e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

International Journal Of Multidisciplinary Research In Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704014 |

Fake News Detection System using Machine Learning and Web Extraction

S.Ashwin Balaji, Assistant Professor. Dr.N.Mahendiran

UG Student, Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu, India

Assistant Professor, Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu, India

ABSTRACT: The Fake News Detection System is an innovative web-based application designed to combat the proliferation of misinformation and deceptive content in the digital age. This system employs advanced natural language processing and machine learning techniques to analyze news articles and classify them as either authentic or potentially fabricated. Users can input news text into the platform, triggering a predictive model that evaluates the likelihood of the news being genuine. The application provides transparency by offering insights into the features influencing the model's decision. The system utilizes a combination of traditional and deep learning models, including Logistic Regression, Decision Trees, to enhance prediction accuracy. Text preprocessing techniques, such as vectorization and tailored word optimization, contribute to the robustness of the model. The web interface enables users to choose from multiple models and receive real-time predictions.In addition to the prediction functionality, the system incorporates model evaluation metrics, fostering an understanding of the model's performance. It addresses imbalanced data concerns, implements hyperparameter tuning for model optimization, and integrates model explanation methods like SHAP values. The platform ensures a seamless user experience by incorporating error handling, loading indicators, and handling edge cases.With the Fake News Detection System, users can make informed decisions about the authenticity of news articles, contributing to a more informed and discerning society. The application aims to mitigate the impact of fake news by providing a reliable tool for news verification.

I. INTRODUCTION

People are spending more and more time interacting on social media, as the wide adoption of smartphones makes their access available almost anytime and anywhere, which is not the case with traditional media. In addition, they facilitate interaction with friends, families, and even strangers through the comment chains, be it through comments, discussions, or simply like and dislike buttons. This has made social media a main channel for the dissemination of news. However, new technologies and features can be used through social media platforms to spread fake news on a large scale. Such inaccurate information might result either from a deliberate attempt to deceive or mislead (disinformation) or from an honest mistake (misinformation). Rumours can fall into either of these two categories, depending on the intent of the source, given that rumours are not necessarily false but may turn out to be true. Unlike rumours, fake news is, by definition, always false and, thus, can be seen as a type of disinformation. Therefore, credible and reliable sources of information are needed so that the public does not fall prey to the intentions of those interested in manipulating reality.

Zhou et al. [6] proposed a theory-driven model for fake news detection. Fake news detection is then conducted within a supervised machine learning framework which enhances the interpretability of fake news feature engineering, and studies the relationships among fake news, deception/disinformation, and click baits. Experiments conducted on two real-world datasets indicate the proposed method can outperform the state-of-the-art and enable fake news early detection when there is limited content information. Datasets consisting of the ground truth of, e.g., both fake news and clickbait, are invaluable to understanding the relationships among different types of unreliable information; however, such datasets are so far rarely available. Furthermore, it should be pointed out that effective utilization of rhetorical relationships and utilizing news images in an explainable way for fake news detection are still open issues. Kaliyar et al. [7] proposed coupled matrix–tensor factorization method to get a latent representation of both news content as well as social context. To classify news content and social context-based information individually as well as in combination, a deep neural network was employed with optimal hyper-parameters. For the task of fake news detection, a feature set can never be considered complete and sound. Jiang et al. [8] evaluated the performance of five machine learning models and three deep learning models on two fake and real news datasets of different sizes withholding out cross-

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704014 |

validation. Moreover, the detection of fake news with sentiment analysis is required for different machine learning and deep learning models.

II. LITERATURE REVIEW

Data preprocessing can be thought of as text mining, in which texts are unstructured data and contain numerous impurities [9, 10], including HTML tags, single characters, ads, non-English characters, numbers, or an apostrophe [9]. This is why the process of expressing textual data in natural language is very difficult.Many techniques convert unstructured data into structured data that a machine can accommodate. This study applied the stopword technique to the process of cleaning the dataset that was classified. Stopword technology is a common technique used in data filtering, information retrieval, and text classification, removing certain valueless words (e.g., the, in, a, an, with, etc.). That is, items are removed that are not keywords that affect categorization [10], [11]. The Python Standard Library was used to remove HTML tags by employing the (remove_tags) function. Next, the preprocess text function removed non-English characters.

Extracting features is a method in which text data are converted into Vectors 0 and 1, and new vectors were created from the sample text file. There are several techniques by which one can create vectors:

TF-IDF vectorizer: One of the most widely used feature extraction techniques is the term frequency-inverse document frequency (TF-IDF) [12]. This technique is divided into two stages, in which the term frequency (TF) is calculated first, and the inverse document frequency (IDF) is calculated in the second stage.

$$TF(t) = \frac{\text{No of times the t term appears in a doc.}}{\text{Total No of terms in the document}}$$

$$IDF(t) = \text{Log}\left(\frac{\text{total No. of documents}}{\text{no. of documents containing term t}}\right)$$

A. Logistic Regression

It is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes and success) or 0 (no and failure). Uma Sharma, Sidarth Saran, Shankar M. Patil developed a Fake News Detection using Machine Learning Algorithms [13]. They used liar dataset for detecting if fake news by Naive Bayes Classifier, Logistic Regression, Random Forest. They used Bag-Of-Words, N-Grams, TFIDF. Logistic regression shows better results with accuracy of 65%. Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf and Muhammad Ovais Ahmad implemented a model for fake news detection on social media [11]. They worked on Logistic Regression, SVM and KNN models using social media and fake news datasets. LIWC method is used for feature extraction. On their experiment they found that Logistic Regression shows high accuracy 91% compared to SVM 67% and KNN-68%. Vanya Tiwari, Ruth G. Lennon and Thomas Dowling developed some Machine learning Algorithms for Fake news detection [14].

B. Decision Tree

One of the most common algorithms used in classification algorithm. It is based on the algorithm in which all the data to be studies must of the type numeric and categorical kind. Therefore, a continuous type of data will not be examined [16].This algorithm utilizes two different pruning ways. The first method, named subtree replacement, which denotes the possibility of replacement nodes in a decision tree with its leaves to minimize the number of tests in the convinced path. Usually, the subtree raising is of a modest impact on the models of the decision tree. Typically, there is no exact way to predict an option's utility, although it can be advisable to turn it off when the induction procedure takes longer because of the subtree's raising being relatively computationally complicated

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704014 |

III. METHODOLOGY OF PROPOSED SURVEY

The proposed methodology for the fake news detection project involves a multi-step approach aimed at developing a robust and accurate system for identifying misinformation. Firstly, the project will commence with data acquisition and preprocessing. This stage involves gathering a diverse dataset consisting of both genuine and fake news articles from reputable sources and fake news repositories. The collected data will then undergo thorough preprocessing steps, including text normalization, removal of irrelevant information such as HTML tags and special characters, and tokenization.

Following data preprocessing, the project will focus on feature extraction and engineering. Techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) will be employed to convert the textual data into numerical features, capturing the importance of words in distinguishing between real and fake news. Additionally, other feature engineering methods, such as word embeddings and n-grams, may be explored to enhance the representation of textual data.

Next, the project will proceed to model selection and training. Various machine learning algorithms, including but not limited to Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Neural Networks, will be evaluated to identify the most effective classifier for fake news detection. Hyperparameter tuning and cross-validation techniques will be employed to optimize the performance of selected models.

Next, the project will proceed to model selection and training. Various machine learning algorithms, including but not limited to Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Neural Networks, will be evaluated to identify the most effective classifier for fake news detection. Hyperparameter tuning and cross-validation techniques will be employed to optimize the performance of selected models.

Subsequently, the developed models will undergo rigorous evaluation using appropriate metrics such as accuracy, precision, recall, and F1-score. The evaluation process will involve testing the models on a separate validation dataset to assess their generalization capabilities and robustness in real-world scenarios. Moreover, techniques such as k-fold cross-validation and stratified sampling will be utilized to ensure unbiased evaluation results.

In parallel with model development and evaluation, the project will also explore advanced techniques for enhancing the performance of fake news detection systems. This may include ensemble methods, deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), and incorporating external knowledge sources such as fact-checking databases and social network analysis.

.IV. CONCLUSION AND FUTURE WORK

Detecting fake news is a critical endeavor in today's digital age, where misinformation can spread rapidly and have significant societal consequences. Through the utilization of advanced technologies and methodologies, such as machine learning algorithms and natural language processing techniques, significant strides have been made in the field of fake news detection. This has enabled researchers and practitioners to develop robust frameworks and models capable of identifying deceptive or misleading information with a high degree of accuracy.

One of the primary conclusions drawn from the exploration of fake news detection is the importance of data preprocessing and feature engineering. Text preprocessing techniques, such as lowercasing, tokenization, and removing stop words and special characters, are essential for transforming raw text data into a format suitable for analysis. Additionally, feature engineering methods, such as TF-IDF (Term Frequency-Inverse Document Frequency), play a crucial role in capturing the semantic meaning and importance of words within a document, thereby enhancing the effectiveness of machine learning models in distinguishing between real and fake news.web extraction is a vital component of the fake news detection system, enabling the acquisition of textual data from diverse online sources for analysis and classification.

Furthermore, the training and evaluation of machine learning models constitute another key aspect of fake news detection. Models such as logistic regression and decision trees have demonstrated impressive performance in classifying news articles based on their authenticity. Evaluation metrics such as accuracy, precision, recall, and F1-score provide valuable insights into the model's performance, enabling researchers to assess its effectiveness in real-

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704014 |

world scenarios. Moreover, techniques such as cross-validation and hyperparameter tuning help optimize model performance and generalization to unseen data.

Web extraction techniques also play a vital role in fake news detection, allowing researchers to gather information from online sources and analyze their content for potential misinformation. By extracting text from news articles and websites, researchers can apply machine learning models to identify suspicious patterns or inconsistencies indicative of fake news. However, challenges such as the dynamic nature of online content and the prevalence of sophisticated misinformation campaigns underscore the need for continuous innovation and adaptation in the field of fake news detection.

In conclusion, while fake news remains a persistent challenge in the digital landscape, advancements in technology and data-driven approaches offer promising solutions for identifying and combatting misinformation. By leveraging machine learning algorithms, text preprocessing techniques, and web extraction methods, researchers and practitioners can develop robust frameworks capable of detecting fake news with a high degree of accuracy. Moving forward, interdisciplinary collaboration and ongoing research efforts will be essential in addressing the evolving nature of fake news and safeguarding the integrity of information dissemination in society.

REFERENCES

[1] 1. Ahmad I, Yousaf M, Yousaf S, Ahmad M. Fake news detection using machine learning ensemble methods. *Complexity.* 2020;2020:1–11.
[2] Hannah Nithya S, Sahayadhas A (2022) Automated fake news detection by LSTM enabled with optimal feature selection. J Inf Knowl Manag 21(03). 10.1142/s0219649222500368
[3] Akinyemi B. Department of computer science and engineering, Obafemi Awolowo University, Ile-Ife, Nigeria, Adewusi O, Oyebade a. an improved classification model for fake news detection in social media. *Int J Inf Technol Comput Sci.* 2020;12(1):34–43. doi: 10.5815/ijitcs.2020.01.05.
[4] Alonso MA, Vilares D, Gómez-Rodríguez C, Vilares J. Sentiment analysis for fake news detection. *Electronics (Basel)* 2021;10(11):1348. doi: 10.3390/electronics10111348
[5] Harb JG, Ebeling R, Becker K. (2020) A framework to analyse the emotional reactions to mass violent events on Twitter and influential factors. Inform Process Manag 57.
[6] Zhou X, Jain A, Phoha VV, Zafarani R. Fake news early detection: a theory-driven model. *Digital Threats: Research and Practice.* 2020;1(2):1–25. doi: 10.1145/3377478.
[7] Kaliyar RK, Goswami A, Narang P. EchoFakeD: improving fake news detection in social media with an efficient deep neural network. *Neural Comput Appl.* 2021;33(14):8597–8613. doi: 10.1007/s00521-020-05611-1
[8] Jiang T, Li JP, Haq AU, Saboor A, Ali A. A novel stacking approach for accurate detection of fake news. *IEEE Access.* 2021;9:22626–22639. doi: 10.1109/access.2021.3056079
[9] M. Rahul R., et al. "Identification of Fake News Using Machine Learning." 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). IEEE, 2020.
[10] K. Sawinder, P. Kumar, and P. Kumaraguru. "Automating fake news detection system using multi-level voting model." Soft Computing 24.12 (2020): 9049-9069.
[11] B. Kwadwo Osei. "Weighted Accuracy Algorithmic Approach In Counteracting Fake News And Disinformation." arXiv preprint arXiv:2008.01535 (2020).
[12] G. Samujjwal, and M. Sankar Desarkar. "Class specific tf-idf boosting for short-text classification." Proc. of SMERP 2018 (2018).
[13]Uma Sharma, Sidarth Saran, Shankar M. Patil developed a Fake News Detection using Machine Learning Algorithms.
[14] A. Kumar, R. S. Umurzoqovich, N. D. Duong, P. Kanani, A. Kuppusamy, M. Praneesh, and M. N. Hieu, ‘‘An intrusion identification and prevention for cloud computing: From the perspective of deep learning,’’ Optik, vol. 270, Nov. 2022, Art. no. 170044
[15] Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf and Muhammad Ovais Ahmad implemented a model for fake news detection on social media.
[16] Vanya Tiwari, Ruth G. Lennon and Thomas Dowling developed some Machine learning Algorithms for Fake news detection.
[17] Suhad A. Yousif, Islam Elkabani, "The Effect of Combining Different SemanticRelations on Arabic Text Classification", Vol. 5, No. 6, 112-118, 2015
[18] Praneesh, M., and R. Annamalai Saravanan. "Deep Stack Neural Networks Based Learning Model for Fault Detection and Classification in Sensor Data." Deep Learning and Edge Computing Solutions for High Performance Computing (2021): 101-110.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com