil' exist, these TLD is assigned a credibility score above 8 (i.e.) 8 to 10.

b) Unrestricted Domain Features: There is some unrestricted domain which anyone can register [3], like '.net' '.org' '.com' and some Country Code domains. As stated as unrestricted, there is a chance of misleading hence these domain features may or may not be credible. Consequently, it needs some investigation to glimpse for indications of reliability. Therefore, these unrestricted domain features are assigned a credibility score between 5 and 7.

c) Occurrence of Top Level Domain: An unusual domain such as '.com.co' is often a non-credible version of legitimate news sources. There is a need to verify the occurrence of the Top Level Domain. In the '.com.co' domain, '.com' and '.co' is two separate TLDs that occur coherently. If more than one Top Level Domain appears in a site_url, there is a chance for the source to be non-credible. Therefore, these occurrence features are assigned a credibility score below 5.

The algorithm for calculating the credibility score of a multi-source based on the site_url features explained below:

Input: Site_url features
Output: Trustworthiness and Credibility Score of multi-Source

if TLD has restricted domain features = '.edu', '.mil' & '.gov'
then assign Credibility score (CS_{source}) for the site_url (source, S) = 8 to 10
if score > 8
then site-url is likely to be credible source
else
if TLD has unrestricted domain features = '.com', '.org' & '.net'
then assign Credibility score (CS_{source}) for the site_url (source, S) = 5 to 7
if score = 5 to 7
then site-url is likely to be both credible and non-credible source

 # Occurrence of Top Level Domain
else
if TLD > 1
then assign Credibility score (CS_{source}) for the site_url (source, S) > 5
if score > 5
then site_url is likely to be non credible source
else
credible source
end

From the site_url features and TLD, the sample of some of the source publishing real news, fake news, and sources publishing both real and fake news is identified as shown in Fig. 8, Fig. 9, and Fig. 10.

Fig. 8. Sample of some site_url (sources) publishing real news.

Fig. 9. Sample of some site_url (sources) publishing fake news.

Fig. 10. Sample of some site_url (sources) publishing both real and fake news.

F. *Integration of Text- based Features and Credibility Score of Multi-Source*

The text-based features were extracted from the site_url features of the news articles. Then, the credibility score for the source domain is computed based on site_url features. The next step is to integrate both the text-based features (F text) and credibility score of the sources (CS source) to find the patterns in deciding the trustworthiness of the news articles (C news). The trustworthiness of the jth news article computed as follows in (1):

$$C^{j}_{news} = CS^{j}_{source} + F^{j}_{text} \quad (1)$$

Legitimate sources have a chance of delivering fake content and its content needs verification. In contrast, if the source itself is non-credible, then automatically it delivers the fake content. Hence, the verification of content and the source is processed here the patterns are recognized resulting in predicting the degree of credibility for the new unlabeled news article.

The algorithm for calculating the credibility of news articles based on text-based features and credibility score of multi-source explained below:

Input: Text-based features, a Credibility score of multi-source
Output: Credibility of news articles, 0 – genuine, 1 - fake

if text-based features = 'genuine' & Credibility Score of source > 8
then news article is genuine
else
if text-based features = 'fake' & Credibility Score of source = 5 to 7
then news article is fake
else
if text-based features = 'genuine' & Credibility Score of source = 5 to 7
then news article is genuine
else
if text-based features = 'fake' & Credibility Score of source < 5
then news article is fake
end

IV. Classification Model

After assimilating the text-based features and Credibility Score of Multi-Source, it is fed to the Machine Learning classification model for training and to test the effective prediction of our proposed framework towards new unlabeled news information.

Machine Learning algorithms such as SVM, KNN, Naïve Bayes, Logistic Regression, Random Forest, AdaBoost, XGBoost, Decision Tree had been used previously to detect fake news. In this proposed approach, the Gradient Boosting classifier is used as a type of Ensemble boosting for the classification purposes.

A. *Gradient Boosting*

Gradient Boosting is an Ensemble Machine Learning algorithm used for news classification. Gradient boosting classifier trains multiple models in additive, gradual, and sequential manner. Generally, Boosting is an approach that converts weak learners (typically Decision trees) into strong learners by minimizing the error. Many weak learners are aggregated to produce strong learners. Besides, Gradient Boosting will improve the performance of the model with superior accuracy.

V. EXPERIMENTAL STUDY

A. *Performance Evaluation*

In this section, the effectiveness of our proposed framework was evaluated based on the following evaluation metrics:

Accuracy [9][13] of a classifier is the measure of summation of correct classification to the summation of all classifications. It is defined as follows in (1),

$$\text{Accuracy} = \frac{TN + TP}{TP + FP + TN + FN} \quad (1)$$

From (1), TP, TN, FP, and FN denote True Positive, True Negative, False Positive, and False Negative. True Positive indicates the number of accurate classifications of the positive samples. True Negative indicates the number of accurate classifications of the negative samples. False Positive indicates the number of misclassification of positive samples as a negative sample. False Negative indicates the number of misclassification of negative samples as a positive sample.

F1-Score is a harmonic mean of recall and precision [14, 15]. It is defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1 Score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \quad (4)$$

Performance evaluation of the proposed framework based on the evaluation metrics tabulated in Table. 1. From the table, it is examined that the Gradient Boosting classifier achieves 99.5% Accuracy, 1.00 Precision, 0.99 Recall, and 0.99 F1 Score. This classifier outperforms all other classifiers in terms of Accuracy, Precision, Recall, and F1 Score which are the evaluation metrics employed in the proposed framework to validate the developed model.

B. *Experimental Results*

In this section, the performance analysis of the proposed framework was demonstrated by applying many classifiers namely, SVM, KNN, Naïve Bayes, Logistic Regression, Random Forest, AdaBoost, Decision Tree, and Gradient Boosting. The proposed framework along with the Gradient Boosting classifier achieves remarkable accuracy of about 99.5% which is superior to other classifiers. The performance analysis of multiple classifiers based on accuracy, precision, recall, and F1 Score is depicted in Fig. 11 and Fig. 12.

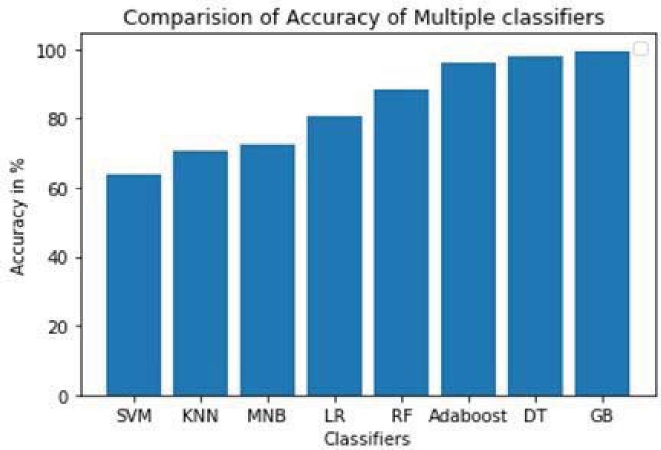

Fig. 11. Performance Analysis of Multiple classifiers based on Accuracy.

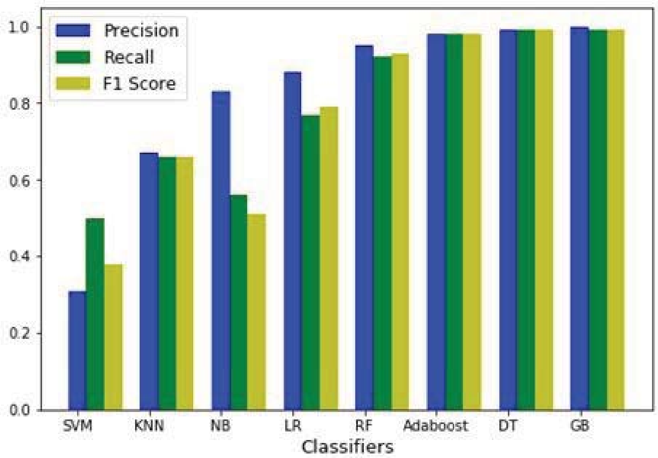

Fig. 12. Performance Analysis of Multiple classifiers based on Precision, Recall, and F1 Score.

From Fig. 11, it is shown that Gradient Boosting achieves high accuracy of about 99.5%. From Fig. 12, it is shown that Gradient Boosting achieves high precision, recall, and F1 score of about 1.00, 0.99, and 0.99. Besides, the tree-based algorithms like Random Forest, AdaBoost, and Decision Tree are suitable for this dataset and achieve the best performance than SVM, KNN, Logistic Regression, and Naïve Bayes. All the classifiers presented in the existing works except XGBoost are compared along with the Gradient Boosting to evaluate their prediction performance in a multi-source platform. XGBoost classifier left for future work.

TABLE I. CLASSIFICATION RESULTS OF MULTIPLE CLASSIFIERS

Classifiers	Accuracy (%)	Precision	Recall	F1 Score
Support Vector Machine	64	0.31	0.50	0.38
K-Nearest Neighbors	70.6	0.67	0.66	0.66
Multinomial Naïve Bayes	72.3	0.83	0.56	0.51
Logistic Regression	80.7	0.88	0.77	0.79
Random Forest	88.3	0.95	0.92	0.93
AdaBoost	96	0.98	0.98	0.98
Decision Tree	98	0.99	0.99	0.99
Gradient Boosting	**99.5**	**1.00**	**0.99**	**0.99**

VI. CONCLUSION

In Fake news detection, the major issue is the credibility of the news originating from multi-source. To determine the credibility of news articles, it is needed to examine the purpose of the source. To overcome the aforementioned issues and to build an efficient classification model, A Novel Score-based Multi-Source Fake news Detection Framework has been devised. The top real and fake features were extracted from the news articles using the TF-IDF technique. Based on the Site_url features obtained from the source, the Credibility Score of the sources got evaluated. The extracted text-based features and the Credibility Score of multi-source were integrated to estimate the reliability of the news. The effectiveness and feasibility of the proposed framework are evaluated and compared with the multiple classifiers. As a future research direction, the proposed framework could extend to XGBoost and deep learning techniques with better accuracy. Besides, datasets with more articles can be supplied to this framework.

REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD explorations newsletter, 19(1), pp. 22-36, 2017.

[2] D. V. Singh, R. Dasgupta, and I. Ghosh, "Automated fake news detection using linguistic analysis and machine learning," In International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS), pp. 1-3, 2017.

[3] K. Stahl, "Fake news detection in social media," California State University Stanislaus, 6, 2018.

[4] https://www.mytechlogy.com/IT-blogs/9754/5-uncommon-top-level-domain-names-and-why-you-should-consider-them/#.X-s_xNgzbIV

[5] S. Helmstetter, and H. Paulheim, "Weakly supervised learning for fake news detection on Twitter," In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) pp. 274-277, IEEE, 2018.

[6] S. Gaonkar, S. Itagi, R. Chalippatt, A. Gaonkar, S. Aswale, and P. Shetgaonkar, "Detection Of Online Fake News: A Survey," In 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), pp. 1-6, IEEE, 2019.

[7] K. Xu, F. Wang, H. Wang, and B. Yang, "Detecting fake news over online social media via domain reputations and content understanding," Tsinghua Science and Technology, 25(1), pp. 20-27, 2019.

[8] H. Liu, L. Wang, X. Han, W. Zhang, and X. He, "Detecting Fake News on Social Media: A Multi-Source Scoring Framework," In 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA) pp. 524-531. IEEE, April 2020.

[9] M. Z. Khan, and O. H. Alhazmi, "Study and analysis of unreliable news based on content acquired using ensemble learning (prevalence of fake news on social media)," International Journal of System Assurance Engineering and Management, 11(2), pp. 145-153, 2020.

[10] M. Granik, and V. Mesyura, "Fake news detection using naive Bayes classifier," In 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 900-903. IEEE, 2017.

[11] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," In International conference on intelligent, secure, and dependable systems in distributed and cloud environments (pp. 127-138). Springer, Cham, 2017.

[12] V. Agarwal, H. P. Sultana, S. Malhotra, and A. Sarkar, "Analysis of Classifiers for Fake News Detection," Procedia Computer Science, 165, 377-383, 2019.

[13] S. Smys, B. Abul, and W. Haoxiang, "Hybrid Intrusion Detection System for Internet of Things (IoT)," Journal of ISMAC 2, no. 04, 190-199, 2020.

[14] W. Haoxiang. "Emotional Analysis of Bogus Statistics in Social Media." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 03, 178-186, 2020.

[15] S. S. Birunda, and R. K. Devi, “Improving Energy Efficient Aspect of Spam Classification Framework using Ensemble Machine Learning,” Solid State Technology, 63(5), 9032-9039, 2020.