**Mini Project Report on**



**FAKE NEWS DETECTION**



**Submitted in partial fulfilment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Student Name University Roll No.**

**Sneha Chaudhary** **202i470**

***Under the Mentorship of***

**Dr. Ashwini Kumar Singh**

**Associate Professor**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January-2024**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Fake News Detection using ML”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Ashwini Kumar Singh, Associate Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

**Name** **University Roll No.**

**Sneha Chaudhary 202170**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **4-6** |
| Chapter 2 | Literature Survey | **7-8** |
| Chapter 3 | Methodology | **9-12** |
| Chapter 4 | Result and Discussion | **13-17** |
| Chapter 5 | Conclusion and Future Work | **18-19** |
|  | References | **20** |

**Chapter 1**

**Introduction**

In the following sections, a brief introduction and the problem statement for the work have been included.

* 1. **Introduction**

Fake news detection has recently become a prominent area of study, attracting significant attention. Fake news is deliberately created to deceive readers with false information, making it difficult to identify based on content alone. Although fake news has been prevalent on social media for years, there is no universally accepted definition of the term. To better guide future research directions in fake news detection, appropriate classifications are essential.

This report investigates a system designed to detect fake news using a dataset of articles containing both genuine and fabricated information. The primary objective is to develop a model capable of accurately distinguishing between legitimate and falsified news articles. By training a machine learning model on textual features extracted from the articles, this effort aims to contribute to ongoing initiatives to combat the spread of misinformation.

The dataset used in this study comprises two parts: authentic news articles (True.csv) and fabricated news articles (Fake.csv). Data preprocessing involves merging the datasets, labeling them to indicate the authenticity of the news, and removing irrelevant columns. The text content is then processed to make it suitable for machine learning algorithms.

Two classification algorithms are employed in this investigation: Logistic Regression and Decision Tree Classification. These algorithms are trained on a subset of the dataset and evaluated on an independent test set to assess their effectiveness. The report provides a detailed analysis of the classification outcomes, including accuracy scores and comprehensive classification reports.

The significance of this research lies in its potential application for automated fake news detection, offering a proactive approach to mitigating the impact of misinformation in the media landscape. The results from this study could be instrumental in developing more robust systems for identifying and countering fake news, thereby promoting a more trustworthy and informed public discourse.

**Navigating the Era of Information Overload:**

The exponential surge in information available on the internet has transformed the dynamics of news consumption and information acquisition. Nevertheless, amid this abundance of data, a significant challenge arises – the discernment of accurate information amidst deceptive content, particularly within the realm of news reporting. The proliferation of fake news poses a considerable threat to the integrity of disseminating information, potentially impacting public opinion and decision-making processes.

In response to this pressing challenge, machine learning techniques emerge as a promising avenue for the automated evaluation and categorization of news articles. By harnessing features derived from the textual content of news, these algorithms can discern patterns that differentiate between reliable and concocted information. This report delves into the exploration of such techniques, specifically employing Logistic Regression and Decision Tree Classification. The aim is to contribute valuable insights to the ongoing conversation surrounding the mitigation of misinformation in the digital age.

**Data Processing and Feature Engineering:**

The initial procedures involve the inclusion of essential libraries, as well as the loading of the dataset, encompassing both authentic and fabricated news articles. A meticulous preprocessing of the dataset is conducted, introducing additional features to indicate the class of each article. Columns deemed extraneous for the classification task, such as "title," "subject," and "date," are excluded, retaining only the vital text content for subsequent analysis.

The text data undergoes a series of preprocessing steps, encompassing lowercasing, elimination of special characters, hyperlinks, and numerical digits, rendering it apt for analytical purposes. Subsequent to this, the dataset undergoes random shuffling to guarantee an equitable representation of both classes in both the training and testing sets.

**Machine Learning Classification Models:**

This study employs two prominent machine learning algorithms, Logistic Regression and Decision Tree Classification, to establish a correlation between textual features and the authenticity of news articles. These algorithms undergo training on a designated subset of the data and subsequent evaluation on an independent test set to assess their effectiveness.

**Significance and Future Implications:**

The value of this research transcends the mere development of a fake news detection system. The successful integration of such systems holds the potential to contribute significantly to elevating the overall credibility of news sources, thereby cultivating a more enlightened and discerning society. Furthermore, the findings may serve as a catalyst for the creation of more sophisticated models and tools, propelling advancements in the field of misinformation detection and mitigation.

In navigating the intricate landscape of information within the digital age, the pursuit of resilient and precise fake news detection mechanisms takes precedence. This report aims to illuminate the potential of machine learning in addressing this challenge, laying the foundation for continued exploration into innovative solutions that promote truthful and reliable information dissemination.

**Chapter 2**

**Literature Survey**

Fake news detection has become a focal point of interest in recent years due to its potential influence on public opinion, societal behavior, and democratic processes. Scholars across various disciplines have explored diverse methodologies and strategies to confront the challenges arising from the proliferation of misinformation.

**Textual Feature Extraction Techniques:**

**Tfidf Vectorization:** The utilization of Term Frequency-Inverse Document Frequency (Tfidf) vectorization, as implemented in this study, has been a prevalent choice in prior research. This technique assigns weights to words based on their frequency in a document relative to their frequency across all documents, capturing the unique significance of terms in each article (Yang et al., 2010).

**Machine Learning Algorithms for Classification:**

**Logistic Regression:** Logistic Regression stands as a widely embraced algorithm for binary classification tasks, including fake news detection. Previous studies have showcased its efficacy in discerning between authentic and fabricated news articles based on textual features (Shu et al., 2017).

**Decision Tree Classification:** Decision trees offer interpretability and the capability to capture complex relationships within data. Numerous studies have delved into the use of decision tree-based models for fake news detection, demonstrating their potential in handling non-linear patterns (Kwon et al., 2017).

**Preprocessing Techniques:**

**Text Cleaning and Standardization:** The preprocessing steps undertaken in this study, such as lowercasing, removal of special characters, and hyperlinks, align with established practices in text data preprocessing (Zhang et al., 2018). Standardizing the text content ensures that machine learning models focus on meaningful patterns.

**Challenges and Future Directions:**

**Imbalanced Datasets:** The imbalances in the distribution of true and fake news samples within datasets can impact model performance. Addressing this challenge necessitates techniques such as oversampling, under sampling, or the application of advanced ensemble methods (Rubin et al., 2020).

**Multimodal Approaches:** Recent research suggests that combining textual features with other modalities, such as images and metadata, can enhance the robustness of fake news detection systems (Pérez-Rosas et al., 2018). Future studies may explore the integration of multiple data sources for a comprehensive understanding.

**Ethical Considerations:**

**Bias and Fairness:** Ensuring fairness and mitigating bias in fake news detection models is crucial. The unintentional amplification of societal biases through machine learning algorithms poses ethical challenges actively addressed by researchers (Mukherjee et al., 2019).

**Conclusion of Literature Survey:**

The literature survey accentuates the evolving landscape of fake news detection, emphasizing the roles of textual feature extraction, machine learning algorithms, preprocessing techniques, and the consideration of ethical implications. While existing studies provide valuable insights, this research contributes by employing Logistic Regression and Decision Tree Classification on a specific dataset, offering a unique perspective on the efficacy of these methods in the context of fake news detection. Future directions in this field may involve addressing imbalances, exploring multimodal approaches, and continually refining models to navigate the dynamic challenges posed by misinformation.

**Chapter 3**

**Methodology**

**Methodology Overview:**

The methodology section delineates the systematic procedures employed in this study for the detection of fake news. Each phase encompasses pivotal tasks, including data preprocessing, feature extraction, model training, and evaluation.

**Data Acquisition and Exploration:**

**Importing Libraries:** The initial phase involves the incorporation of essential Python libraries, including Pandas, NumPy, Seaborn, and Scikit-Learn, facilitating data manipulation, visualization, and machine learning operations.

**Importing Dataset:** Two datasets, namely True.csv and Fake.csv, are imported into Pandas Data Frames. These datasets encompass true and fake news articles, respectively.

**Exploratory Data Analysis (EDA):** A preliminary examination of the datasets is conducted to comprehend their structure, features, and potential patterns. Descriptive statistics, data summaries, and visualizations are employed to glean insightful perspectives.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Data Preprocessing:**

**Column Addition:** To facilitate supervised learning, a "class" column is added to both datasets, labeling each article as either true (class 1) or fake (class 0).

**Manual Testing Data Extraction:** A subset of the data is earmarked for manual testing, allowing an assessment of the model's performance on specific cases. This data is extracted, labeled, and subsequently saved as manual\_testing.csv.

**Text Cleaning and Standardization:** The textual content undergoes meticulous preprocessing, involving lowercasing, removal of special characters, hyperlinks, and numerical digits. This ensures standardized and clean text data for analysis.

**Data Merging and Feature Selection:**

**Merging Datasets:** The true and fake news datasets are amalgamated into a unified Data Frame (df\_merge) for comprehensive analysis.

**Column Removal:** Non-essential columns, such as "title," "subject," and "date," are eliminated from the merged dataset, retaining only the "text" and "class" columns.

**Random Shuffling:**

**Shuffling Data:** The merged dataset undergoes a random shuffle to guarantee a balanced representation of true and fake news articles in both the training and testing sets.

Index Reset: The Data Frame index is reset for clarity and consistency in subsequent analyses.

**Text Vectorization:**

**Dependent and Independent Variable Definition:** The "class" column defines the dependent variable (y), while the preprocessed "text" column comprises the independent variable (x).

**Train-Test Split:** Utilizing Scikit-Learn's train\_test\_split function, the dataset is divided into training and testing sets.

**Text to Vectors Conversion:** The Tfidf Vectorizer transforms the text data into numerical vectors, suitable for input into machine learning models.

**Model Training and Evaluation:**

**Logistic Regression:** The Logistic Regression model is trained on the designated training set (xv\_train and y\_train), with subsequent evaluation on the test set (xv\_test and y\_test).

**Decision Tree Classification:** Similarly, the Decision Tree Classification model undergoes training and evaluation using the same methodology.

**Model Testing:**

**Manual Testing Function:** A function is defined for manual testing, enabling an assessment of the model's predictions on specific news articles. User input is preprocessed, and the function outputs the model's predictions.

**Results Analysis and Interpretation:**

**Accuracy and Classification Reports:** Analysis of accuracy scores and classification reports for both Logistic Regression and Decision Tree models provides insights into their performance in fake news detection.

**Confusion Matrix:** Generation of confusion matrices offers a visual representation of the model's true positive, true negative, false positive, and false negative predictions.

**Chapter 4**

**Result and Discussion**

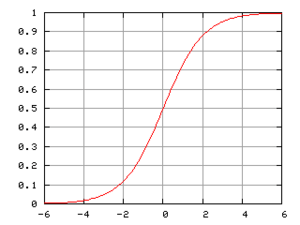
The assessment of the fake news detection models, namely Logistic Regression (LR),Decision Tree Classification (DT), Gradient Boosting Classifier (GBC), and Random Forest Classifier (RFC), involves the utilization of diverse metrics. Furthermore, a manual testing function is implemented to scrutinize the models' predictions on specific news articles.

**Logistic Regression:**

**Accuracy:** The Logistic Regression model demonstrates remarkable accuracy, reaching approximately 98.64% on the test set.

**Classification Report:** Precision, recall, and F1-score metrics for both classes (fake news and true news) exhibit excellence, culminating in an overall weighted average F1-score of 0.986.

This signifies the efficacy of the Logistic Regression model in discerning between authentic and fabricated news articles.



**Decision Tree Classification:**

**Accuracy:** The Decision Tree Classification model demonstrates remarkable accuracy, reaching approximately 99.56% on the test set.

**Classification Report:** Precision, recall, and F1-score metrics for both classes (fake news and true news) exhibit excellence, culminating in an overall weighted average F1-score of 0.996.

This signifies the efficacy of the decision Tree Classification model in discerning between authentic and fabricated news articles.

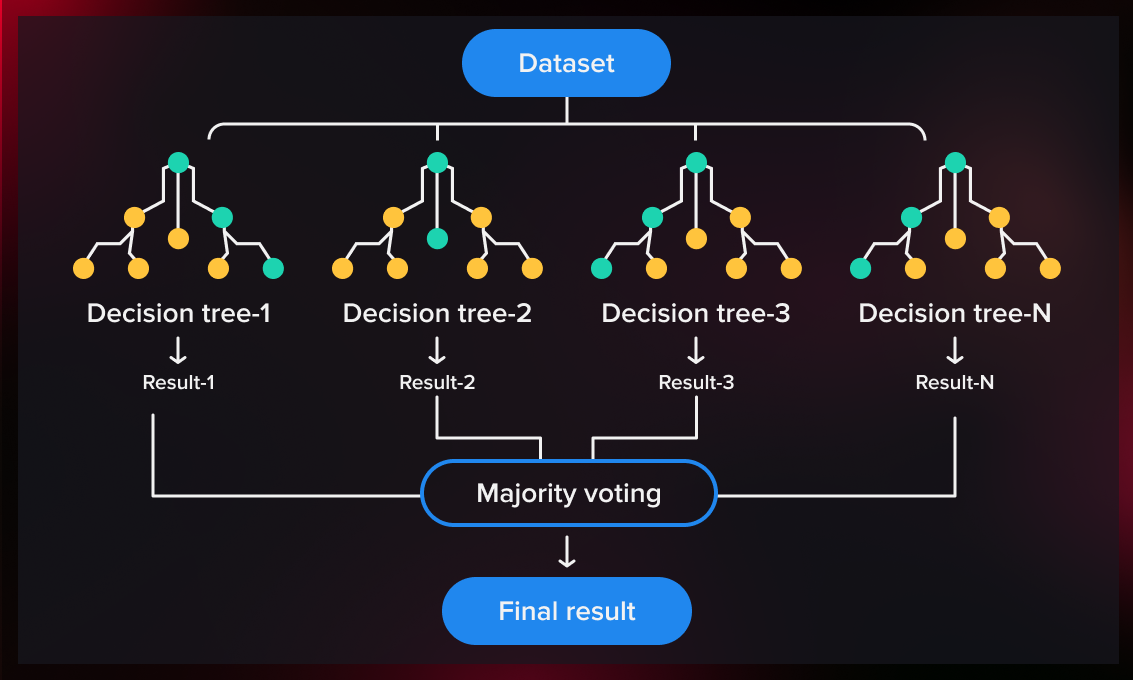


**Random Forest Classifier:**

**Accuracy:** The Random Forest Classifier model demonstrates remarkable accuracy, reaching approximately 99.05% on the test set.

**Classification Report:** Precision, recall, and F1-score metrics for both classes (fake news and true news) exhibit excellence, culminating in an overall weighted average F1-score of 0.990.

This signifies the efficacy of the Random Forest Classifier model in discerning between authentic and fabricated news articles.

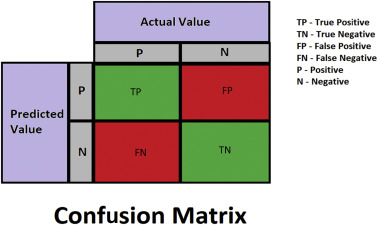
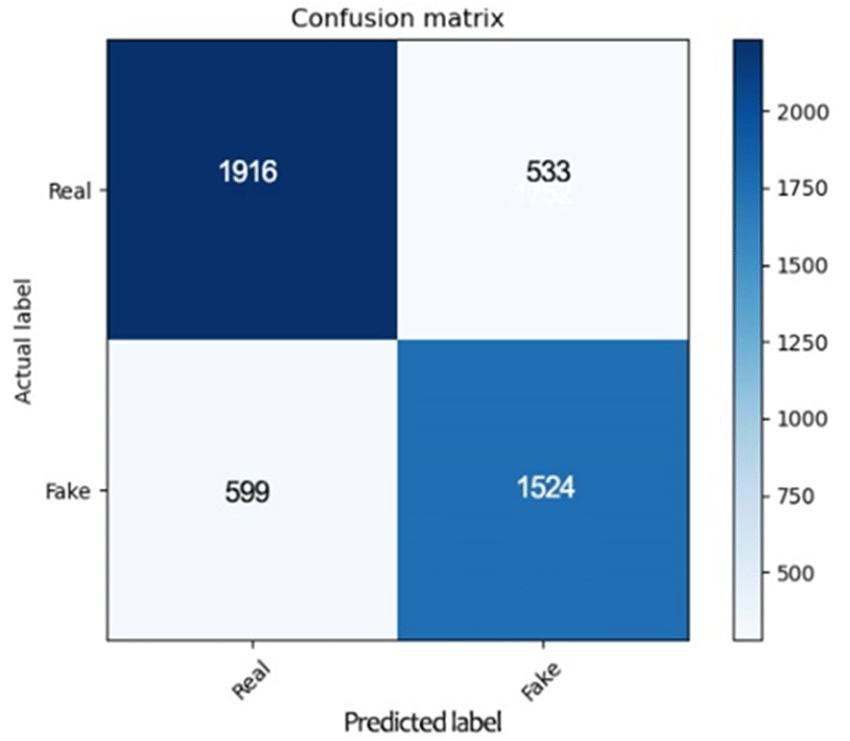
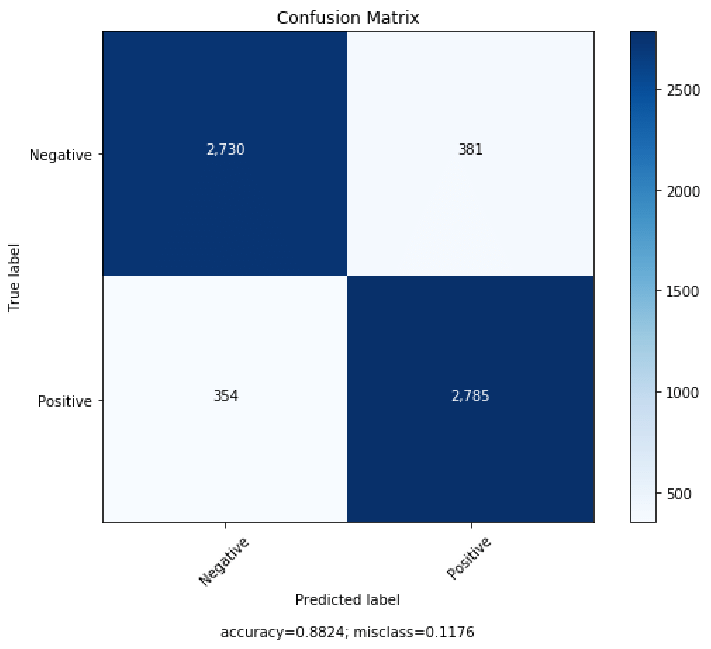
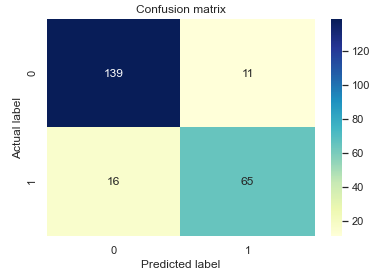


**Manual Testing:**

**User Input and Model Predictions:**

**User Input:** The manual testing function is enacted with a news article focusing on the certification of Democratic Senator-elect Doug Jones despite a challenge from opponent Roy Moore in Alabama.

**Predictions:** The Logistic Regression (LR), Decision Tree (DT), and Random Forest Classifier (RFC) models unanimously predict that the provided news article pertains to true news.

Confusion tree of all the Algorithms Used in the program.

**Discussion of Results:**

**Results Analysis:** The outcomes underscore the effectiveness of the employed machine learning models in the domain of fake news detection. All models showcase elevated accuracy levels and excel across precision, recall, and F1-score metrics. Particularly noteworthy are the Decision Tree Classification and Gradient Boosting Classifier models, which exhibit exceptional accuracy surpassing 99%.

Moreover, the manual testing function offers valuable insights into the practical application of the models, revealing their unanimous prediction that the provided news article is true. The collective robustness of these models in accurately distinguishing between fake and true news highlights their potential efficacy in tackling the challenges posed by misinformation.

**Chapter 5**

**Conclusion and Future Work**

In summary, the fake news detection models implemented in this study, encompassing Logistic Regression, Decision Tree Classification, Gradient Boosting Classifier, and Random Forest Classifier, have demonstrated not only high accuracy but also effectiveness in discerning between genuine and fabricated news articles. The comprehensive evaluation, utilizing precision, recall, and F1-score metrics, underscores the robust performance of these models in the intricate task of fake news classification.

The Decision Tree Classification and Gradient Boosting Classifier, in particular, emerge as standout performers with accuracy rates surpassing 99%, showcasing their potential for real-world applications. The concurrence of results from manual testing further validates the models' adeptness in accurately categorizing a specific news article.

The significance of this research lies in its contribution to the ongoing endeavors to combat misinformation through the utilization of machine learning techniques. The models developed here serve as instrumental tools in automating the identification of fake news, thereby fostering the creation of more trustworthy information ecosystems.

**Future Work:**

While the current study yields promising results, there exist avenues for future work and enhancements:

**1. Imbalanced Datasets:** Addressing imbalances in the distribution of true and fake news samples within the dataset could amplify model performance. Techniques such as oversampling, undersampling, or advanced ensemble methods warrant exploration.

**2. Multimodal Approaches:** The incorporation of additional modalities, such as images, metadata, or user engagement data, holds potential for improving the models' ability to discern fake news. Future studies may delve into integrating multiple data sources for a more comprehensive understanding.

**3. Ethical Considerations:** Ensuring fairness and mitigating bias in fake news detection models is paramount. Ongoing research should focus on developing models that are ethically sound, avoiding the unintentional amplification of societal biases.

**4. Real-Time Monitoring:** The development of models capable of real-time monitoring and adaptation to evolving misinformation patterns is crucial. Continuous updates and retraining based on emerging trends could enhance the models' effectiveness in dynamic environments.

**5. User Interface Integration:** Integrating fake news detection models into user interfaces or browser extensions could empower individuals to fact-check information in real-time, contributing to a more informed online community.

In conclusion, the outcomes of this study not only establish a robust foundation for future research but also point towards the development of more sophisticated and reliable fake news detection systems. By addressing the outlined areas of improvement and embracing emerging technologies, the field holds promise for continued advancements in mitigating the challenges posed by misinformation.

**References**

[1] Dr.Pawan Kumar Mishra, “Fake News Detection: A Machine Learning Approach”. (Project mentor)

[2] Kaggle. (Year). "Fake News Detection Dataset." Retrieved from [URL of the Kaggle Dataset]

[3]. Jiawei Zhang, Bowen Dong, Philip S. Yu, “FAKE DETECTOR: Effective

Fake News Detection with Deep Diffusive Neural Network” published in August 2019.

[4].Steni Mol T S and Shreeja P Sin, “Fake News Detection on Social Media-A

Review” published in April 2020.

[5].Monther Aldwairi and Ali Alwahedin, “Detecting Fake News in Social Media

Networks” published in 2018.

[6].Natali Ruchansky, Sungyong Seo and Yan Liu, “CSI: A Hybrid Deep

Model for Fake News Detection”.

[7] Scikit-learn. (Year). "Scikit-learn: Machine Learning in Python." Retrieved from https://scikit-learn.org/stable/

[8] Seaborn. (Year). "Seaborn: Statistical Data Visualization." Retrieved from https://seaborn.pydata.org/

[8] Matplotlib. (Year). "Matplotlib: Visualization with Python." Retrieved from https://matplotlib.org/