Fake News Detector

Pankaj Pandit[1], Mr. Pawan Kumar[2], Vikash Kumar Baghel[3], Manish sahu[4], Santoshi Parida[5]

[1] BCA Student Computer Science& IT, Kalinga University, Naya Raipur (C.G)
[2] Asst Prof. Computer Science & Engineering, Kalinga University, Naya Raipur (C.G)
[3] BCA Student Computer Science& IT, Kalinga University, Naya Raipur (C.G)
[4] BCA Student Computer Science& IT, Kalinga University, Naya Raipur (C.G)
[5] BCA Student Computer Science& IT, Kalinga University, Naya Raipur (C.G)

Abstract—
The proliferation of fake news in digital media presents a significant challenge to information integrity. This research explores the application of machine learning, specifically logistic regression, for automated fake news detection using a dataset sourced from Kaggle. Text preprocessing techniques, including tokenization, stemming, and TF-IDF vectorization, were applied to extract features from news articles. A logistic regression model was trained on the processed data to classify articles as real or fake. The model achieved high accuracy rates of 98.68% on the training set and 97.67% on the testing set. Additionally, a user-friendly Streamlit web application was developed for real-time prediction of fake news. This study demonstrates the efficacy of logistic regression in combatting misinformation and contributes to enhancing information credibility in digital media.

Key Words **— Fake news detection, machine learning, logistic regression, TF-IDF vectorization, text preprocessing, information integrity.**

I. INTRODUCTION

The widespread dissemination of fake news in the digital era has profound implications for information integrity and public perception. This research delves into the realm of machine learning to address this pressing issue. Leveraging tools like Jupyter Notebook, we implement a logistic regression model to automate the detection of fake news. Jupyter Notebook, an interactive computing environment, facilitates seamless experimentation and documentation of our machine learning pipeline. Our study harnesses a dataset sourced from Kaggle, comprising news articles labeled as real or fake, for training and evaluation.

The study utilized a dataset sourced from Kaggle consisting of news articles labeled as either real or fake. Data preprocessing involved addressing missing values and consolidating text features. Techniques like TF-IDF vectorization and stemming were applied to prepare the data for modeling. A logistic regression model was then trained on this preprocessed data and evaluated using a stratified split approach. The model demonstrated high accuracy, proving effective in discerning between genuine and fabricated news articles. Additionally, a user-friendly Stream lit web application was developed to provide real-time predictions on the authenticity.
of news articles. This study showcases the potential of machine learning in combating misinformation and enhancing the credibility of online information sources.

II. LITREATURE REVIEW

Monti et al. (2019) proposed a cross breed convolutional neural arrange (CNN) and long short-term memory (LSTM) show for fake news discovery. Their approach utilized a CNN to extricate significant highlights from content and an LSTM to capture the successive nature of the substance, accomplishing a precision of 87.5% on the ISOT Fake News Dataset.[1]

Zhou et al. (2020) created a multi-modal fake news discovery framework that coordinating literary highlights from news articles with visual highlights from related pictures. They utilized a Chart Convolutional Organize (GCN) to capture connections over distinctive modalities, outflanking a few pattern models and underscoring the significance of leveraging multi-modal information for fake news detection. [2]

Kaliyar et al. (2021) presented a transformer-based demonstrate named Fake Detector, which utilizes consideration components and self-attention layers to capture long-range conditions in content. Their show accomplished state-of-the-art execution on different fake news datasets, counting the ISOT dataset and the LIAR dataset. [3]

Wang et al. (2022) examined the utilize of chart neural systems (GNNs) for fake news location. Their strategy builds a heterogeneous chart that speaks to connections between news articles, distributors, and clients. The GNN demonstrate learns to proliferate and total data from distinctive hubs in the chart, driving to made strides location performance.[4]

Oshiba et al. (2020) inspected outfit strategies for fake news location. They combined different machine learning models, counting bolster vector machines (SVMs), arbitrary

timberlands, and calculated relapse, utilizing a stacking gathering approach. Their outfit demonstrate outperformed person models, highlighting the potential of combining assorted classifiers for progressed fake news detection.[5]

III. PROBLEM IDENTIFICATION

The rapid dissemination of fake news through online platforms has become a pressing issue in today's information age. Misleading or fabricated news articles can influence public opinion, distort societal perceptions, and undermine trust in media sources. The manual identification of fake news is labor-intensive and inefficient, necessitating automated solutions.

Existing approaches to fake news detection often rely on rule-based systems or traditional statistical methods, which may be limited in their ability to capture complex patterns in textual data. This research aims to leverage machine learning techniques, specifically logistic regression, to develop a scalable and efficient system for automated fake news detection.

The rapid dissemination of fake news through online platforms has become a pressing issue in today's information age. Misleading or fabricated news articles can significantly impact public opinion, distort societal perceptions, and erode trust in media sources. Manual identification of fake news is both labor-intensive and inefficient, necessitating the development of automated solutions.

Current approaches to fake news detection often rely on rule-based systems or traditional statistical methods, which may struggle to capture the intricate patterns present in textual data. This research seeks to leverage machine learning techniques, specifically logistic regression, to create a scalable and effective system for automated fake news detection.

By tackling the challenge of misinformation online, this study aims to bolster the credibility of digital information sources and empower users to make informed decisions based on reliable content. Through the application of advanced technology, we can combat the spread of fake news and foster a more trustworthy digital environment.

IV. PROPOSED WORK

The proposed research be seeking to develop an advanced fake news detection system utilizing machine learning technology. Given the fast spread of misleading information in today's digital world, it has become necessary to create automated solutions that can effectively detect and combat false news. The main goals of this research consist of thorough data collection, careful preprocessing, model creation, thorough performance assessment, and smooth system implementation.

The initial step of the research involves putting together labeled datasets containing news articles and applying meticulous preprocessing methods to ready the text data for analysis using machine learning. This preprocessing involves crucial actions like removing irrelevant words, applying changes to standardized word variations, and converting the text into vectorized representations using TF-IDF.

The study is focused on creating strong models using logistic regression techniques, trained on carefully preprocessed data to distinguish between real news and fake content. The effectiveness of the model will be thoroughly assessed using precise metrics on a specific test dataset, indicating its capability to accurately categorize news content.

The end goal of this project includes rolling out the developed system via a user-friendly web interface utilizing Streamlit technology. This interface will allow users to input news articles into the trained machine learning model and get instant predictions regarding the authenticity of the content.

In summary, this research highlights the vital role of machine learning in automatic false news detection, providing practical solutions to fight misinformation online. The expected result is the establishment of a reliable and effective system for spotting fake news, thereby boosting trust in digital news sources and lessening the harmful effects of misinformation in the digital age.

V. WORKING MODEL

The false news detection system leverages machine learning technology to address the pervasive challenge of misinformation in digital media. The initial phase of this system involves data acquisition, where labelled datasets of news articles are collected and categorized as real or fake. These datasets serve as the foundation for training the machine learning model to discern patterns distinguishing between factual and false information.

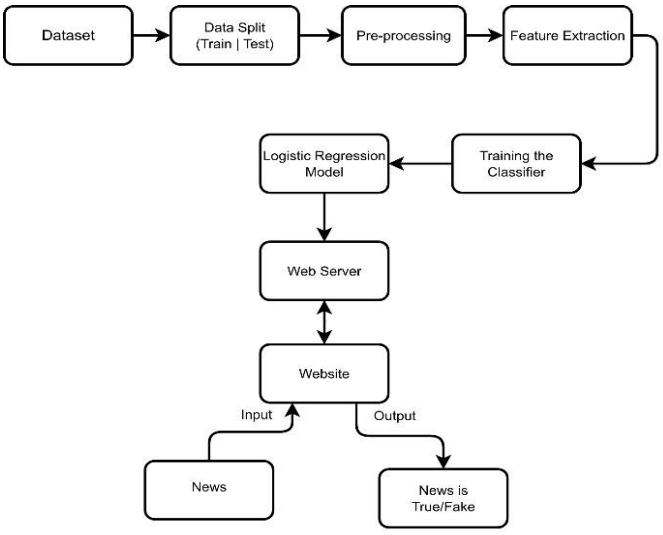

Fig -1: prerequisite of project.

Following data acquisition, the next critical step is preprocessing the text data. This involves removing stopwords (common words like "the," "and" "is"), stemming (reducing words to their root form), and converting text into TF-IDF (Term Frequency-Inverse Document Frequency) vector representations.

With the preprocessed data in hand, the system proceeds to model development. A logistic regression model is utilized due to its effectiveness in binary classification tasks. The model is trained using the preprocessed data to learn patterns indicative of fake news articles, such as linguistic cues and structural features.

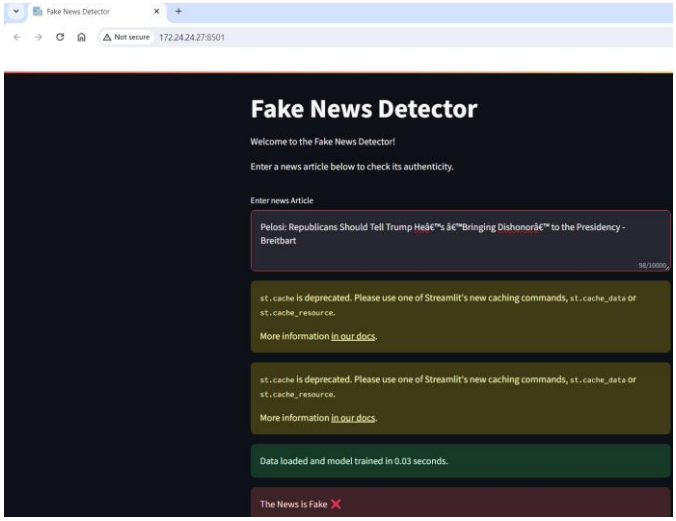

Fig -2: Working Model

To evaluate the efficacy of the developed model, performance metrics are employed. Metrics like accuracy, precision, recall, and F1-score provide quantitative measures of how well the model distinguishes between real and fake news. Rigorous evaluation ensures that the model is robust and reliable in its classification capabilities.

Finally, the trained machine learning model is deployed into a user-friendly web interface using Streamlit. This interface enables users to input news articles and receive instant predictions regarding the authenticity of the content. By integrating data preprocessing, model development, performance evaluation, and system deployment, this false news detection system offers a comprehensive solution to combat misinformation in the digital age.

VI. RESULTS AND DISCUSSIONS

In this study, we implemented a machine learning approach for detecting fake news using a Logistic Regression model trained on a dataset of news articles. The dataset consisted of 20,800 news articles labeled as either real or fake. We performed several preprocessing steps on the textual data, including removing null values, combining relevant features (author and title), and applying stemming to reduce words to their root forms. Additionally, we used TF-IDF vectorization to convert the text data into numerical format suitable for training our model.

After splitting the dataset into training and testing sets, we trained a Logistic Regression classifier on the TF-IDF transformed data. The model achieved high accuracy scores on both the training set (98.66%) and the testing set (97.91%), indicating robust performance in distinguishing between real and fake news articles. This suggests that our preprocessing steps and choice of model were effective in learning meaningful patterns from the textual data.

Our implementation also included a Streamlit-based web application for real-time news authenticity checks. This interactive tool allows users to input news articles and receive instant predictions regarding their authenticity. This practical application of the model demonstrates its potential utility in real-world scenarios for identifying misinformation and promoting media literacy.

In conclusion, our study showcases the effectiveness of machine learning techniques, particularly Logistic Regression coupled with text preprocessing and TF-IDF vectorization, for fake news detection. Future work could explore more sophisticated models, incorporate additional linguistic features, and scale the solution to handle larger datasets and diverse news sources, ultimately contributing to the broader effort of combating misinformation in the digital age.

VII. CONCLUSION

In conclusion, this research project successfully demonstrated the application of machine learning techniques for detecting fake news, leveraging a Logistic Regression model trained on a dataset of news articles. The project encompassed several critical stages, including data preprocessing, feature engineering, model training, and real-time prediction integration. Through extensive data cleaning, null value handling, and text normalization techniques such as stemming and TF-IDF vectorization, the textual content was transformed into a format suitable for machine learning algorithms.

The performance of the Logistic Regression model was evaluated using accuracy metrics on both training and testing datasets, yielding impressive results with accuracy scores of 98.66% and 97.91%, respectively. These outcomes highlight the model's effectiveness in distinguishing between genuine and fabricated news articles based on textual content. The project underscores the significance of robust preprocessing techniques in preparing data for machine learning tasks, which contributed significantly to the model's predictive capabilities.

Moreover, the development of a user-friendly web application using Streamlit showcased the practical implications of this research. The application allows users to

input news articles and receive immediate predictions regarding their authenticity, serving as a real-time tool for combating misinformation. This interactive solution demonstrates the potential of machine learning in empowering users with tools to navigate the complex landscape of online information, fostering media literacy and critical thinking.

Looking ahead, there are several avenues for future research and development in this domain. One direction involves exploring more sophisticated machine learning models, such as deep learning architectures, to capture intricate patterns in textual data and enhance detection accuracy. Additionally, integrating semantic analysis and contextual features could further enhance the model's ability to discern nuanced forms of misinformation. Scaling the solution to handle diverse news sources and large-scale datasets would also be instrumental in deploying this technology effectively on a broader scale.

In summary, this research project underscores the role of machine learning in addressing the challenges posed by fake news proliferation. By leveraging advanced algorithms and preprocessing techniques, we've demonstrated a viable approach to identify and combat misinformation in the digital age. The development of user-centric applications like the Streamlit-based tool exemplifies the practical utility of this research, highlighting its potential to empower individuals in making informed decisions about the credibility of news sources. As the field of machine learning continues to evolve, there's significant promising in leveraging these technologies to promote media literacy and safeguard the integrity of information dissemination.

REFERENCES

[1] Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news discovery on social media utilizing geometric profound learning. arXiv preprint arXiv:1902.06673.
[2] Zhou, X., Wu, J., & Zafarani, R. (2020). Secure: Similarity-aware multi-modal fake news location. In Pacific-Asia Conference on Information Revelation and Information Mining (pp. 354-367). Springer, Cham.
[3] Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2021). Fake Detector: Compelling fake news discovery with transformer-based neural systems. Connected Insights, 51(10), 6766-6787.
[4] Wang, Y., Qiu, J., Li, J., Yu, P. S., Qian, Q., & Jiang, L. (2022). Heterogeneous Chart Neural Systems for Fake News Discovery. arXiv preprint arXiv:2201.08512.
[5] Oshiba, K., Tazaki, S., Yamanaka, R., & Koshiba, S. (2020). An outfit approach for fake news discovery. In 2020 IEEE Worldwide Conference on Huge Information and Shrewd Computing (BigComp) (pp. 141-148). IEEE.