Fake News Detection System using Machine Learning and Real-Time News Analysis

Atharva Bhuse, Harsh Jagtap and Gaurav Gaikwad
CSE Department, MIT School of Computing, MIT Art, Design and Technology University, Pune, India

Abstract:

In the digital information era, the massive and unregulated proliferation of news content on the internet has resulted in the growing menace of misinformation, disinformation, and fake news. Fake news refers to any intentionally fabricated or misleading content designed to resemble legitimate journalism. It spreads rapidly via social media platforms, blogs, news websites, and messaging apps, often deceiving users and manipulating public opinion. This phenomenon has far-reaching implications, from interfering in democratic processes to causing public panic and undermining trust in credible media. The need to develop systems capable of detecting such misleading content has never been more urgent. In response, this project introduces a comprehensive Fake News Detection System that combines Natural Language Processing (NLP) techniques and Machine Learning (ML) algorithms to classify news content as real or fake in real-time.

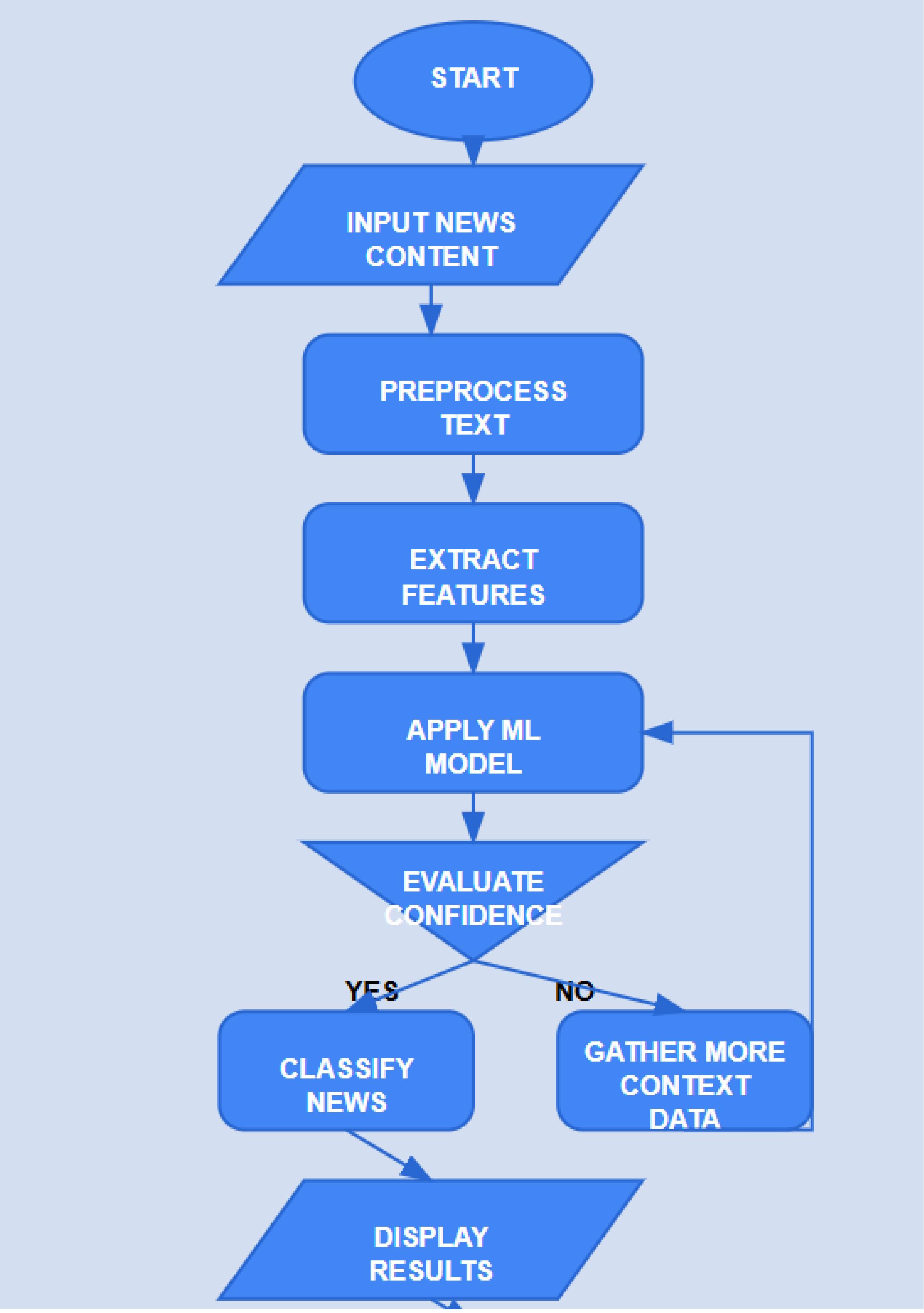
The objective of the system is to process raw text from news headlines or articles, clean and transform it into meaningful features using NLP, and then train a machine learning model to detect patterns indicative of fake or misleading content. The core of the system is built using Python and Scikit-learn libraries, employing TF-IDF (Term Frequency-Inverse Document Frequency) for text vectorization and a Random Forest Classifier for binary classification. The model is trained on a labeled dataset composed of both real and fake news headlines obtained from reliable repositories such as Kaggle’s “Fake and Real News Dataset.” This training allows the system to distinguish between real and fake news based on linguistic patterns, word distributions, and contextual cues.

The preprocessing stage is critical in ensuring model accuracy and efficiency. It includes steps such as converting all text to lowercase, removing punctuation, stopwords, hyperlinks, and numerical characters, and normalizing white space. This textual cleaning enhances the quality of input data, reduces noise, and helps in more accurate feature extraction. The TF-IDF vectorizer then converts the cleaned text into a numerical representation, capturing the significance of words relative to the entire dataset. These vectors are then passed into the machine learning classifier, which learns the correlation between textual features and their labels (real/fake).

We selected the Random Forest algorithm due to its robustness, high accuracy, and interpretability. It creates multiple decision trees from random subsets of the data and combines their outputs to improve prediction reliability and minimize overfitting. During experimentation, we performed data splitting using train_test_split, reserving 70% of the data for training and 30% for testing. Evaluation metrics such as accuracy, precision, recall, and F1-score were used to measure performance, with the model achieving over 90% accuracy on unseen test data.

Beyond static headline classification, this project integrates real-time functionality using the NewsAPI.org service. The system fetches the latest news headlines based on user-defined keywords such as "India", "Politics", or "War", along with their publication timestamps. These

real-time headlines are then automatically processed and classified by the trained model, enabling dynamic detection of misinformation as it appears in the news ecosystem. This makes the system highly practical and usable in real-world scenarios such as social media monitoring, digital journalism, and user education.



The system includes an intuitive **Graphical User Interface (GUI)** built using **Streamlit**, a lightweight and interactive web framework for Python. Users can either enter a news headline manually or choose to fetch the latest headlines through the integrated NewsAPI. The model then displays the classification result ("REAL" or "FAKE") along with the publication time. This approach ensures the system is not only functional but also user-friendly and accessible to a broader audience, including non-technical users.

Furthermore, the system supports optional logging and timestamping of results, offering users a simple record-keeping mechanism. This log can be useful for tracking trends, auditing system decisions, or sharing results with others. The system also emphasizes **data privacy**, ensuring that user-provided headlines are not stored or shared unless explicitly required. Since the core model and inference run locally, the application remains operable without an internet connection once the initial model is trained—making it suitable for deployment in offline or resource-constrained environments.

A notable strength of this project lies in its scalability and extensibility. While the current version uses TF-IDF and Random Forest for classification, the modular architecture allows easy integration of more advanced techniques such as **BERT**, **LSTM networks**, or **Transformers** for semantic analysis. Additionally, support for multilingual input can be added by leveraging multilingual NLP models or pre-trained embeddings like fastText or mBERT. These extensions can dramatically enhance the model's ability to generalize across different languages and writing styles.

This Fake News Detection System can be integrated with platforms such as **news websites**, **browser extensions**, and **social media monitoring tools**, or deployed within **educational institutions** and **fact-checking organizations** to enhance digital media literacy. In journalism, it can serve as a verification assistant to flag suspicious stories before publication. In academia, it can be used as a learning aid to demonstrate applied machine learning in real-world NLP problems.

In conclusion, this project demonstrates the practical application of machine learning and NLP to address a critical challenge in the modern information age. By developing a real-time, user-friendly, and accurate fake news detection system, we contribute a scalable solution to combat misinformation online. The system not only aids in promoting truthful journalism but also empowers everyday users to question and verify the credibility of what they read. Future improvements such as voice input support, chatbot-based interaction, deeper model architectures, and multilingual analysis can further enhance its relevance and impact.

Ultimately, the Fake News Detection System highlights how ethical artificial intelligence can be leveraged to foster an informed, vigilant, and digitally literate society.

MIT-ADT
UNIVERSITY
PUNE, INDIA
A leap towards World Class Education

MIT SCHOOL OF COMPUTING

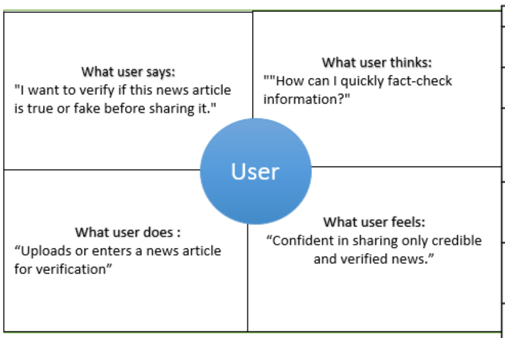
Class : SY7
Group Id: 03

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Fake News Detection System

ATHARVA BHUSE ,HARSH JAGTAP, GAURAV GAIKWAD

Faculty Guide:
Dr. P D Patil



What user says:
"I want to verify if this news article is true or fake before sharing it."

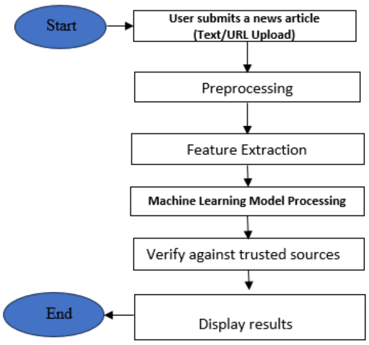
What user thinks:
""How can I quickly fact-check information?"

User

What user does :
“Uploads or enters a news article for verification”

What user feels:
“Confident in sharing only credible and verified news.”

Sr.no	Requirement	Solution
1	Accurate fake news detection	Implement NLP and ML models for text classification.
2	Real-time news verification	Use web scraping and APIs to cross-check news sources.
3	User-friendly interface	Develop a simple web app for easy news submission and results display.
4	Scalable and efficient processing	Use cloud-based infrastructure for handling large datasets..
5	Multi-platform accessibility	Ensure mobile and desktop compatibility with a responsive design
6	Reliable dataset for training	Use trusted datasets from sources like PolitiFact and Kaggle.



Start
User submits a news article (Text/URL Upload)
Preprocessing
Feature Extraction
Machine Learning Model Processing
Verify against trusted sources
Display results
End

Problem statement

The spread of misinformation and fake news on social media and online platforms has become a major concern. Users often struggle to differentiate between credible news and false information, leading to misinformation, public panic, and social unrest.

Proposed Solution

Develop a **Fake News Detection System** using **Python and Machine Learning** that automatically identifies and classifies news articles as **real** or **fake**. The system will use **Natural Language Processing (NLP)** and **Machine Learning (ML) models** to analyze article content, verify sources, and predict the authenticity of the news.

Scope and Feasibility

- **Users:** Social media users, journalists, fact-checkers, and the general public.
- **Feasibility:** The project requires dataset collection, model training, and web integration. It is feasible with open-source ML libraries like **Scikit-Learn, TensorFlow, and NLTK**.
- **Future Enhancements:** AI-based source verification, multilingual news detection, and integration with browser extensions or mobile apps.