Fake News Detection Using Machine Learning

Anirudh Pandey
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
anirudhpandey244@gmail.com

Ritesh Bisht
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
riteshbisht1209@gmail.com

Aditya Verma
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
Adityaverma@gehu.ac.in

***Abstract*—The proliferation of false information in the digital era offers a huge challenge to information reliability, demanding sophisticated detection techniques. This research suggests a thorough method for identifying fake news using cutting-edge machine learning algorithms. We concentrate on three fundamental models, which are well-known for their effectiveness in classification tasks: Random Forest, Logistic Regression, and Support Vector Machine (SVM). As part of our process, we compare these models in-depth in order to determine the best strategy for separating fake news from real news. Our goal is to improve detection accuracy by utilizing the distinct advantages of each model, including the high-dimensional space classification of SVM, the probabilistic approach of Logistic Regression, and the ensemble learning technique of Random Forest. The theoretical underpinnings of each model will be discussed in the article, followed by a suggested feature extraction and model optimization according to the subtleties of data related to fake news. In addition to adding to the body of knowledge in the field of information verification, we hope that our findings will help media outlets and other stakeholders combat the spread of false information by offering them useful solutions. By conducting this study, we hope to set a standard for the identification of false news, opening the door for more responsible and knowledgeable news consumption in the digital era.**

***Index Terms*—Fake News Detection, Machine Learning, Support Vector Machine (SVM), Logistic Regression, Random Forest, Information Credibility, Classification Algorithms, Feature Extraction, Model Optimization, Misinformation Analysis**

I. Introduction

The Fake News epidemic has developed significantly over the past ten years, helped along by social media. There are several reasons why this bogus news might be disseminated. A few are created solely to boost the quantity of clicks and traffic to a website. others, to sway public opinion regarding financial markets or political issues. For instance, by harming organizations' and businesses' online reputations. Social media fake news about health poses a threat to world health. In February 2020, the World Health Organization issued a warning that the COVID-19 pandemic had been accompanied by a large "infodemic," or an excess of information—some of which was correct and some of which wasn't—that made it challenging for people to access trustworthy sources and information when they needed it. Uncertainty, fear, anxiety, and bigotry are spreading as a result of disinformation overload on a magnitude not observed in previous outbreaks [11]. In this study, we introduce a new approach and instrument for identifying false news that makes use of:

• **Text preprocessing:** this includes heating and analyzing the text to eliminate special characters and stop words.
• **Text encoding:** TF-IDF , bag of words, and N-gram are used first.
• **Characteristic extraction:** this makes it possible to precisely identify misleading information. The author, the news's source, the date, and the sentiment conveyed by the text are all considered news features.
• **Support vector machine:** is an approach for supervised machine learning that facilitates the categorization of novel data.
• **Logistic Regression:** When it comes to statistics, logistic regression and binary switches are comparable in that they can both identify the probability that an occurrence will fall into a single of two groups, such as good or bad, gain, or on/off. It estimates the likelihood that a situation will occur compared to not occurring based on the odds principle. It performs exceptionally well in tasks requiring binary outcomes.
• **Random Forest:** Random Forest is like an ensemble of experts educated on different datasets. Instead of relying on a single expert, this team combined their talents. Each team member casts a vote for what reaction they think should be provided, and the vote with the most support determines the final decision. By using the collective knowledge in this way, more accurate and trustworthy results are produced. Furthermore, suggestions get more precise and avoid the common over fitting issue of being unduly specific or favored toward one piece of data as the team's number of specialists increases. In short, the Random Forest produces a comprehensive and reliable forecast by combining collaboration and expertise.

The format of this document is as follows: A few current ideas for detecting bogus news are presented in Section 2. Section 3 explains the various parts of our plan. Section 4 contains the dataset we will use and section 4 is the conclusion.

II. Literature Survey

Many literary works attempt to imitate fake news detection. Authors [3] offer a classification of various truth evaluation techniques that fall into two primary categories: machine learning-based linguistic cue approaches and network analysis

approaches for identifying false news.

A naive Bayesian classifier is used by the authors of [5] to easily identify bogus news. We evaluate our strategy with a collection of information taken from postings on Facebook. According to their claims, they can achieve an accuracy rate of 74%. While several other studies have achieved higher rates with various classifiers, this model's rate is respectable but not the greatest. An analysis of these works may be found here.

Using n-gram analysis and machine learning techniques, the authors of [1] propose a fake news detection model by contrasting two distinct feature extraction methods procedures and six distinct methods of classification. The so-called features extraction approach (TF-IDF) produces the greatest outcomes, according to the performed research' results. They used a classifier with a 92% accuracy rate, the Linear Support Vector Machine (LSVM). Because LSVM is used in this model, it can only handle situations when there are two linearly separated classes.

How social network users can confirm the authenticity of information is explained by the authors of [11]. They also talk about the processes that allow them to be validated, The function of reporters, and expectations from government organizations and academic establishments. Readers of the following content are better equipped to separate the veracity from some of the unreliable news on social media.

The authors of [9] offer a number of approaches and index types that are related to various modalities (text, image, and social information). In order to evaluate and confirm shared information, they also discuss the benefits of integrating and merging these methodologies.

The authors provide a comprehensive study of the performance of various methods on three distinct datasets in [8]. This project overlooks certain details that could have a significant influence on the outcome, such as the author, source, or publication date, in favor of concentrating only on the text of the information and the sensation it conveys. Furthermore, our research will demonstrate that including emotion into the detection procedure yields no useful information.

The authors of [12] developed a fresh public dataset of genuine news stories and suggested a machine learning method based on text processing for the 87% accurate automatic detection of fake news. This piece seems to be more concerned with the emotions that come to light when they read the words than with the text itself.

LIAR, a novel dataset for automatic fake news identification, was introduced by the authors of [14]. In addition, this corpus can be utilized for political natural language processing, argument mining, topic modeling, rumor detection, and stance classification. This norm has been accepted. in the most of the works in this field. It is commonly known, nonetheless, that the latter only includes political data, whereas the former incorporates data from a variety of disciplines.

The primary drawback of these techniques is the encoding of categorical data might not hold true in practice! Furthermore, the conventional classification of false information is restricted to two things (i.e., True or False), although in practice we are only able to determine the news's authenticity to a certain degree of confidence rather than 100%. We believe that this element is crucial for social media news classification.

III. Proposed System

Our suggested approach builds a decision-making model utilizing random forests, support vector machines and logistic regression techniques using a news dataset. Next, the algorithms are applied to categorize new news as either valid or fake.

A. The suggested system's general architecture

The suggested system receives a dataset of comments as input and the details that go with them, such the date, the source, and the writer. After that, it creates a features dataset that may be utilized during the learning process. Preprocessing is the name given to this transformation, which includes a number of tasks like encoding, filtering, and cleaning. There are two sections to the preprocessed dataset: A separate one for testing and another for training. The instruction package creates a a decision-making model that works with the test data by utilizing the support vector machine algorithm and the training dataset. The model can be retained and used after training is completed if it is deemed suitable (i.e., able to achieve a sufficient precision rate).

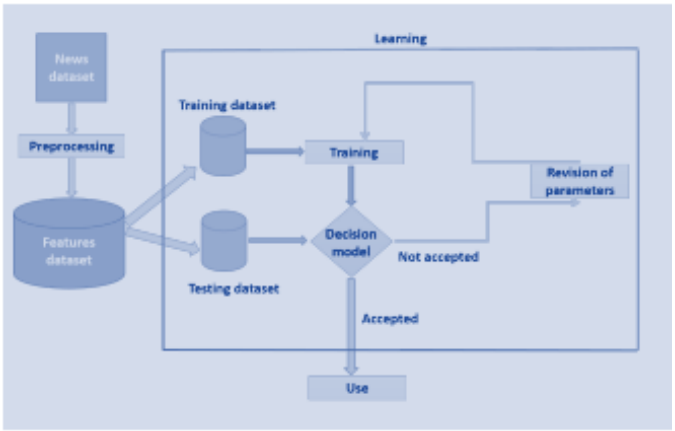

Fig. 1. The suggested system's general architecture

B. Preprocessing

Textual data, category data, and numerical data are the three categories into which news attributes are divided in the news dataset. Preprocessing for each category is carried out using a series of procedures, as shown in Figure 2.

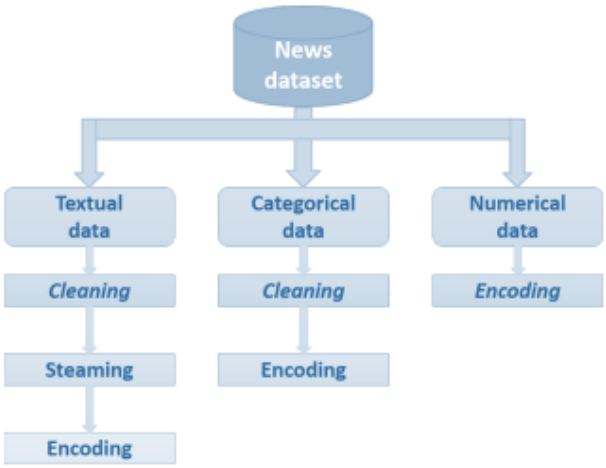

Fig. 2. Preprocessing

• **Textual data** Present the author's written work as it was preprocessed using the following methods in a news article:

1) Cleaning: getting rid of special characters and stop words.
2) Steaming: turning the practical terms into root words.
3) Encoding: converting every word in the comment into a vector of numbers. This requires two steps: first, the outcomes of the TF-IDF approach [12]; second, a mix of two methods, namely N-grams [4] and bag of words [13].
   $TF\text{-}IDF_t = TF_t \cdot IDF_t = \frac{n}{k} \cdot \log \frac{D}{D_t}$

• **Categorical data** Represent the news's author as well as its source, such as the TV station, a magazine or newspaper. There are two phases involved in the pre-treatment of these data: - Cleaning: converting letters to lowercase and removing special characters.
- Encoding: We employed label encoding for sources. To make authors from the same source appear closer to one another than authors from different sources, we developed our personal encoding to translate the authors' names into numerical representations.
We will construct a list with two fields: one for the source and another for its authors. We will then add the total of the dimensions of the prior resources plus one to each author's index number.

• **Numerical data** Reflect the text's mood and the date the comment was posted. We just divided the date into three distinct values—day, month, and year—because it is already represented by a numerical number. We determine the total of the word sentiment degrees for the sentiment conveyed by the text. The specialists claim that every word has a sentimental degree that enables it to be divided into three categories:
- The feeling is neutral if the sum is zero.
- The sentiment is positive if the amount is larger than zero.
- The feeling is negative if the sum is less than zero.

C. Learning

It combines the training and validation modules into one. 1) *Training* We have selected three algorithms support vector machine algorithm, random forest, logistic regression [4] to train our model. This makes it possible to use the judgment function value that is given for an article of news to indicate how confidently it is classified: A real news item and its degree of truth are concurrently designated by a positive decision function value, and vice versa; a false news item and its level of imitation are simultaneously designated by a negative decision functional value. As a result, the decision function's maximum and minimum are determined in the training stage and utilized to assess the degree of error or truth using the following function:

$$\text{p} = \begin{cases} \frac{Dec}{Max_{dec}} \times 100 & if\ Dec > 0 \\ \frac{Dec}{Min_{dec}} \times 100 & else \end{cases}$$

Where:

- p is the percentage of real or fake.
- The value of the decision function is Dec.
- Mar_{dec} and Min_{dec} are the decision function's maximum and minimum values.

2) *Validation:* A portion of the cases were reserved to serve as test models in order to gauge the model's ability to recognize new cases. Next, there are two sections to the features dataset: a training section and an examination segment. Its benefit lies in preventing excessively fitting, which occurs when the model is tested using the identical training set. Using the cross-validation method, the subdivision is carried out based on a specific sample rather than at random [10].

D. Parameter revision

By adjusting the cross-validation variable and fine-tuning or modifying the parameters of the algorithms namely, Cost, Y , and —this operation seeks to increase the accuracy of the model [2].

E. Use

The final and most crucial stage of our system is this one. We will use the best model, which we will build, to predict the classes of unlabeled news—true or false—with a degree of confidence once we have reached the best recognition rate.

F. Dataset

We will use dataset namely "Fake News" from Kaggle[5]. Author of this dataset is William Lifferth. This dataset contains an ID which is unique for a piece of news. Secondly, it contains the title for that piece of news and also the author. It also contains the text for that article which may not be comprehensive and a label which shows piece of news is fake.

IV. Conclusion

When we design our research and get ready to go on our investigation voyage into the world of news integrity, we plan to use a wide range of analytical instruments. The models we have selected Logistic Regression and Random Forest, offer an innovative method for analyzing the 'Fake News' dataset.

We see Random Forest as a group of decision-making elements working together as a team, sharing knowledge and making democratic decisions that guarantee thorough analysis. We may use logistic regression to plot the attributes of our news articles in connection to one another, giving us a statistically sound narrative of what represents truth and what reflects untruth.

References

[1] Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1 (pp. 127-138). Springer International Publishing.

[2] Chang, C.-C., Lin, C.-J. (2018). LIBSVM – A Library for Support Vector Machines.

[3] Conroy, N. J., Rubin, V. L., Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1–4.

[4] Faloutsos, C. (1985). Access methods for text. ACM Computing Surveys (CSUR), 17(1), 49–74.

[5] Lifferth, W. (2018). Fake News. Kaggle. https://www.kaggle.com/williamlifferth/fake-news

[6] Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A., Afroz, S. (2019). A benchmark study on machine learning methods for fake news detection. arXiv preprint arXiv:1905.04749.

[7] Maigrot, C., Kijak, E., Claveau, V. (2018). Fusion par apprentissage pour la détection de fausses informations dans les réseaux sociaux. Document numerique, 21(3), 55–80.

[8] Payam, E., Tang, L., Liu, H. (2009). Cross-validation. Encyclopedia of database systems, 532–538.

[9] Pulido, C. M., Ruiz-Eugenio, L., Redondo-Sama, G., Villarejo-Carballido, B. (2020). A new application of social impact in social media for overcoming fake news in health. International Journal of Environmental Research and Public Health, 17(7), 2430.

[10] Salton, G., McGill, J. M. (1983). Introduction to modern information retrieval.

[11] Sauvageau, F. (2018). Les fausses nouvelles, nouveaux visages, nouveaux défis. Comment déterminer la valeur de l'information dans les sociétés démocratiques? Presses de l'Université Laval.

[12] Schölkopf, B., Smola, A. J. (2018). Learning with kernels: support vector machines, regularization, optimization, and beyond. Adaptive Computation and Machine Learning series.

[13] Singh, D. S. K. R., Dasgupta, R. Automated fake news detection using linguistic analysis and machine learning.

[14] Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648.

[15] Lechevallier, Y. WEKA, un logiciel libre d'apprentissage et de data mining. INRIA-Rocquencourt.