Comprehensive Fake News Detection Using Machine Learning Approaches

Puja Rarhi[1], Sudip Mishra[2], Soumya Kanti Pramanik[3]
[1]Department of Cyber Science & Technology, Brainware University, Ramkrishnapur Road, Barasat, 700125, West Bengal, India. Email:pujapersonal03@gmail.com
[2]Department of Computer Science & Engineering(Data Science), Haldia Institute of Technology, Hatiberia, ICARE Complex, Haldia , 721657, West Bengal , India. Email:mishrasudip02@gmail.com
[3] Department of Computer Science & Engineering, Camellia Institute Of Technology, Digberia, Badu Road, Near NSG Hub, Madhyamgram, Kolkata, 700129, West Bengal, India. Email: cse.soumyakanti@gmail.com@gmail.com

ABSTRACT: Fake news on social media and other platforms is widely disseminated and a major reason for concern because of its potential to have detrimental effects on society and the country. Its detection has already been the subject of much research. The argument made in this paper is analysis of the literature on fake news detection and investigates the best traditional machine learning models to develop a product model with supervised machine learning algorithm that can categorize fake news as true or false using tools like Python Scikit-Learn and Natural Language Processing for textual analysis. This procedure will yield feature extraction and vectorization; we suggest performing tokenization and feature extraction of text data using the Python scikit-learn module, which includes helpful functions like Count Vectorizer and Tiff vectorizer. Next, based on the results of the confusion matrix, we will experiment with feature selection techniques to select the best-fit features in order to get the highest precision.

Keywords: Fake news detection, Supervised machine learning, Natural Language Processing (NLP), Feature extraction and vectorization, Python Scikit-Learn, Confusion matrix and precision

I. INTRODUCTION

In the digital age, the internet has revolutionized the way we access and share information. However, this ease of information dissemination has also given rise to the proliferation of fake news—false or misleading information presented as news with the intent to deceive, manipulate public opinion, or generate monetary gains. Fake news has emerged as a significant challenge, influencing political landscapes, public health policies, financial markets, and even societal harmony. The advent of social media platforms has further exacerbated the problem by amplifying the spread of fake news at an unprecedented rate.

Fake news is a broad term that encompasses various forms of misinformation and disinformation. Misinformation refers to false information spread without malicious intent, while disinformation involves the deliberate dissemination of false or misleading content to achieve a specific agenda. Common types of fake news include hoaxes, propaganda, satire presented as truth, and outright fabrications. These can be propagated through social media posts, blogs, fabricated news websites, or even manipulated images and videos.

The consequences of fake news can be severe and far-reaching. For instance:

1. **Political Manipulation**: Fake news has influenced elections, swayed public opinion, and undermined democratic processes.
2. **Public Health Crises**: During the COVID-19 pandemic, the spread of false information about vaccines and treatments led to widespread confusion and hesitancy.
3. **Economic Ramifications**: False rumors can impact stock markets and business reputations, leading to financial losses.
4. **Social Divisions**: Fake news often exploits sensitive topics such as religion, race, and culture, fueling polarization and conflicts.

Detecting fake news is crucial to mitigate its impact and preserve the integrity of information. Manual fact-checking by journalists and fact-checking organizations, while essential, cannot keep pace with the sheer volume of content generated daily. This has led to the development of automated fake news detection systems powered by Artificial Intelligence (AI) and Machine Learning (ML). These systems can analyze and classify news articles, social media posts, and other forms of content to determine their authenticity.

2. RELATED WORK

2.1 Natural Language Processing :

Considering one or more system or algorithm specializations is the primary motivation for using natural language processing. An algorithmic system's Natural Language Processing (NLP) rating makes it possible to combine voice creation and comprehension. It could also be used to identify actions in different languages.Using a variety of language pipelines, including Emotion Analyzer and Detection, Named Entity Recognition (NER), Parts of Speech (POS) Taggers, Chunking, and Semantic Role Labeling, [5] proposed a new optimal system for extracting actions from English, Italian, and Dutch speeches. This made NLP a good search topic. [6][7].
Sentiment analysis [8] extracts feelings about a specific topic. Extracting a particular term for a subject, determining the sentiment, and combining it with connection analysis make up sentiment analysis. Two languages are used in the sentiment analysis. Analysis resources: Meaning and Sentiment Glossary

A database that looks for words that are constructive or destructive and tries to classify them on a scale of -5 to 5. Parts of speech taggers for European languages are being investigated in order to create parts of language taggers for other languages, including Arabic, Hindi, and Sanskrit [9, 10]. Capable of being effective Label and classify words as verbs, adjectives, names, and so on. The majority of speech patterns work well in European languages, but not in Arabic or Asian languages. The tree-bank approach is notably used in a portion of the Sanskrit word "speak". Vector is used in Arabic.Machine (SVM) [11] employs a technique to automatically recognize speech patterns and symbols, as well as to automatically reveal simple words in Arabic text [12].

2.2 Data Mining
The two primary categories of data mining techniques are supervised and unsupervised. To predict the concealed actions, the supervised approach makes use of the training data. The goal of unsupervised data mining is to identify hidden data models without training.Data, such as input label and category pairings. Aggregate mines and syndicate bases are prime examples of unsupervised data mining [13].

2.3 Machine Learning (ML) Classification
A class of techniques known as machine learning (ML) enables software systems to provide more accurate outcomes without requiring direct reprogramming. Changes or traits that the model must examine and use to generate predictions are described by data scientists.The learnt levels are divided into fresh data by the method [12]. In order to classify the bogus news, this paper uses six different algorithms.

2.4 Decision Tree
A crucial tool for categorization problems, the decision tree is built on a structure resembling a flow chart. A condition or "test" on an attribute is specified by each internal node of the decision tree, and branching is carried out based on the test conditions and outcome. Lastly, the leaf node has a class label that is determined by adding together all of its qualities. The classification rule is represented by the distance between the root and the leaf. Its ability to function with both dependent and categorical variables is astounding. They do a fantastic job of highlighting the most crucial elements and accurately illustrating how they relate to one another. They play an important role in developing new variables and characteristics that are helpful for exploring data and accurately forecasting the desired variable.

Decision Tree Pseudo-code
GenerateDecisionTree(Sample s, features F)
1. If stop _conditions(S,F) = true then
a. leaf = create_Node()
b. Leaf.lable= classify(s)
c. Return leaf
2. root = create_Node()
3. root.testcondition = find_bestSplit(s,f)
4. v = { v l v a possible outcome of root.testconditions)
5. for each value v∈ V:
6. sv: = {s | root.testcondition(s) = v and s∈ S};
7. child = Tree_Growth(Sv ,F) ;
8. Grow child as a descent of roof and label the edge (root□child) as
v
Return root

Tree based learning algorithms are widely with predictive models using supervised learning methods to establish high accuracy. They are good in mapping non-linear relationships. They solve the classification or regression problems quite well and are also referred to as CART [14][15][16].

2.5 Random Forest

The foundation of Random Forest is the idea of creating numerous decision tree algorithms, each of which produces a distinct outcome. The random forest adopts the outcomes that a majority of decision trees predicted. In order to guarantee that the decision trees are varied, the random forest randomly chooses a subset of each group's attributes [17][18]. When applied to uncorrelated decision trees, Random Forest is most applicable. The final outcome will resemble a single decision tree if applied to related trees. Bootstrapping and feature randomness can be used to create uncorrelated decision trees.

Random Forest Pseudo-code

To make n classifiers:
For i = 1 to n **do**
Sample the training data T randomly with replacement for Ti output
Build a Ti-containing root node, Ni
Call BuildTree (Ni)
end For
BuildTree (N):
If N includes instances of only one class, then returns
else
Select z% of the possible splitting characteristics at random in N
Select the feature F with the highest information gain to split on
Create f child nodes of N, Ni ,..., Nf , where F has f possible values (F1, ... , Ff)
For i = 1 to f do
Set the contents of Ni to Ti, where Ti is all instances in N that match Fi
Call Buildtree (Ni)
end for
end if [18]

2.6 Support Vector Machine (SVM)

The SVM algorithm is based on the layout of each data item in the form of a point in a range of dimensions n (the number of available properties), and the value of a given property is the number of specified coordinates [12]. Given a set of n features, SVM algorithm uses n dimensional space to plot the data item with the coordinates representing the value of each feature. The hyper-plane obtained to separate the two classes is used for classifying the data.

SVM Pseudo-Code

F[0..N-1]: a feature set with N features that is sorted by information gain in decreasing order accuracy(i): accuracy of a prediction model based on SVM with F[0...i] gone set
low = 0
high = N-1
value = accuracy(N-1)
IG_RFE_SVM(F[0...N-1], value, low, high) {
If (high } > low)
Return F[0...N-1] and value
mid = (low + high) / 2
value_2 = accuracy(mid)
if (value_2 < value)
return IG_RFE_SVM(F[0...mid], value_2, low, mid)
else (value_2 > value)
return IG_REF_SVM(F[0...high], value, mid, high) [13]

2.7. Naive Bayes

This algorithm works on Bayes theory under the assuming that its free from predictors and is used in multiple machine learning problems [18]. Simply put, Naive Bayes assumes that one function in the category has nothing to do with another. For example, the fruit will be classified as an apple when its of red color, swirls, and the diameter is close to 3 inches. Regardless of whether these functions depend on each other or on different functions, and even if these functions depend on each other or on other functions, Naive Bayes assumes that all these functions share a separate proof of the apples [14] .
Random Forest (RF) and Naïve Bayes have many differences, the main is their model size. The NB models are not good at representing complex behavior, resulting in low model size and good for a constant type of data. In contrast, the model sixe for Random Forest model is very large and it might

results in over fitting. NB is good for dynamic data and can be reshaped easily when new data is inserted while using a RF may require a rebuild of the forest every time a change is introduced.

2.8. KNN (k- Nearest Neighbors)
KNN classifies new positions based on most of the sounds from the neighboring k with respect to them. The position assigned in the class is highly mutually exclusive between the nearest neighbors K, as measured by the role of the distance [16]. KNN falls in the category of supervised learning and its main applications are intrusion detection, pattern recognition. It is nonparametric, so no specific distribution is assigned to the data or any assumption is made about them. For example GMM, assumes a Gaussian distribution of the given data.

3.METHODOLOGY

This section presents the methodology used for the classification. Using this model, a tool is implemented for detecting the fake articles. In this method supervised machine learning is used for classifying the dataset. The first step in this classification problem is dataset collection phase, followed by preprocessing, implementing features selection, then perform the training and testing of dataset and finally running the classifiers [1] [2] [3] [4] [5].The methodology is based on conducting various experiments on dataset using the algorithms described in the previous section named Random forest, SVM and Naïve Bayes, majority voting and other classifiers.

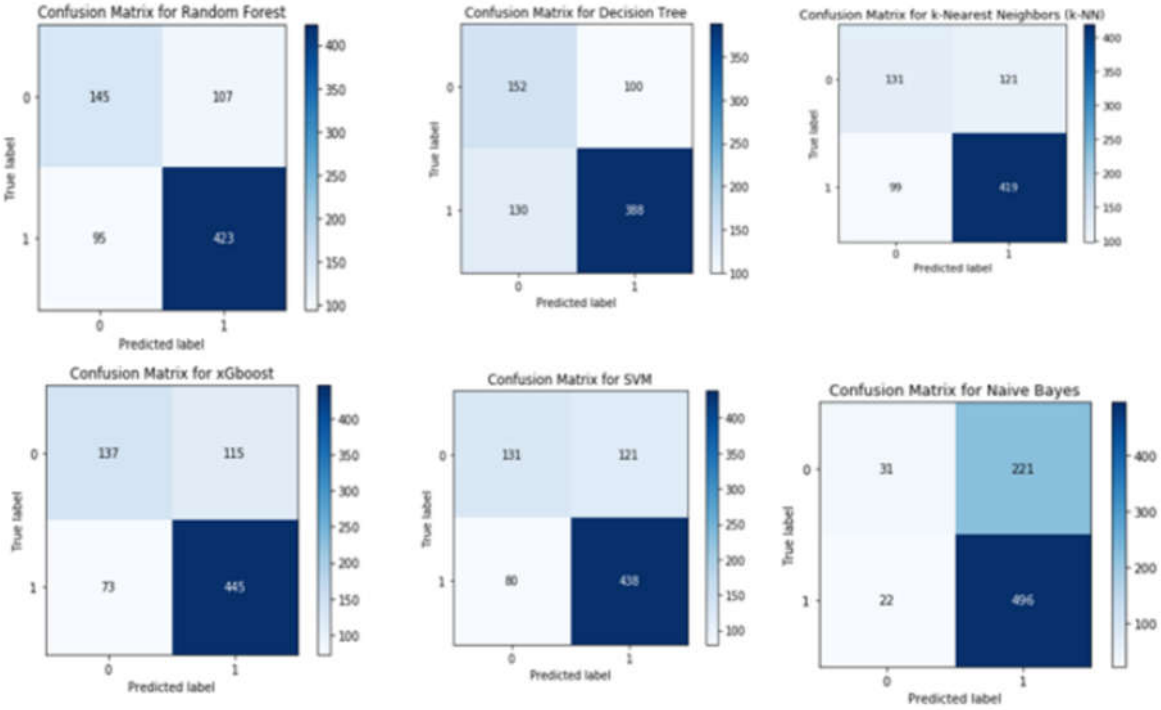
The main goal is to apply a set of classification algorithms to obtain a classification model in order to be used as a scanner for a fake news by details of news detection and embed the model in python application to be used as a discovery for the fake news data [19][20]. Also, appropriate refactorings have been performed on the Python code to produce an optimized code [21][22].
The classification algorithms applied in this model are k-Nearest Neighbors (k-NN), Linear Regression, XGBoost, Naive Bayes, Decision Tree, Random Forests and Support Vector Machine (SVM). All these algorithms get as accurate as possible. Where reliable from the combination of the average of them and compare them.
In the process of model creation, the approach to detecting political fake news is as follows: First step is collection political news dataset, (the Liar dataset is adopted for the model), perform preprocessing through rough noise removal, the next step is to apply the NLTK (Natural Language Toolkit) to perform POS and features are selected. Next perform the dataset splitting apply ML algorithms (Naïve bays and Random forest) then create the proposed classifier model. The Fig 2 shows that after the NLTK is applied, the Dataset gets successfully preprocessed in the system, then a message is generated for applying algorithms on trained portion. The system response with N.B and Random forest are applied, then the model is created with response message. Testing is performed on test dataset, and the results are verified, the next step is to monitor the precision for acceptance. The model is then applied on unseen data selected by user. Full dataset is created with half of the data being fake and half with real articles, thus making the model's reset accuracy 50%. Random selection of 80% data is done from the fake and real dataset to be used in our complete dataset and leave the remaining 20% to be used as a testing set when our model is complete. Text data requires preprocessing before applying classifier on it, so we will clean noise, using Stanford NLP (Natural language processing) for POS (Part of Speech) processing and tokenization of words, then we must encode the resulted data as integers and floating point values to be accepted as an input to ML algorithms. This process will result in feature extraction and vectorization; the research using python scikit-learn library to perform tokenization and feature extraction of text data, because this library contains useful tools like Count Vectorizer and Tiff Vectorizer.

4. RESULTS

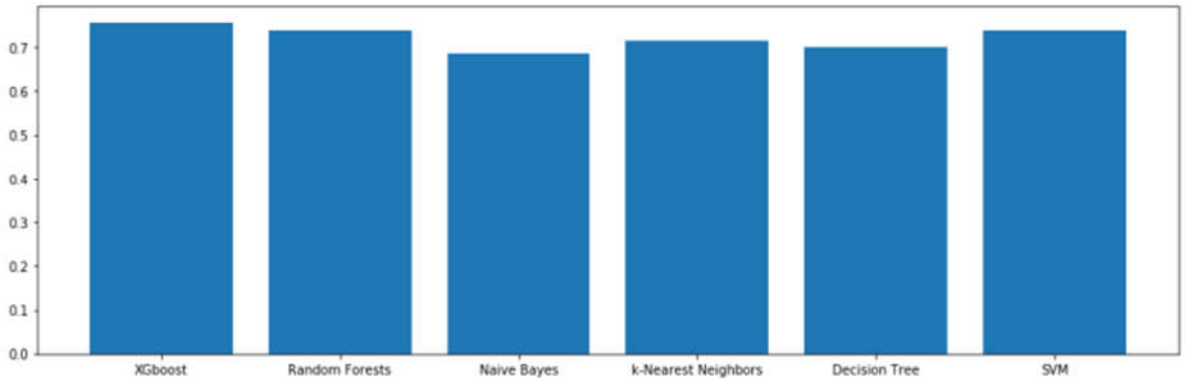
The scope of this project is to cover the political news data, of a dataset known as Liar-dataset, it is a New Benchmark Dataset for Fake News Detection and labeled by fake or trust news. We have performed analysis on "Liar" dataset . The results of the analysis of the datasets using the six algorithms have been depicted using the confusion matrix. The six algorithms used for the detection are as:

- XGboost.
- Random Forests.
- Naive Bayes.
- K-Nearest Neighbors (KNN).
- Decision Tree.
- SVM

The confusion matrix is automatically obtained by Python code using the cognitive learning library when running the algorithm code in Anaconda platform.



The XGBOOST is depicting the highest accuracy with more than 75%, next is SVM and Random forest with approximately 73% accuracy.



5. CONCLUSION

The research in this paper focuses on detecting the fake news by reviewing it in two stages: characterization and disclosure. In the first stage, the basic concepts and principles of fake news are highlighted in social media. During the discovery stage, the current methods are reviewed for detection of fake news using different supervised learning algorithms.

As for [20] the displayed fake news detection approaches that is based on text analysis in the paper utilizes models based on speech characteristics and predictive models that do not fit with the other current models. From [21] they utilized Naive Bayes classifier to detect fake news from different sources, with results of accuracy of 74%. [22] Used combined ML algorithms, but they depend on unreliable probability threshold with 85-91% accuracy. Naive Bayes to detect fake news from different social media websites, but the results were not accurate for the untruthful sources. They got their data from Kaggle with average accuracy of 74.5%. Used Naive Bayes algorithms to detect Twitter spam senders, with accuracy rated from 70% to 71.2%. They tried different approaches with accuracy of 76%. Three common methods are utilized through their researches: Naïve Bayes, Neural Network and Support Vector Machine (SVM). Naïve Bayes has an accuracy of

96.08% for detecting fake messages. The neural network and the machine vector (SVM) reached an accuracy of 99.9 0%. They used the combination of KNN and random forests that gave the final results improved by up to 8% using a mixed false message detection model.They worked on 2012 Dutch elections fake news on Twitter, they examine the execution of 8 supervised machine learning classifiers in the Twitter dataset. And they assume that the decision tree algorithm works best for the data set used with a F score of 88%. Presented a counterfeit detection model using N-gram analysis achieved the highest accuracy in use contains a unigram and a linear SVM workbook. The highest accuracy is 92%.

In the aforementioned research summary and system analysis, we concluded that most of the research papers used naïve bays algorithm, and the prediction precision was between 70-76%, they mostly use qualitative analysis depending on sentiment analysis, titles, word frequency repetition. In our approach we propose to add to these methodologies, another aspect, which is POS textual analysis, it is a quantitative approach, it depends on adding numeric statistical values as features, we thought that increasing these features and using random forest will give further improvements to precession results. The features we propose to add in our dataset are total words (tokens), Total unique words (types), Type/Token Ratio (TTR), Number of sentences, Average sentence length (ASL), Number of characters, Average word length (AWL), nouns, prepositions, adjectives etc.

REFERENCES

[1]. Sharma, Karishma, et al. "Combating fake news: A survey on identification and mitigation techniques." ACM Transactions on Intelligent Systems and Technology (TIST) 10.3 (2019): 1-42. [2]. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter, 19(1), 22-36. [3]. Shu, Kai, et al. "Fake news detection on social media: A data mining perspective." ACM SIGKDD explorations newsletter 19.1 (2017): 22-36. [4]. Khanam Z. and Agarwal S. Map-reduce implementations: Survey and Performance comparison, International Journal of Computer Science & Information Technology (IJCSIT) Vol 7, No 4, August 2015. [5]. Zhang, Jiawei, Bowen Dong, and S. Yu Philip. "Fakedetector: Effective fake news detection with deep diffusive neural network." 2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE, 2020.
[6]. Alkhodair S A, Ding S H.H, Fung B C M, Liu J 2020 “Detecting breaking news rumors of emerging topics in social media” Inf. Process. Manag. 2020, 57, 102018. [7]. Jeonghee Yi et al. “Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. ”In Data Mining, 2003. ICDM 2003. Third IEEE International Conference (pp. 427-434). http://citeseerx.ist.psu.edu. 200).2003

[8]. Tapaswi et al. “Treebank based deep grammar acquisition and Part-Of-Speech Tagging for Sanskrit m sentences.” Software Engineering (CONSEG), on Software Engineering (CONSEG), (pp. 1-4). IEEE. 2012
[9]. Ranjan et al. “Part of speech tagging and local word grouping techniques for natural language parsing in Hindi”. In Proceedings of the 1st International Conference on Natural Language Processing (ICON 2003). Semanticscholar. 2003
[10]. MonaDiab et al. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. Proceedings of HLT-NAACL 2004: Short Papers (pp. 149–152). Boston, Massachusetts, USA: Association for Computational Linguistics. 2004
[11]. Rouse, M. https://searchenterpriseai.techtarget.com/definition/machine-learning-ML May 2018
[12]. Sumeet Dua, Xian Du. “Data Mining and Machine Learning in Cybersecurity”. New York: Auerbach Publications.19 April 2016.
[13]. RAY, S. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/ 2017, September
[14]. Huang, T.-Q. (n.d.) https://www.researchgate.net/figure/Pseudo-code-of-information-gain-basedrecursive-feature-elimination-procedure-with-SVM_fig2_228366941 2018
[15]. Researchgate.net. Available at: Available at: https://www.researchgate.net/figure/Pseudocode-ofnaive-bayes-algorithm_fig2_325937073. 2018.
[16]. Researchgate.net. Available at: https://www.researchgate.net/figure/Pseudocode-for-KNNclassification_fig7_260397165, 2014.

[17]. NaphapornSirikulviriya; SukreeSinthupinyo. “Integration of Rules from a Random Forest.” International Conference on Information and Electronics Engineering (p. 194 : 198). Singapore: semanticscholar.org. 2011.
[18]. Jasmin Kevric et el. “An effective combining classifier approach using tree algorithms for network intrusion detection.” Neural Computing and Applications , 1051–1058. 2017.
[19]. Bovet, Alexandre, and Hernán A. Makse. "Influence of fake news in Twitter during the 2016 US presidential election." Nature communications 10.1 (2019): 1-14. The science of fake news.
[20]. https://science.sciencemag.org/content/359/6380/1094.summary Science 09 Mar 2018:Vol. 359, Issue 6380, pp. 1094-1096 DOI: 10.1126/science.aao2998.
[21]. Khanam, Z., Ahsan, M.N.”Evaluating the effectiveness of test driven development: advantages and pitfalls.”International. J. Appl. Eng. Res. 12, 7705–7716, 2017
[22]. Khanam, Z. “Analyzing refactoring trends and practices in the software industry.” Int. J. Adv. Res. Comput. Sci. 10, 0976–5697, 2018.