Fake News Detection

Project By:-

1. Sanju Kumari 23MCA048
2. Akansha Bharadwaj 23MCA004

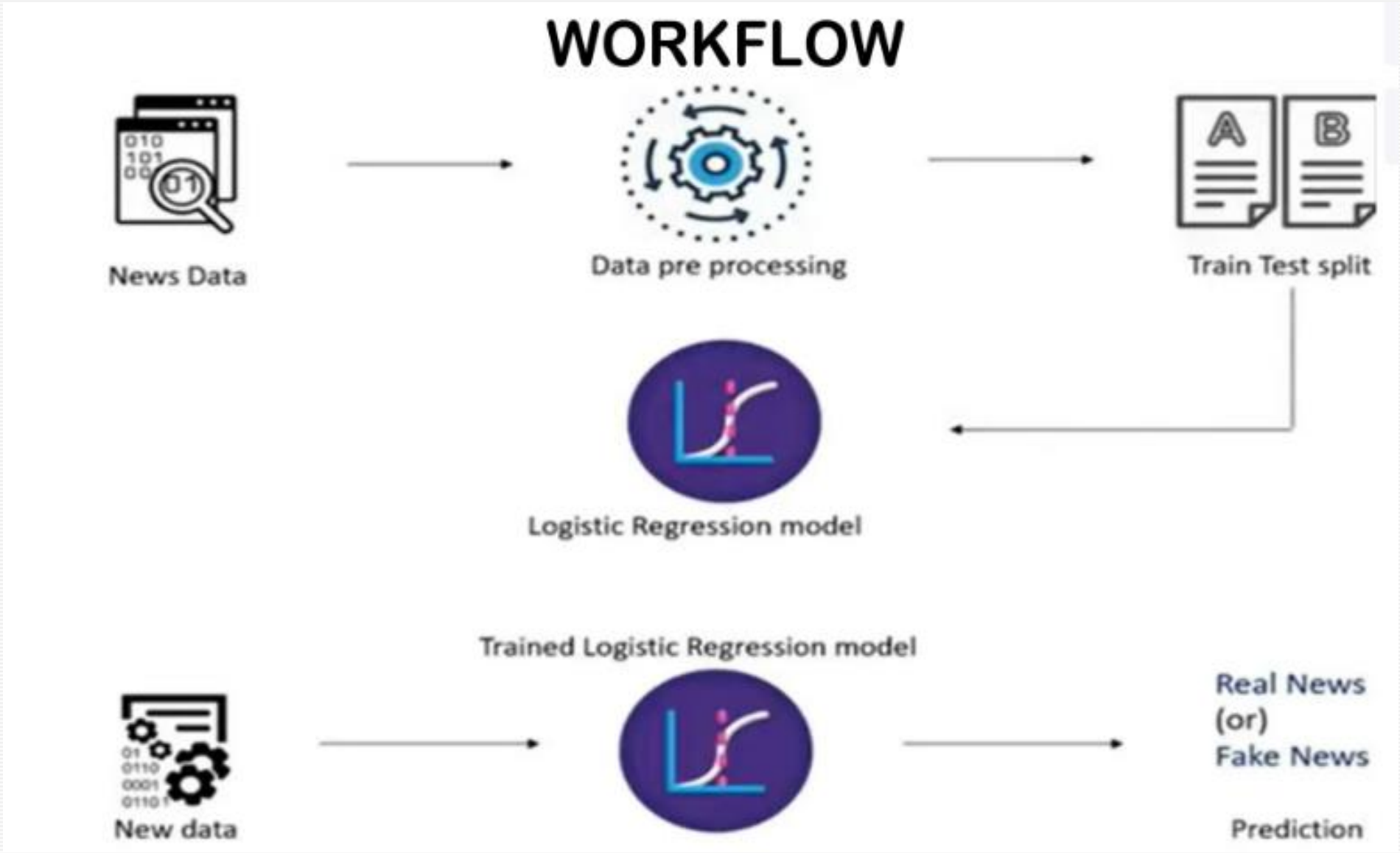
About Project

This project aims to combat the spread of misinformation using machine learning techniques. By analyzing linguistic patterns and features, we've developed a system to automatically detect fake news articles. Our goal is to promote a more informed society by filtering out unreliable information.

Introduction to Fake News

- Fake news is deliberately fabricated information presented as factual news to mislead readers.
- It can significantly influence public opinion, fuel anxiety, and manipulate political landscapes.
- Social media platforms exponentially amplify the reach and impact of fake news.

WORKFLOW


News Data
Data pre processing
A
B
Train Test split
Logistic Regression model
Trained Logistic Regression model
New data
Real News
(or)
Fake News
Prediction

Libraries Used in Fake News Detection Project

NumPy

Facilitates numerical operations and array handling for efficient data manipulation.

Pandas

Enables data loading, manipulation, and preprocessing in a tabular format.

scikit-learn

Offers machine learning algorithms and evaluation metrics for model development.

NLTK

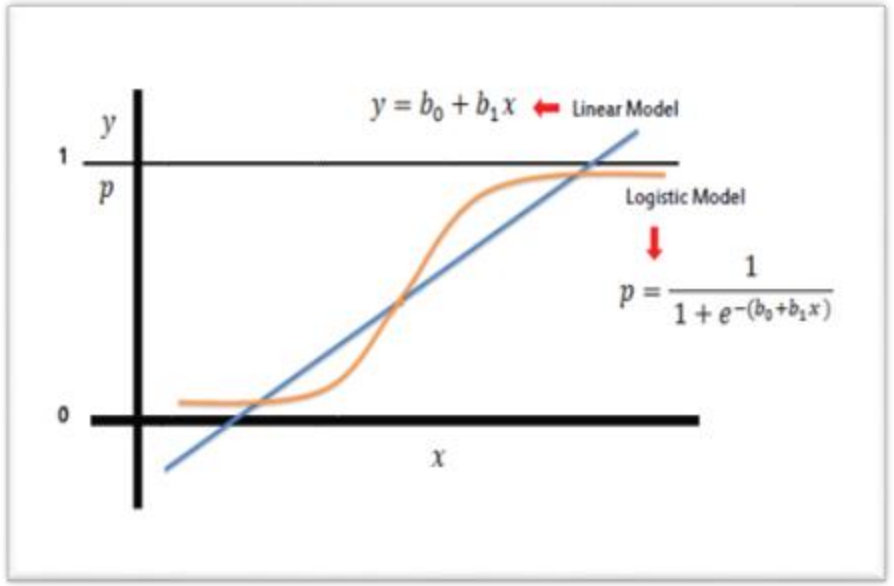
Provides essential tools for natural language processing tasks like tokenization and stemming.

Data Preprocessing:

Steps involved in preprocessing:

- Text Cleaning : Remove special characters, URLs, and convert text to lowercase for consistency.
- Tokenization : Break down text into individual words or tokens for further processing.
- Stop Word Removal : Eliminate common words like "the" and "and" that don't contribute to classification.
- Stemming : Reduce words to their root form to normalize variations (e.g., "running" to "run").
- TF-IDF Vectorization: Convert text data into numerical features, capturing word importance across documents

Model Selection : Logistic Regression



- Binary Classification

Ideal for fake news detection's two-class problem: real or fake.

- Probability Output

Provides likelihood scores, allowing flexible decision-making and threshold adjustments.

- Efficiency and Interpretability

Computationally efficient and easy to interpret, suitable for large datasets.

Model Training and Evaluation:

Data Split :
Divide dataset into training (80%) and testing (20%) sets for robust evaluation.

Model Training :
Feed training data to Logistic Regression model, optimizing parameters iteratively.

Result

- A high accuracy score suggests that the model is effective in detecting fake news within the dataset used. This validates the approach of using text-based features and Logistic Regression for binary classification.
- However, accuracy alone doesn’t capture all aspects of performance. Precision and recall, which measure false positives and false negatives, could also be explored to understand how well the model handles borderline or ambiguous cases.

Limitations And Future Work

- **Dataset Bias :** Model performance may vary on news from different sources or time periods.
- **Feature Limitation :** Reliance on content feature alone may miss important contextual or metadata cues.
- **Model Complexity :** More advanced models like deep learning could capture nuanced language patterns better.
- **Real-time Detection :** Developing a system for instant fake news detection on social media platforms.

Conclusion

This project demonstrates how machine learning, particularly Logistic Regression, can effectively classify news articles as fake or real based on their textual content. By processing and analyzing large volumes of text data, the model helps in distinguishing patterns associated with unreliable information.

- **Effective Detection**: The model provides an efficient and accurate method to identify fake news, supporting efforts to combat misinformation.
- **Practical Application**: This approach has real-world potential, as it could be integrated into news platforms and social media to filter content before it reaches a wider audience.
- **Future Improvements**: The model's performance can be enhanced by exploring advanced techniques like deep learning, expanding the dataset, or incorporating additional linguistic features.

References

- Dataset of fake news detection from **Kaggle.**
- **Text Mining : IBM** and **Qualtrics** emphasize the use of text mining techniques, such as NLP and ML, to transform unstructured text (like news articles and social media posts) into structured data. These methods help identify patterns and trends in text, making it easier to detect fake news.
- Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches by Andrei Korobeinikov, article on wiley online library website.

THANK YOU