Methodologies For Fake News Detection Using Natural Language Processing and Machine Learning

SHAMLA MANTRI[1], GAYATRI GATTANI[2], SARVESH DHAPTE[3], YASH JAIN[4]
[1]*Associate Professor, School of Computer Engineering and Technology, Dr. Vishwanath Karad's MIT World Peace University, Pune, India*
[2, 3, 4]*Student, School of Computer Engineering and Technology, Dr. Vishwanath Karad's MIT World Peace University, Pune, India*

Abstract: ***Having easy access to the internet and an intuitive user interface has made e-reader's growth extremely tremendous. These led to the gradual increase of fake news activity over social media and other websites. Using NLP an expeditiously emerging method of detecting fake content with help of machine learning algorithms is done. As part of this paper, we provide a recapitulation of the methods for collecting and classifying fake news, as well as a discussion of future directions for research in this area. In this experiment, data preprocessing is a first step, where data is created and transformed in a format used to model training. That preprocessed data is then programmed for feature extraction. Next a pipeline is created for all ML algorithms namely, Naive Bayes, SVM, Logistic Regression, KNN, LGBM Classifier, Random Forest. An analysis of all algorithms' performance is conducted in a comparative study. Detailed analyses of each model are provided, with an emphasis on its performance. According to our experiment, Random Forest is an overfitted model for this purpose. With an accuracy of 99.09%, the SVM classifier performs best followed by the LGBM classifier with a 99.79% accuracy.***

Indexed Terms: News, Fake News identification, NLP, Machine Learning, Social media, Information legitimacy.

I. INTRODUCTION

In today's digital era, over 50% of readers are e-readers. Fast response time, low cost, and a large data storage capacity makes the internet a popular source of news and information. Today, it is one of the most important sources of knowledge in people's lives. As the term implies, fake news is inaccurate and specious information masquerading as news. Besides hurting people's feelings, it may cause damage to their image, mislead public opinion, or cause major conflicts. In some cases, authors and websites lurk at people in order to monetize their content or gain media coverage by using their influence and clickbait. So, it is reasonable to consider them as one of the most significant menaces to community and confraternity. As the Internet and social communication sites have become more widely available, the rate of generating this fake news has increased dramatically. This news is produced in bulk which makes it difficult to detect in real-time analysis.

Since the world population has been expanding on a massive scale from the past decade, it is robustly important for people to fathom actual authenticated and hoax news. Using natural language processing and machine learning, we examine different methods for detecting fake news, and provide direction for something impeccable coming soon. 59% of people are concerned about the impact fake news has at work. This can be incredibly damaging to business [1]. Facebook and Twitter, two social media platforms that debuted in 2004 and 2006, respectively, are among those attempting to lower these barriers in the false news detection sector and are expanding to function intelligently with the most legitimate material feasible [2].

It is now simpler to discern between fake and legitimate news, courtesy to machine learning algorithms. Using Natural Language Processing facilitates text-based work. Machine learning and natural language processing are the two methods for text-based detection that are most frequently utilised since they automate the model-building process and offer real-time predictions. Our paper synthesizes and analyzes all the previous work done on fake news detection by leveraging these emerging technologies. In turn, we choose the ideal model, feature generating method, and accuracy checking criteria after we analyze their performance.

II. LITERATURE SURVEY

A. Fake News:

Fake news is often used as a source in order to destroy the notoriety of a person or entity, or make

money through marketing and other forms of advertising. It also uses unintentional and unconscious mechanisms and also by the users that are bulged in the high profile that use it for their unfavourable intentions.

Industry	% of Fake news
Politics	71
Communalism	22
International	16.7
Education	15.6
Crime	10
Economy	6.7
Historical	4
Entertainment	3
Health	3
Sports	1
Others	5.6

Table I. Distribution of fake news with respect to different industries [3].

B. Fake News Detection :

Throughout the world, fake news is being combated, but multiple approaches and analytics are available to combat it, as well as identify what types of fake news exist. It is imperative to regulate social media and web search engines both by self-regulation and by law. In addition to natural language processing, machine learning algorithms are often employed as classification tools, such as SVM, Logistic Regression, Random Forest, Naive Bayes, ANN, Gradient Boosting, and Decision Trees.

Year	Author	Classification/Technique	Dataset	Highest Accuracy
2018	Monther Aldwairi, Ali Alwahedi	ML techniques provided by WEKA tool	-	Logistic Regression=99.4%
2019	Junaed Younus Khan, Md. Tawkat Islam Khondaker, Anindya Iqbal, Sadia Afroz2	Traditional Machine Learning Models, Neural Network-Based and Deep Learning Models	Liar, Fake and Real news Kaggle, Combined Corpus	Naïve Bayes, Bi-LSTM, C-LSTM=95 %
2019	Anjali Jain, Harsh Khatter AvinashShakya	Naive Bayes, Support Vector Machine (SVM), NLP	-	93.50%
2020	Pranav Ashtaputre, Ashutosh Nawale, Rohit Pandit, Savita Lohiya	Logistic Regression, Multinomial Naïve Bayes, Term Frequency-Inverse Document Frequency (TFIDF), Count Vectorizer	Reddit	Logistic Regression=85%
2020	Abdulaziz Albahr, Marwan Albahar	Naive Bayes, Random Forest, Artificial Neural Network, Decision Trees	Liar	Naive Bayes=99 %
2021	Pragnesh Bugade, Pooja Sarode, Tanve	Logistic Regression, Decision Tree, Random Forest	-	Decision Tree=99.6 %

	e Pimple			
2021	Dr. S. Rama Krishna, Dr. S. V. Vasantha, K. Mani Deep	Logistic Regression, Support Vector Machine, Naïve Bayes model, Random Forest, Decision tree	Liar	Random Forest=65.6%
2021	Mayank Singh Rawat, Aviral Srivastava, Shubh Aggarwal	Tfidf Vectorizer and Passive-Aggressive Classifier	7796× 4 news set	92.82%

Table II. Some previously used techniques in news detection and its highest accuracy.

A. Existing Strategies

With internet accessibility becoming increasingly ubiquitous, e-readers are growing exponentially. The proliferation of data has been found to be directly related to the commotions networks and information transmission applications. With the increasing use of digitalization in reading, we are also seeing an increase in fake content. Through machine learning, numerous studies have been conducted to identify fake and real news.

As detection strategies, there have been developed three approaches, namely knowledge-primarily-based, style-primarily-based, and visual-primarily-sourced detection on social networking platforms. With Tf Idf vector feature extraction, a passive-aggressive classifier achieved 92.82% accuracy on the dataset "news.csv". Large datasets can be handled well by the Passive-Aggressive Classifier. The results in this experiment were evaluated using a confusion matrix and a simple accuracy score [4]. Term frequency/inverse document frequency (TF-IDF) approaches are methods for discovering how important and unique each word within a document is. They contemplate the relevance of words, categories, and documents [5].

In [6], Natural language processing has been used to develop a machine learning model for identifying fake news. The authors proposed scraping articles from the "Subreddit" platform using PushShift API and creating a visualization based on the data obtained. Using GridsearchCV techniques and pipelines enabled automatic identification of the foremost co-occurrence of model and feature vectors. The conclusion was that combining Logistic Regression and Multinomial Naive Bayes algorithms along with Tf-IDF and count vectorizer feature extraction methods offered them the finest results. The highest rate of detection accuracy was for Logistic Regression with Tf-IDF at approximately 85%. An additional feature for self-confirmation where a user can himself check the validation of news, they consolidated the Selenium Webdriver.

There is a concept of "Semantic Fake News Detection" that seeks to reveal the news that is false by analyzing analogous attributes such as sentimentality, accuracy, and veracity in [7]. Their methodology began with the assemblage of metadata, followed by the evocation of relations, and embracement of embedding to classifiers. The pipeline that created the metadata was based on Natural Language processing that incorporated the named entity recognition, named entity links, and sentiment analysis based on the accumulated sentence level polarity. This experiment employed the "Liar" dataset and 5 Deep Learning algorithms -- CNN, BasicLSTM, BiLSTM, GRU, and CapsNetLSTM. According to their findings, the addition of semantic features to Deep Learning models increased accuracy by 5 to 6%.

There are also ensemble techniques that can be applied to identify right and wrong news based on linguistic features. In [8], linguistic inquiry and word count were used for identifying textual features from articles. The LIWC tool helps in gaining 93 varieties of characteristics from any given text. They used three datasets [9,10,11] that comprised the data of almost all the categories. The textual linguistic features containing stop words, frequency of words, punctuation, formal-informal words, functional words are converted to numerical

form to use as an input criterion for the model. To prevent the model from underfitting or overfitting, each algorithm is trained several times with different parameters using GridSearch. Ensemble methods of working like voting have been investigated to assess the performance on the diverse datasets. In which, log regression and random forest are combined with KNN and logistic regression is combined with linear SVM and classification and regression trees (CART). Other ensembling strategies like bagging and also two boosting methodologies were used. Evaluation of the model was performed using Confusion matrix, cross-validation technique.

Some authors proposed a work that can be brought into play by the users to identify and refine sites that contain deceptive information or news. While detection, the features like keywords in different languages, numerical starting title, case sensitive words, punctuation frequency count, user spent time on the site makes a great difference. The proposed methodology is based on the syntactical structure of links. The use of algorithms such as Bayes Net, Logistic Regression, Random Forest Tree, and Naive Bayes is done. They concluded that 99.4% accuracy was achieved using the Cross-Validation score methodology and the Logistic Regression algorithm [12].

In [13], the study used Signal Media data as well as OpenSources.co's database of almost 11000 sources. The feature extraction was done using bi-grams tf-idf methodology and probabilistic context-free grammar. The algorithms like Random Forest, SVM, Bounded Decision Tree, Gradient Boosting, and Stochastic Gradient Descent. As a means of speeding up the generation of features , the spacy python was used for natural language processing tasks instead of the nltk package. Stochastic Gradient Descent model based on probabilistic context-free grammar achieves an impressive 77.2% accuracy.

The authors in [14] suggested a system that can accept the input from the user as a statement or the link to an article to identify if it is fake or not. The system they built detects wrong content on the basis of stance detection with the help of LSTM. The data is collected from various news sources and websites and is preprocessed with the use of Microsoft Azure and IBM natural language processing. MLP is used as a classifier here. Fake news types are discussed in [15]: visual, stance, post-based, user-based, style-based, knowledge-based. They revised the classification of detection methods as the linguistic basis, non-text cue methods, clustering, content-cue methods, and predictive modeling. Rapid dissemination has been described in the form of click baiting. Algorithms were developed based on inductive logic, content-based logistic regression, and boolean label crowdsourcing.

In [16], the authors confronted the following problems, which may be useful to aspiring researchers in the future: Multi-Modal Dataset, Multi-Modal Verification Method, Source Verification, Author Credibility Check. They also compared different datasets, including BuzzFeedNews, LIAR, CREDBANK, and PHEME, with features classified as Content-based: linguistic and visual and Context-based: user, post, and network. In [17], the authors validated the datasets using public posts, posts' likes in different facebook pages. Based on Facebook API data, the authors assembled a text corpus of the actual text content of the post, the title of the post if the link was shared, and the title of the post when the link was shared.They established a cutoff point and categorised the postings according to their content and social media strategies. The authors in [18], focused their attention on clickbait. They discussed methods that can detect clickbaits from lexical choices to complex languages.

Authors in [19], proposed an experiment which used Machine Learning supervised classification algorithms like Decision Tree, Random Forest, SVM, Naive Bayes, KNN, XGboost with Tfidf and count vectorizer. Applying a set of these classification algorithms was the major goal in order to create a scanner for false news identification. In [20], The authors integrated the factors to model tri-relationship and give a semi-supervised discovery frame in the end. Rich auxiliary information is provided by the user engagements on social media that helps in the detection of fake news. Credibilities of the users and their shared new pieces are considered to inherit a relationship.

III. METHODOLOGY

For this experiment, the Fake News Detection dataset from Kaggle was utilised. The comparative analysis was done for Machine Learning algorithms,

namely Logistic regression, Naive Bayes, Support Vector Machine, LightGBM(Light Gradient Boosting Machine), Random Forest, and KNN.

Preparing the dataset was the first step because analysing the raw data was not practical. Data preprocessing is essentially the act of turning unclean data into clean data. In this first step, superfluous columns were removed, missing data was handled, and the data was scaled appropriately. Further EDA (Exploratory Data Analysis) was carried out, in order to analyse data through visual techniques.

Transforming features text into feature vectors and converting to lowercase was carried out with the help of TF-IDF (Term Frequency - Inverse Document Frequency). With the help of TF-IDF, quantification of the important terms could be done, which is widely used in text mining.

The dataset was trained and tested on the following Machine Learning algorithms :

1. Logistic Regression:

Classification issues are addressed by the use of logistic regression. The dependent variable is modeled using a logistic function. The dependent variable has a dialectical character. It classifies unidentified recordings quickly and trains quickly.

2. Naive Bayes:

Bayes theorem is the foundation of the Naive Bayes classifier. Every pair of features being classified should be independent of one another according to this principle. Additionally, it makes the assumption that each attribute contributes equally and independently to the final classification.

3. Support Vector Machine:

SVM also comes under supervised learning. Essentially, it identifies an n-dimensional space hyperplane that categorises the data points with clarity. The amount of features affects how big the hyperplane is. Because it just takes a portion of the support vector training points from the decision function, it performs well in high-dimensional settings and requires little memory.

4. LightGBM:

While LightGBM (Light Gradient Boosting Machine) breaks trees into individual leaves, boosting algorithms generally create trees level by level. It chooses the leaf with the highest delta loss. While the model may get more sophisticated and the likelihood of overfitting may increase, the leaf-wise approach has less loss than level-wise techniques. Thus, fake news detection is carried out more quickly and with extremely little runtime memory utilisation, making this approach unique.

5. Random Forest:

Subsets of data are used to generate random forests, and the results are based on average or majority ranking. It can handle datasets with categorical variables and handles the overfitting issue as well.

6. KNN:

The KNN algorithm, which uses "feature similarity" to anticipate the values of incoming data points, assigns a value to a new data point based on how much it resembles the points in the training set. The model uses a k value of 5, which KNN is effective even with enormous datasets.

IV. RESULTS AND DISCUSSION

A description of the results of our proposed experiment is provided in this section. In addition to implementation, a subtle comparison of different machine learning algorithms is demonstrated in the proposed research to demonstrate its eminence. As a part of classification of fake news, preparing the dataset was the first step because analysing the raw data was not practical. Next, transforming features text into feature vectors and converting to lowercase was carried out with the help of TF-IDF (Term Frequency - Inverse Document Frequency). A proposed algorithm is used to classify news as fake or not-fake in the final stage. Data in this experiment is used for training the model 80% of the time, and test data 20% of the time.

A comparison of the different algorithms used in the experiment above is based on accuracy and precision. For comparing algorithms, other metrics are calculated, including f1-score, error rate, and recall. Following Table 3 shows the complete evaluation metrics for all the algorithms proposed in this experiment.

Algorithm	Accuracy	Precision	Recall	F1-score
Logistic	94.45	90.9	92.59	91.74

Regression	%	%	%	%
Naive Bayes	89.93 %	79.2 %	97.3 %	87.33 %
Random Forest	99.99 %	86.81 %	94.03 %	90.28 %
SVM	99.09 %	93.51 %	93.6 %	93.55 %
KNN	58.01 %	91.54 %	11.37 %	20.23 %
LGBM	97.79 %	92.91 %	94.6 %	93.75 %

Table III. Evaluation metrics for all algorithms in proposed experiment

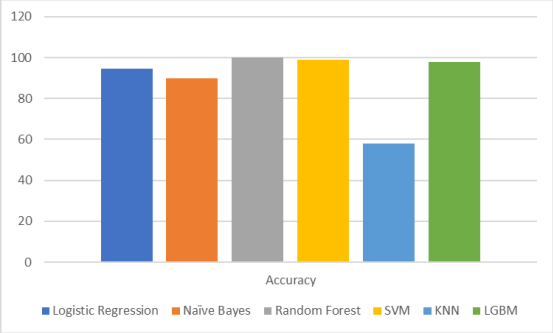

Fig I. Accuracy for all algorithms

Figure 1 shows that KNN has the lowest accuracy while Random Forest Classifier has the best accuracy, followed by SVM and LGBM Classifier.

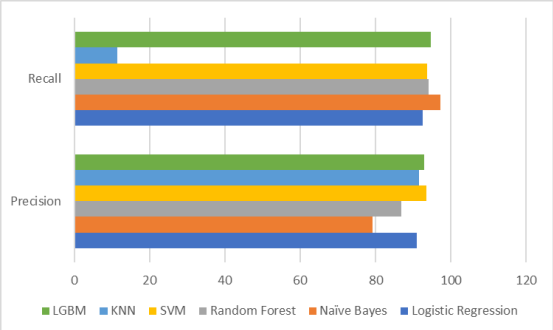

Fig II. Recall and Precision of all algorithms

From fig 2, it can be seen that, even if accuracy of Random forest is highest, its recall and precision is very low, that shows that Random forest model is overfitted.

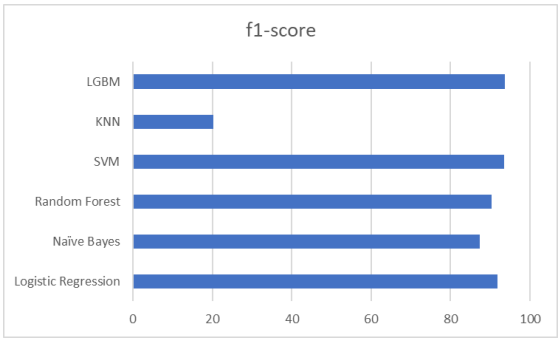

Fig III. F1 score for all algorithms

Fig 3 shows the f1-score for all algorithms proposed. F1 score is the harmonic mean of precision and recall to decide which algorithm performs better. From Fig 3, it can be seen that LGBM Classifier has highest F1-score of 93.75% followed by SVM that has F1-score of 93.55%. Thus, considering all the above results, we can conclude that though Random forest has highest accuracy but its precision and recall scores show that model is overfitted. Support Vector Machine followed by LGBM are good classifiers for the purpose of fake news classification. SVM is correct when there are two classes because it determines the optimum hyperplane that separates each class into its component components. SVM hence performs better in this scenario. Trees are divided by the LightGBM (Light Gradient Boosting Machine) leaf by leaf. It chooses the leaf with the highest delta loss. Thus it is memory efficient and gives good accuracy. KNN model's accuracy is observed as 58.01%, which is the lowest. One of the reason for this low accuracy can be KNN's sensitivity to the scale of data.

CONCLUSION

In today's digital era, over 50% of readers are e-readers. The review in this paper focuses on earlier work in the fake news detection system utilising machine learning and natural language processing as well as the drawbacks of false news. Supervised Machine learning algorithms are commonly used for this purpose. The experiment's main objective is to classify the news as fake or not on the basis of author and title details and compare the different algorithms. As the first step of the experiment, the dataset is preprocessed using different NLP techniques like lemmatization, tokenization and sequencing. Further, for feature extraction Tf-Idf vectorizer is used. Afterwards, a pipeline for machine learning algorithms Logistic Regression, Naive Bayes, Random Forest, SVM, KNN, LGBM

Classifier is created. Comparative study is done where performance of all algorithms is examined. It is seen that Random Forest is an overfitted model for this purpose. LGBM classifier comes in second with an accuracy of 97.79%, followed by SVM with a performance of 99.09%.

Future Work

The detection of fake news is an evolving topic for research. There is a requirement to find external features that can be added to the used techniques to not only speed up the automation of identification for real time analysis but also increase the accuracy of prediction. Deep Learning algorithms can be applied and compared with other algorithms to come up with finest accuracy and less overfitting model. Automated web extension tools can be developed that can be used by readers so they can check the validation of the content they read in real time.

REFERENCES

[1] "How to spot real and fake news", MindTools, 2021.

[2] Nathaniel.Hoy2, Theodora.Koulouri, 'A Systematic Review On The Detection Of Fake News Articles', ACM-TIST, 2021

[3] Rubal Kanozia, Ritu Arya, Satwinder Singh, Garima Ganghariya, Sumit Narula, "A Study On Subject Fake News Matter, Presentation Elements, Tools of Detection, and Social Media Platforms in India" , Asian Journal for Public Opinion Research, Volume 9, Issue 1, 2021.

[4] Mayank Singh Rawat , Aviral Srivastava , Shubh Aggarwal, "Detection of Fake News using Machine Learning", International Journal of Engineering and Applied Physics (IJEAP) , Vol. 1, No. 2, May 2021, pp. 205~209.

[5] Yun-tao, Z., Ling, G. & Yong-cheng, W. "An improved TF-IDF approach for text classification", J. Zheijang Univ.-Sci. A 6,2005,pp 49–55 .

[6] Pranav Ashtaputre, Ashutosh Nawale, Rohit Pandit, Savita Lohiya, "A Machine Learning Based Fake News Content Detection Using NLP" , International Journal of Advanced Science and Technology, Vol. 29, No. 7, (2020), pp. 11219-11226.

[7] Adrian M. P. Bra¸ Soveanu and R˘azvan Andonie, "Semantic Fake News Detection: A Machine Learning Perspective" , Springer Nature Switzerland AG 2019, IWANN 2019, LNCS 11506, pp. 656–667.

[8] Iftikhar Ahmad , Muhammad Yousaf, Suhail Yousaf , and Muhammad Ovais Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods", 17 October 2020, Hindawi Complexity Volume 2020, Article ID 8885861, 11 pages.

[9] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," Security and Privacy, vol. 1, no. 1, 2018.

[10] Kaggle, Fake News, Kaggle, San Francisco, CA, USA, 2018, https://www.kaggle.com/c/fake-news.

[11] Kaggle, Fake News Detection, Kaggle, San Francisco, CA, USA, 2018, https://www.kaggle.com/jruvika/fake-news-detection.

[12] Monther Aldwairi, Ali Alwahedi, "Detecting Fake News in Social Media Networks" , The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2018),pp 215–222.

[13] Shlok Gilda, "Evaluating Machine Learning Algorithms for Fake News Detection" , 2017 IEEE 15th Student Conference on Research and Development, pp 110-115.

[14] Dr. Rachna Somkunwar, Anil Kumar Gupta, Faizan Shikalgar, Shubham Maral, Arpit Paithankar, Rohan Kapdi, "Fake News Detection: A Survey" ,IJIRT, Volume 7 Issue 12, May 2021, pp 664-669.

[15] Syed Ishfaq Manzoor, Dr Jimmy Singla, Nikita, "Fake News Detection Using Machine Learning approaches: A systematic Review" , Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART, pp 230-234.

[16] Shivam B. Parikh and Pradeep K. Atrey, "Media-Rich Fake News Detection: A Survey", 2018 IEEE Conference on Multimedia Information Processing and Retrieval, pp 436-441.

[17] Della Vedova, M. L., Tacchini, E., Moret, S.Ballarin, G., DiPierro, M., & de Alfaro, L.

(2018, May). Automatic Online Fake News Detection Combining Content and Social Signals. In 2018 22nd Conference of Open Innovations Association (FRUCT) (pp. 272-279). IEEE.

[18] Chen, Y., Conroy, N. J., & Rubin, V. L. (2015, November). Misleading online content: Recognizing clickbait as false news. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection (pp. 15-19). ACM.

[19] Z Khanam , B N Alwasel , H Sirafi and M Rashid, "Fake News Detection Using Machine Learning Approaches" , IOP Conference Series: Materials Science and Engineering, pp 1-13.

[20] K. Shu, S. Wang, and H. Liu, "Exploiting Tri-Relationship for Fake News Detection," arXiv:1712.07709 [cs], Dec. 2017.