International Journal of Safety and Security Engineering
Vol. 12, No. 3, June, 2022, pp. 311-318
Journal homepage: http://iieta.org/journals/ijsse

Fake News Identification Using Regression Analysis and Web Scraping

Sandeep Dwarkanath Pande[1,2*], Suresh Rathod[3], Rahul Joshi[3], Gurunath T. Chvan[4], Digambar Jadhav[5], Pravin Phutane[6], Sudhanshu Gonge[3], Kalyani Kadam[3]

[1] School of Computer Engineering and Technology, MIT, Academy of Engineering, Alandi, Pune 412105, India
[2] Agasti Publishers, Talegaon Dabhade, Pune 410507, India
[3] Department of CSE/IT, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 411042, India
[4] Department of Computer Engineering, Pimpri Chichwad College of Engineering and Research, Ravet, Pune 412101, India
[5] Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune 411018, India
[6] Department of Information Technology, Vishwakarma Institute of Information Technology, Pune 411037, India

Corresponding Author Email: sandeep.pande@mitaoe.ac.in

https://doi.org/10.18280/ijsse.120305

Received: 14 January 2022
Accepted: 23 April 2022

Keywords:
data mining, fake news detection, regression analysis, machine learning, web scraping, data structuring

ABSTRACT

From the last few years, the use of social media has increased resulting into the rise of fake news and their spreading on a large scale. Recent political events have increased the spread of fake news. As seen by the widespread impact of the huge beginning of fake news, people are inconsistent in the absence of effective fake news detectors. This work has made an attempt to automate the fake news detection process by employing the logistic regression (LR) and latest and modified word embedding technique. In this paper, we worked on the fake news recognition mechanism for 2 different datasets, viz. dataset comprising online traditional news articles and news collected from a wide range of sources. The results are compared with long short-term memory (LSTM) and traditional machine and deep learning methods for both the datasets. It reveals that the traditional mechanism for attention does not function as expected. With the help of word2vec embedding, we modified the original attention mechanism, which is more effective in dealing with this issue. The proposed method is compared with several outstanding approaches and the results are presented. Our work outperforms these methods in many parameters. This approach has created a framework that captures various fake news indicators and classifies the news as genuine or fake and makes decisions.

1. INTRODUCTION

Nowadays, people are vastly relied on social networking platforms to get knowledge, information and current happenings as it provides the same in a quite new, fluent, speedier and portable manner. The content or news has significant influence on their day today life. It is even capable of changing their perspective, opinion or mindset. US election in 2016 is best example of it to describe. There arises a high chance to modify the view of person to exploit the opportunity in their favor. Sometimes media may handle the information differently to achieve someone's personal gain. Several activists, groups, or political parties may false propagate information on these social networking platforms. This led to development a new research domain known as fake news detection. The fake news is going to mold or rephrase people's judgement on particular issues, which are often aimed at pleasing individual agendum which majorly includes political intentions. The best-known way is to blacklist the unreliable sources and authors. It requires to develop an effective model or tool that considers more difficult cases where authors and sources publish fake news as these tools are convenient, in order to generate a more perfect understanding about the authenticity of the news. Further, the sentiment or emotion following the articles from news is a prime thing to take into account fake news. The ease with which people may use the Internet has resulted in an expeditious and unmanageable spread of various kinds of misinformation, such as rumor, spam opinion, fake news, deception, and forgeries. A news-article can be twisted in context of particular political agenda therefore, it might be one sided information. The threat of misinformation to democracy, justice, and public confidence has become a global issue, prompting an increase in research interest in both detecting it and combating its spread across the world. In the current Covid 19 pandemic situation, it has been noticed that several fake news and misbeliefs are spread in the social media regarding unnecessary medications, unnecessary precautions, false information regarding vaccination effects, and so on. The genuine news must be factual, impartial, not blaming or declarative. It's never easy to stop the spread of false information. The identification of information reliability is a critical topic in anti-misinformation. Nowadays, information's credibility fluctuates greatly between truthful, false, or of varying degrees of reliability. The spread of misinformation produces adverse effects in the society. Therefore, there is a need to propose the system where, not only the fake news can be identified on social media and internet but also the sources of such spread should be traced. This approach has developed a technique to overcome this problem.

Several approaches have been presented in this domain to identify the sources of fake news and genuineness of the news.

However, there exists some major research gaps in the existing systems such as false feature extraction, extensive feature vector sizes, poor pre-processing techniques, higher complexity and the larger training. Various trends are observed on social media over time. Therefore, there is always a need of an effective, versatile, and computationally efficient system to perform semantic analysis. In our approach we have focused on these major research gaps and presented a new system for depression detection.

The proposed approach involves two modules in which first module is news classification based on Kaggle dataset of fake news [1] whereas, the second module is designed to extract the web crawlers and scrapers [2] and label them as 0 and 1 and then measure the accuracy. This approach is aimed with finding and removing misleading web pages with the intention of preventing readers from misleading content. In order to achieve this goal, the approach employs some factors in deciding whether or not to classify a webpage as fake news. To achieve this, this approach implements two models. The input news is simultaneously given as an input to the classification web scraping module. In classification module preprocessing and data cleaning steps such as handling missing values and null values, lemmatization, tokenization, stop words removal are applied. Further the input is given to the LR classifier for classification. In the web scarping stage automated methods are applied to obtain huge data from website URLs related to input news. The data present on the web URLs is formless. It helps to get this formless data and store it in a proper form. This data is further analyzed to check genuineness of the input news. Both of these module works in parallel. To get improved results and accuracy, the results of both the classification and scraping module are combined with a properly plotted graph based on the experimental values. The overall training time of this approach is reduced due to the parallel execution of the modules. As per the obtained findings, the analysis here shows there is 97% accuracy for the fake news classifier. The proposed method has given prominent and satisfactory results. The user must download and install the tool on his/her computer before using its services. It is hoped that this technique will be compatible with the most commonly used browsers worldwide. Remainder of this paper is organized as: Section 2 outlines the state-of-the-art techniques reported in fake news detection. Methodologies used and the details of the implementation and measures undertaken to design this system is described in Section 3. Further, the working, requirements and results are discussed in depth in Section 4. Finally, Section 5 offers the concluding remarks.

2. RELATED WORK

This section outlines the recent works in detection of fake news. The following Figure 1 depicts the classification of several existing fake news detection techniques. Basically, there are 4 main types of fake news detection techniques viz. Data, Feature, Model and application-oriented. The data-oriented approach primarily focusses on psychological and temporal data. The Feature oriented approaches try to find out novel features from the news context or through social context. However, the model oriented one tries to fit some supervised, semi-supervised or unsupervised classifier to predict the genuineness of the news whereas, the application oriented one tries to predict the diffusion or intervention of the news. Current researches primarily focus on the use of social-features and speaker-information to improvise the overall quality of recognition. The study of Torabi and Maite [3] proposed a deep learning approach for detecting fake news that make use several features like sequential arrangement between 'n' users and 'm' news articles over time span. They not only generate fake news labels for classification but also reports the suspicious user's score. A method is proposed that obtains hoax information from the information through social network such as user, like count [4]. A stacked ensemble classifier is proposed by Zhou et al. [5] to tackle the issue of recognition of fake news. It is a fact of determining whether an article disagrees, agrees or just states a fact. Shu et al. [6] used Naïve Bayes to classify news from datasets of buzz feed. In addition to texts and social features, Figueira and Oliveira [7] and Zhang and Ghorbani [8] employed visual features like images and used convolutional neural network (CNN) and adversarial neural networks for detecting fake news respectively.

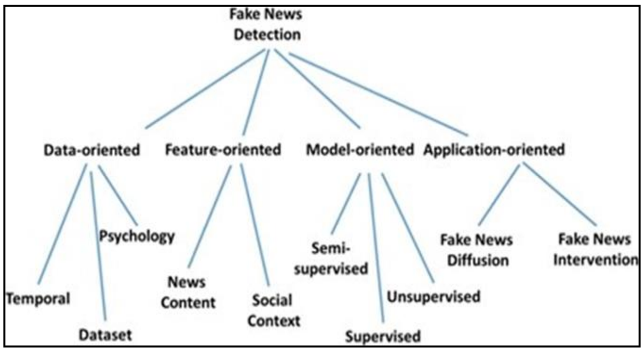


Figure 1. Different approaches to detect fake news

Yadollahi et al. [9] and Kwak and Cho [10] have presented a detailed review of existing fake news techniques. Aldwairi and Ali [1] have extracted novel features from the input and employed WEKA based classifiers (Bayes Net, Logistic, Random Tree, Naive Bayes) for fake news detection. Various researchers have used different classifiers such as TriFN [11], LSTM [12], Deep CNN [13], LR [14], Hierarchical Propagation [15], and SVM [16] for detecting the fake news. Pande and Chetty [17] have employed CapsNet in the equivalent domain.

Thorne et al. [18] proposed a stack of different classifiers for handling the fake news. It employed a ReLu activation based multilayer perceptron (MLP) with word2vec for handling headline and tf-idf for handling the article body, secondly, average word2vec for handling headlines and article body, tf-idf bigrams and unigram for handling article body. Then it used LR with L2 regularization and concatenation of word2vec for handling the headlines and article body with MLP and dropout. However just applying tf-idf is inefficient for fake news identification

Xu et al. [19] have used content understanding and domain reputations for detecting the fake news. They have believed on the popularity of domain, registrations for the domain and time duration for which the news exists. Whereas the research of Dong et al. [20] focusses on timely identification of fake news. In this work they have proposed a semi-supervised learning approach with two paths viz. supervised and unsupervised. These paths are realized with CNN.

Shrivastava et al. [21] have tried to resolve the issue of spreading fake messages on social media regarding the COVID-19 pandemic. A system of differential equations is used to develop the model. Verma et al. [22] developed a two-phase model termed as WELFake with word embedding on linguistic features to detect the for fake news.

Ghosh et al. [23-25] used text context and content information of the news for detecting fake news. Pan et al. [26, 27] employed CNN for effectively detecting the fake news. Umer et al. [28] have used combination of CNN and LSTM for fake news detection. However, these approaches require huge dataset with labelled data for training the model which results in labelling and time overhead. Therefore, this approach undertakes linear regression method to detect fake news.

For sophisticated text classification, the recent researches include Attention Mechanism [29], LSTM [30], IndRNN [31], Attention-Based Bidirected LSTM [32], Adversarial Training techniques [33], Hierarchical Attention Networks [34], various supervised Techniques like CNN [35], Random Multimodal Deep Learning (RMDL) [36]. The authors of [37] have examined how people perceive risk in the Pompeii Archaeological Park, with a focus on the emotional aspects and the help of a semantic analysis of the textual materials found on Twitter. These techniques have equivalent performances.

3. METHODOLOGIES

The motive of project is to identify a solution that may be used to detect and remove sites with fake news to help users to avoid being enticed by clickbait's. It is our good fortune that such solution could be discovered and it would be much helpful for technical corporations and readers who are plagued by fake news. The proposed idea solves problems related to fake news which includes the tool usage that finds and filters out the fake news from the result provided by a search engine or a feed from social networking news. The aim is to use AI to identify articles and claim that are likely to contain false or highly biased information.

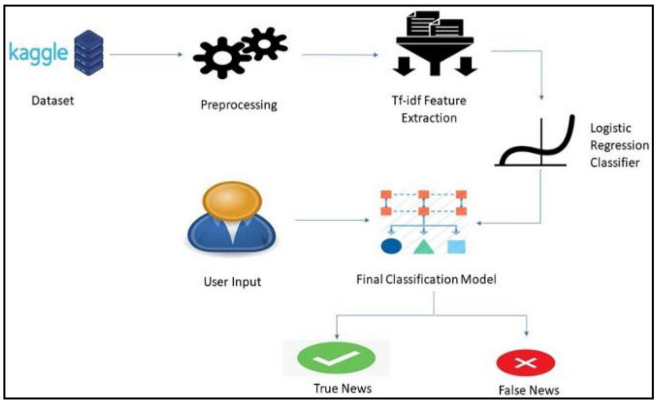

kaggle
Dataset
Preprocessing
Tf-idf Feature Extraction
Logistic Regression Classifier
User Input
Final Classification Model
True News
False News

Figure 2. System architecture

The challenge of fake news detection involves some of the following tasks: extracting and matching text-as it involves searching the several sources of data and then relating it to the input news, named entity recognition as it involves generation of token level fine-grained output, document, and entity-level sentiment analysis - as it involves the identification of the polarity, emotions negations, sarcasms, tone and bias of the sentences, document classification, stance classification, among others. The approach can be downloaded and attached to the browser or application used by the end user to receive news feeds. So that the application starts its work by using several methods like the methods related to the feature extraction so that we can verify the content we are seeing is valid one or false. This approach is implemented by integrating 2 modules: the web scraping and the LR module for detecting fake news as depicted in Figure 2.

3.1 Pre-processing

It involves the pre-processing steps needed to handle all the input texts and feeds. Initially, it reads the train, test and validation data and performs the processing such as tokenization and stemming. The exploratory data analysis such as response variable distribution and data quality checks like missing values or null values are then accomplished.

NULL Value: We will start data pre-processing with checking the null values. We need to tackle with missing values during data analysis phase. So, handling the missing values is most important aspect before starting to work on data. However, before doing anything about missing values, we need to understand the pattern of missing values occurring. The techniques are useful in the early phases of the analysis of exploratory data.

```
title 0
text 0
subject 0
date 0
Target 0
```

The following Figure 3 shows the heat map that indicates there were no null values present. So, no need to pre-process on this NULL values.

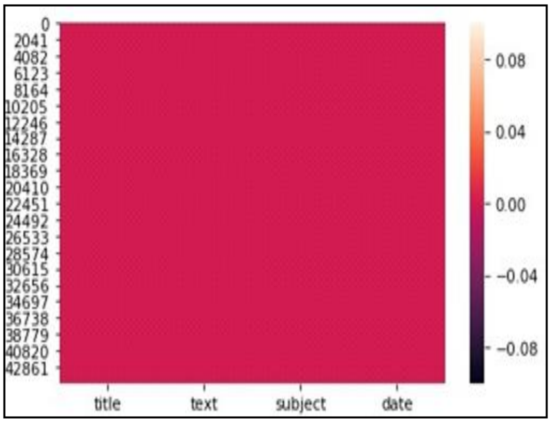


Figure 3. Heat map

Lemmatization: Lemmatization in linguistics is that the method of grouping along the inflected kinds of a word so that they are analyzed as one item, known by the word's lemma, or wordbook kind for example, run, runs, ran and running area unit kinds of an equivalent set of words that are unit connected through inflection, with run because of the lemma.

Tokenization: It is a method which is used to divide the text or knowledge in a very strict singular kind of entity. It suggests that, as an example, if we tend to tokenize a text into words, then every word is going to be the tiniest entity to be processed in our logic. If we have a tendency to tokenize sentences, then every sentence is going to be the tiniest entity (or token) to be processed. However, what is the necessity of tokenization? Some of you'll say, to word tokenize the text, we are able to simply split the text by whitespaces and take away all the punctuation marks. Let's see this via an associate degree example. Mr. O'Neill agrees that the boys' tales about Chile's investment aren't funny. The breaking of the above sentence into tokens (or words) produces the following different outcomes for the words for "O'Neill" -

o, neill
oneill
o'neill
o', neill
neill.
And in a similar way for aren't -
are, n't
arent
aren't
aren, t

As we can see, separating the terms merely by whitespace isn't going to produce the best outcomes for the North American nation. It crates ambiguity in further steps. These problems are area unit language specific, and it is very tough to write rules for area unit in these cases. The English language employs hyphenation for a variety of purposes, ranging from indicating word grouping following cacophonous up vowels in words to connecting nouns as names. It is simple to feel that the primary example ought to be considered one token (and is so a lot of normally engraved as basically coeducation), the last ought to be into separate words, which the center situation is uncertain. Handling hyphenates mechanically will therefore be complicated as it will either be like a classification downside, or will generally follow some empirical rules, such as allowing brief combined word prefixes, however, are no longer aggregated forms.

3.2 Logistic regression

It is one of the ML algorithms used for classification on the basis of the estimated probability of categorial dependent variable. The dependent variable in this classification algorithm is a variable of binary type that encode data as 0 (no, failure) or 1 (yes, success). Put differently, the LR model predicts P (Y = 0 or 1) as a function of X.

If the decision threshold is taken into account, then this algorithm becomes a classification method. Managing the threshold value is a very important part of LR and it depends on the classification problem. As per feature selection used for the dataset here, the best threshold value for this algorithm is 0.6. Further, 80% data part of the data is used for training and 20% data is used for testing on the LR classifier. It gives us mean score of 0.93 and best score of 0.94. The logit function that is trained by the BuzzFeed dataset is employed in this approach for training the model that forecasts the falsification possibility related to the news reported through the data filtering. The probability modelling of unreliability relating to our predictive variables is calculated using maximum likelihood estimation and logit transformation. Here, the basic version of linear regression equation using dependent variable considered in a relation to function that is used for LR:

$$h(y) = \beta 0 + \beta(Fake) \tag{1}$$

Our interest in LR rests in the possibility of outcome related variable (failure or success in relation to whether the news is fake or genuine). As discussed earlier, h () is the link function. It is specified using 2 factors: Success (p) probability and Loss Probability (1 - p). The value of p will meet the terms so the exponential form could be used as probability must be positive values. For any value of slope and dependent variable, the exponent of this equation can never be negative.

$$f = \exp\big(\beta 0 + \beta(Fake)\big) = e^{(\beta 0+\beta(Fake))} \tag{2}$$

where, a number greater is used to divide so that the probability would be < 1, Its simple form is as given below:

$$\begin{aligned} f &= \exp\big(\beta 0 + \beta(Fake)\big)/\exp\big(\beta 0 + \beta(Fake)\big) + 1 \\ &= e^{(\beta 0+\beta(Fake))/(\beta 0+\beta(Fake))+1} \end{aligned} \tag{3}$$

from above Eq. (3) we can write:

$$f = e^{y/1} + e^{y} \tag{4}$$

where, f is the success related probability it means it is the fake news. If p is the probability related to success, then, 1-p is the probability related failure that means the news is unaffected and written as:

$$q = 1 - f = 1 - \left(e^{y/1} + e^{y}\right) \tag{5}$$

where, failure probability is q, by dividing we get,

$$f/(1 - f) = e^{y} \tag{6}$$

Evaluating both sides using log is,

$$y = log(f/(1 - f)) \tag{7}$$

Eq. (7) is known as the link function. This function is intended to obtain a linear combination and change the values in terms of genuine and fake respectively. To build and train the classifier, a dataset with defined discriminants is used. The model related to classification is then used to evaluate suspicious news data and it responds with a percentage to determine exactly how false the data related to news is.

3.3 Web scraping

It is one of the automated methods applied to obtain huge data from web URLs. Following Figure 4 shows web scraping. As we can see that, the data present on the web URLs is formless. It helps to get this formless data and store it in a proper form. Following some ways by which scraping to the websites can be done:

1. Writing your own code.
2. APIs.
3. Online services.

When open web scraper code is used and run to the URL that you have copied, a request is sent to the server. The sever then sends the data and grant you to read the HTML or XML page, in the response of the request made by us. The code then, finds the data and extracts it, parses the HTML or XML page. In this approach our own code is used for effective collection of data and its genuineness from several websites and social media platforms.

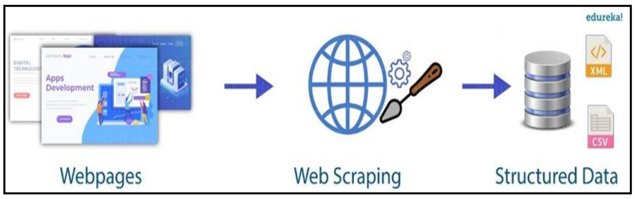


Figure 4. Web scraping

3.4 Combination of the scraping module and the results from the LR module

To get improved results and accuracy, the results of both the LR and scraping module are combined with a properly plotted graph based on the experimental values for the LR and scraping module. For further results generated by the modules, the graph is extrapolated and aggregated by averaging the results attained from the graph and it generates the output in the context of accuracy percentage for the web scraping module.

3.5 Sentimental analysis

Text-Cleaning: NLTK (natural language toolkit) tools are employed for the process of text cleaning. It assists to converts the raw text to a list of words. It obtains words through splitting of texts, by selecting character strings of alphanumeric form (a-z, A-Z, 0-9 and '_'). It then removes all punctuation marks such as quotes, commas as well as the whitespaces.

Term Frequency - Inverse Document Frequency (TF-IDF) Vectorizer: It is a method for computing a word in documents. Usually, the weight of an individual word is computed that represents the importance of the word in the document or news. Term frequency (TF) is the count of a particular word present in a document and is individual to every word and document. Whereas, Document frequency (DF) calculates the importance of document in the whole dataset. The only difference in TF and DF is that TF is used to count frequency of t terms in d documents, whereas the repetition count of term t in the document set N is DF.

$$df(t) = occurrence\ of\ t\ in\ documents \tag{8}$$

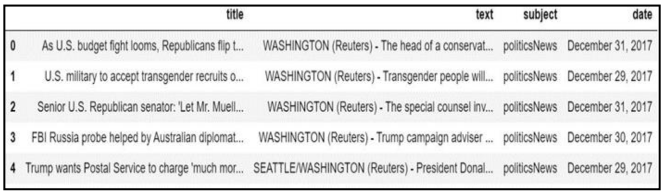
IDF is the reciprocal of the DF which helps to compute the term t showing informativeness. The value of IDF will be very less for the stop words which are most occurring in the raw text. The relative weightage TF-IDF can be computed as:

$$Tfidf(t) = N/df \tag{9}$$

4. RESULT AND DISCUSSION

This section initially discusses the analysis of the data set to explore the data and further it discusses performance of this approach. Counting the texts per class, count for the number of words per sentence is carried out first as statistical analysis. Then, the insights related to distribution of data are presented by using reduction in dimensions and the 2D plots of the data are generated. The dataset employed here is taken from Kaggle. The dataset consists of two files called as fake and real csv files. The data consists of 44898 rows and 4 columns. The below Figure 5 depicts how the dataset looks like.

In the following Figure 6 the plot of the number of true and fake records is plotted. It can be observed that the dataset has more fake instances than true.



	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

Figure 5. Dataset sample

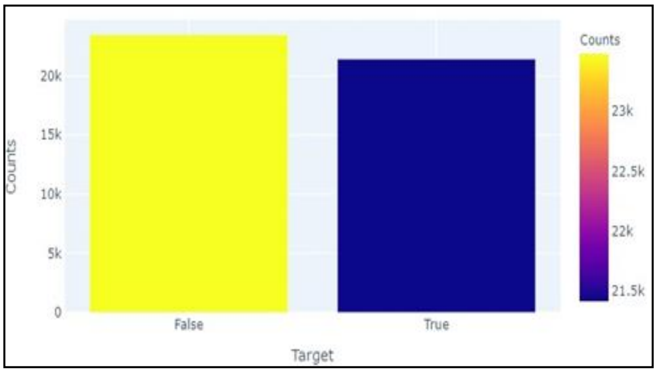


Figure 6. Plot of dataset as true or false based on news

Figure 7 shows topic wise distribution of the news data. From the plot it is clear that we have got a greater number of instances on political news. In order to take into account, the title in our accuracy prediction, we created an extra column that combines text and title. We will not do separate predictions on the title, since it might classify as e. g. fake news, weather the actual text with more explanation tells a real story. After pre-processing the representation of the dataset with respect to different categories is depicted in Figure 8.

Further, we have to add a prefix to each word with its type (Noun, Verb, Adjective,). For e.g.: I drink water => PRP-I, VBD-drink, NN-water. Also, after lemmatization it will be 'VB-drink NN-water', which shows the semantics and differentiate the purpose of the sentence. This will be useful for the classifier to differentiate between different types of sentences. The sample output after adding prefix is represented in Figure 9.

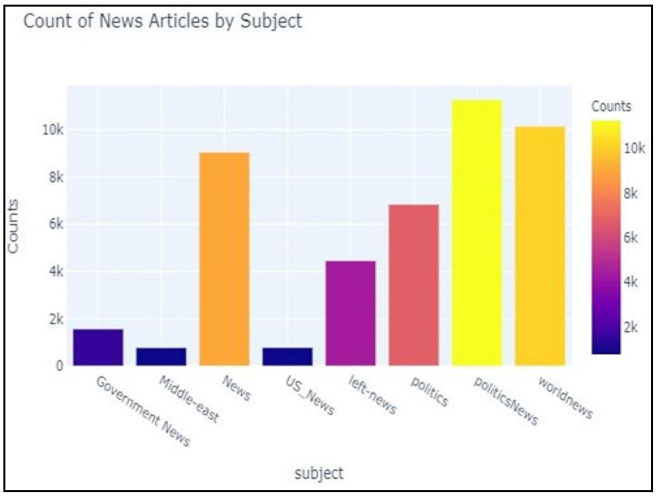
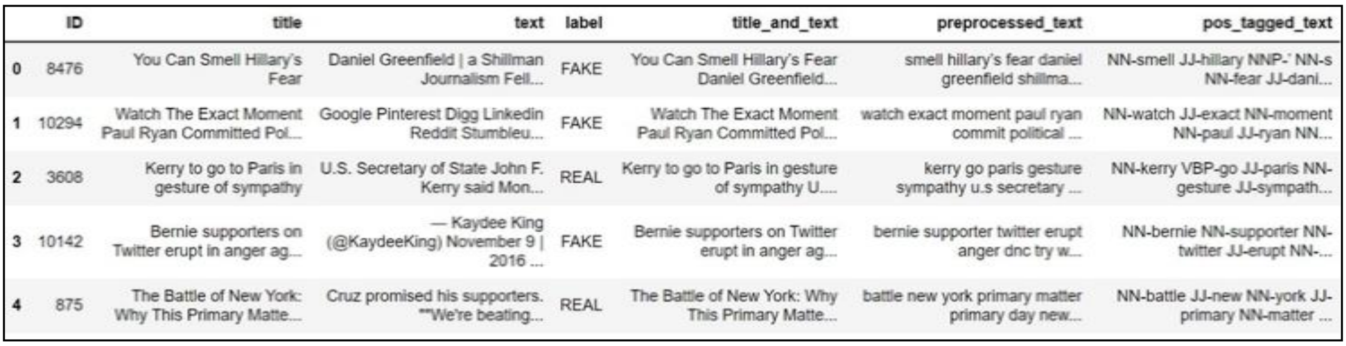
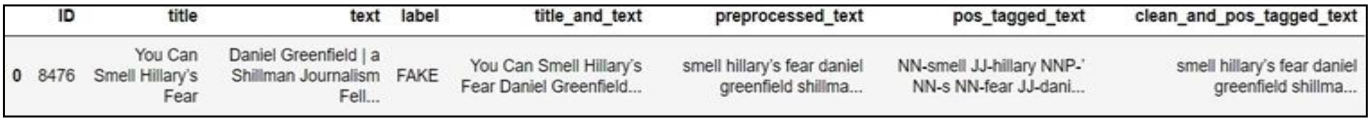


Figure 7. Topic wise distribution of news data



	ID	title	text	label	title_and_text	preprocessed_text	pos_tagged_text
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield | a Shillman Journalism Fell...	FAKE	You Can Smell Hillary's Fear Daniel Greenfield...	smell hillary's fear daniel greenfield shillma...	NN-smell JJ-hillary NNP-' NN-s NN-fear JJ-dani...
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE	Watch The Exact Moment Paul Ryan Committed Pol...	watch exact moment paul ryan commit political ...	NN-watch JJ-exact NN-moment NN-paul JJ-ryan NN...
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL	Kerry to go to Paris in gesture of sympathy U....	kerry go paris gesture sympathy u.s secretary ...	NN-kerry VBP-go JJ-paris NN-gesture JJ-sympath...
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9 | 2016 ...	FAKE	Bernie supporters on Twitter erupt in anger ag...	bernie supporter twitter erupt anger dnc try w...	NN-bernie NN-supporter NN-twitter JJ-erupt NN-...
4	875	The Battle of New York: Why This Primary Matte...	Cruz promised his supporters. ""We're beating...	REAL	The Battle of New York: Why This Primary Matte...	battle new york primary matter primary day new...	NN-battle JJ-new NN-york JJ-primary NN-matter ...

Figure 8. Dataset after pre-processing



	ID	title	text	label	title_and_text	preprocessed_text	pos_tagged_text	clean_and_pos_tagged_text
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield | a Shillman Journalism Fell...	FAKE	You Can Smell Hillary's Fear Daniel Greenfield...	smell hillary's fear daniel greenfield shillma...	NN-smell JJ-hillary NNP-' NN-s NN-fear JJ-dani...	smell hillary's fear daniel greenfield shillma...

Figure 9. Sample output after adding prefix

The TFIDF score of each term in a piece of text is initially calculated. Further, the text is tokenized into sentences and each sentence is then considered as a text item. These steps are also applied on the cleaned text and the concatenated POS_tagged text. In this step, the texts related news is distributed as the fake or genuine through classification techniques. This classification is executed through LR algorithms step by step. In first step, the TFs and count vectorizers are obtained and are taken as attributes related input for the particular classification model and the targeted attribute that is defined on it will work as the attribute specific to output. To combine the vectorizer count, classification and TF-IDF method together, this approach employs a pipeline method. It is handled by enabling a data sequence to be converted and together correlated in a single model that will go through testing phase to obtain the results. This step classifies the text related news through the model named LR and evaluates evaluation parameters for performance. Several stages of this approach such as scraping and classification are carried out in parallel that reduces the training time.

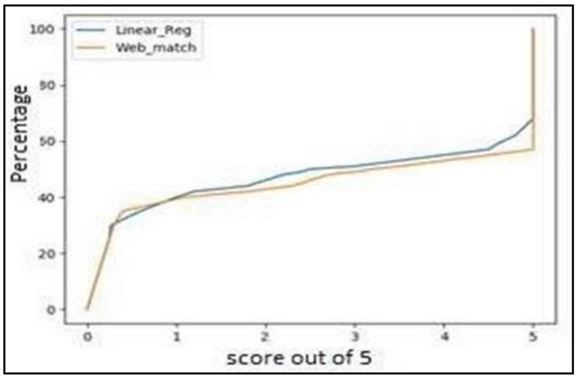


Figure 10. Results from LR and web scraping

The comparison between LR and web matching is done with respect to a score out of 5 metrics in percentage (%) and is given in Figure 10. Here, the scores 5 and 0 are the highest and lowest scores respectively. The evaluation metrics such as recall, precision, F1-score and accuracy are utilized to measure the performance [38, 39]. This approach's performance is compared with several modern semantic analysis methods known as SVM [16], NB [1], KNN [40], CNN [13] and LSTM [12]. Numerous iterations are carried out on the datasets with diverse samples. Then, the results are averaged and compared using the evaluation metrics mentioned earlier and shown in Table 1. The evaluation results reveal that this technique has outperformed in all metrics as compared to other techniques.

Table 1. Results comparison

Approach	Precision	Recall	F-Score	Accuracy
SVM	0.75	0.73	0.739	0.70
NB	0.92	0.91	0.915	0.92
KNN	0.73	0.71	0.72	0.86
CNN	0.94	0.95	0.945	0.96
Proposed	0.96	0.97	0.965	0.97

4.1 Confusion matrix

This model is capable of classifying news as fake or not. The throughput of the model is further evaluated on the basis of confusion matrix as shown in Table 2. Total 8980 sample are used in which model has given correct output of 8719 samples and 261 were incorrect. It makes our model 97.09% accurate. Here 120 instances are true negative, 4574 are true positive, 141 are false positive and 4145 are false negatives.

Table 2. Confusion matrix

Predicted Actual	Positive	Negative
Positive	4574	141
Negative	120	4145

5. CONCLUSION

As social media is continuously growing, more people easily get news through social media as compared to traditional forms of news mediums. Networking though online modes now used to spread distorted news, which may lead to significant or adverse effects on consumers side and wider to the community. Through this work, a technique is proposed wherein there is a combination of two different approaches for achieving improved accuracy in recognition of fake news. This approach uses effective data cleaning and categorization which helps in improved accuracy. The web scraping module is combined with classification module further improves the accuracy of the proposed approach. Both of these modules work in parallel therefore, the overall training time is reduced. As per the obtained findings, the analysis here shows there is 97% accuracy for the fake news classifier. The proposed method has given prominent and satisfactory results. However, it has also been suggested that new discriminators should be included in further research so that the accuracy of the fake news classification will be beyond 97%. Although the proposed approach performs better, still there exists a scope for some enhancements. In social media, verities of posts like images with some text, comments, videos, posts related to fake news, are generated in vast and are required to be handled effectively to get the truthfulness of the news or a post. A detailed classification needs to be considered as the recent news are sometimes partially true and partially fake. These are some of the future scopes of our work which we will try to address in near future.

REFERENCES

[1] Aldwairi, M., Ali, A. (2018). Detecting fake news in social media networks. Procedia Computer Science, 141: 215-222. https://doi.org/10.1016/j.procs.2018.10.171

[2] Singrodia, V., Mitra, A., Paul, S. (2019). A review on web scrapping and its applications. International Conference on Computer Communication and Informatics (ICCCI), pp. 1-6. https://doi.org/10.1109/ICCCI.2019.8821809

[3] Torabi, F., Maite, T. (2019). Big data and quality data for fake news and misinformation detection. Big Data & Society, 6(1): 1-12. https://doi.org/10.1177/2053951719843310

[4] Granik, M., Mesyura, V., (2017). Fake news detection using naive Bayes classifier. IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 900-903. https://doi.org/10.1109/UKRCON.2017.8100379

[5] Zhou, X., Reza, Z., Kai S., Huan, L. (2019). Fake news: Fundamental theories, detection strategies and challenges. Twelfth ACM International Conference on Web Search and Data Mining, pp. 836-837. https://doi.org/10.1145/3289600.3291382

[6] Shu, K., Mahudeswaran, D. Huan, L. (2019). FakeNewsTracker: A tool for fake news collection, detection, and visualization. Computational and Mathematical Organization Theory, 25(1): 60-71. https://doi.org/10.1007/s10588-018-09280-3

[7] Figueira, Á., Oliveira, L. (2017). The current state of fake news: Challenges and opportunities. Procedia Computer Science, 121: 817-825. https://doi.org/10.1016/j.procs.2017.11.106

[8] Zhang, X., Ghorbani., A. (2020). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57(2): 102025. https://doi.org/10.1016/j.ipm.2019.03.004

[9] Yadollahi, A., Shahraki, A., Zaiane, O. (2017). Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys (CSUR), 50(2): 1-33. https://doi.org/10.1145/3057270

[10] Kwak, J.A., Cho, S.K. (2018). Analyzing public opinion with social media data during election periods: A selective literature review. Asian Journal for Public Opinion Research, 5(4): 285-301. https://doi.org/10.15206/ajpor.2018.5.4.285

[11] Shu, K., Wang, S., Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 312-320. https://doi.org/10.1145/3289600.3290994

[12] Negi, S., Paul, B. (2017). Inducing distant supervision in suggestion mining through part- of-speech embeddings. arXiv preprint arXiv:1709.07403.

[13] Poria, S., Cambria, E., Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. Knowledge-Based Systems, 108: 42-49. https://doi.org/10.1016/j.knosys.2016.06.009

[14] Vedova, M., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., Alfaro, L. (2018). Automatic online fake news detection combining content and social signals. 2018 22nd Conference of Open Innovations Association (FRUCT), pp. 272-279. https://doi.org/10.23919/FRUCT.2018.8468301

[15] Jin, Z., Cao, J., Jiang, Y., Zhang, Y. (2014). News credibility evaluation on microblog with a hierarchical propagation model. 2014 IEEE International Conference on Data Mining, pp. 230-239. https://doi.org/10.1109/ICDM.2014.91

[16] Al-Ash, H.S., Wibowo, W.C. (2018). Fake news identification characteristics using named entity recognition and phrase detection. In 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 12-17. https://doi.org/10.1109/ICITEED.2018.8534898

[17] Pande, S., Chetty, M. (2019). Bezier curve based medicinal leaf classification using capsule network. International Journal of Advanced Trends in Computer Science and Engineering, 8(6): 2735-42. https://doi.org/10.30534/ijatcse/2019/09862019

[18] Thorne, J., Chen, M., Myrianthous, G., Pu, J., Wang, X., Vlachos, A. (2017). Fake news stance detection using stacked ensemble of classifiers. In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism, pp. 80-83. https://doi.org/10.18653/v1/W17-4214

[19] Xu, K., Wang, F., Wang, H., Yang, B. (2020). Detecting fake news over online social media via domain reputations and content understanding. Tsinghua Science and Technology, 25(1): 20-27. https://doi.org/10.26599/TST.2018.9010139

[20] Dong, X., Victor U., Qian, L. (2020). Two-path deep semisupervised learning for timely fake news detection. IEEE Transactions on Computational Social Systems, 7(6): 1386-1398. https://doi.org/10.1109/TCSS.2020.3027639

[21] Shrivastava, G., Kumar, P., Ojha, R., Srivastava, P., Mohan S., Srivastava, G. (2020). Defensive modeling of fake news through online social networks. IEEE Transactions on Computational Social Systems, 7(5): 1159-1167. https://doi.org/10.1109/TCSS.2020.3014135

[22] Verma, P., Agrawal, P., Amorim I., Prodan, R. (2021). WELFake: Word embedding over linguistic features for fake news detection. IEEE Transactions on Computational Social Systems, 8(4): 881-893. https://doi.org/10.1109/TCSS.2021.3068519

[23] Ghosh, S., Shah, C. (2018). Towards automatic fake

news classification. Proceedings of the Association for Information Science and Technology, 55(1): 805-807. https://doi.org/10.1002/pra2.2018.14505501125

[24] Liu, Y., Wu, Y.F. (2018). 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, 32(1): 1-8. https://ojs.aaai.org/index.php/AAAI/article/view/11268.

[25] Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explor. Newslett., 19(1): 22-36. https://doi.org/10.1145/3137597.3137600

[26] Pan, J.Z., Pavlova, S., Li, C., Li, N., Li, Y., Liu, J. (2018). Content based fake news detection using knowledge graphs. In International Semantic Web Conference, pp. 669-683. https://doi.org/10.1007/978-3-030-00671-6_39

[27] Stein, R., Jaques, P., Valiati, J. (2019). An analysis of hierarchical text classification using word embeddings. Information Sciences, 471: 216-232. https://doi.org/10.1016/j.ins.2018.09.001

[28] Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G., On, B.W. (2020). Fake news stance detection using deep learning architecture (CNN-LSTM). IEEE Access, 8: 156695-156706. https://doi.org/10.1109/ACCESS.2020.3019735

[29] Tuan, N.M.D., Minh, P.Q.N. (2021). Multimodal fusion with BERT and attention mechanism for fake news detection. In 2021 RIVF International Conference on Computing and Communication Technologies (RIVF), pp. 1-6. https://doi.org/10.1109/RIVF51545.2021.9642125

[30] Bahad, P., Saxena, P., Kamal, R. (2019). Fake news detection using bi-directional LSTM-recurrent neural network. Procedia Computer Science, 165: 74-82. https://doi.org/10.1016/j.procs.2020.01.072

[31] Li, S., Li, W., Cook, C., Zhu, C., Gao, Y. (2018). Independently recurrent neural network (INDRNN): Building a longer and deeper RNN. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5457-5466. https://doi.org/10.1109/CVPR.2018.00572

[32] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2: 207-212.

[33] Miyato, T., Dai, A.M., Goodfellow, I. (2016). Adversarial training methods for semi-supervised text classification. arXiv:1605.07725.

[34] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E. (2016). Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480-1489. https://aclanthology.org/N16-1174.pdf.

[35] Chen, Y.H. (2015). Convolutional neural network for sentence classification. UWSpace. http://hdl.handle.net/10012/9592.

[36] Kowsari, K., Heidarysafa, M., Brown, D.E., Meimandi, K.J., Barnes, L.E. (2018). Rmdl: Random multimodel deep learning for classification. In Proceedings of the 2nd International Conference on Information System and Data Mining, pp. 19-28. https://doi.org/10.1145/3206098.3206111

[37] Garzia, F., Borghini, F., Bruni, A., Lombardi, M., Mighetto, P., Ramalingam, S., Russo, S.B. (2020). Emotional reactions to the perception of risk in the Pompeii Archaeological Park. International Journal of Safety and Security Engineering, 10(1): 11-16. https://doi.org/10.18280/ijsse.100102

[38] Pande, S.D., Chetty, M.S.R. (2018). Analysis of capsule network (Capsnet) architectures and applications. J Adv Res Dynam Control Syst, 10(10): 2765-2771.

[39] Mungase, R.G., Shinalkar, S.G., Pagare, S.S., Pansare A.B., Pande, S.D. (2020). Assessment of image annotation techniques and recommendation of neural network. Multidisciplinary Journal of Research in Engineering and Technology, 7(2): 1-12.

[40] Kesarwani, A., Chauhan, S., Nair, A. (2020). Fake news detection on social media using k-nearest neighbor classifier. International Conference on Advances in Computing and Communication Engineering (ICACCE), pp. 1-4. https://doi.org/10.1109/ICACCE49060.2020.9154997