MIT School of Engineering

Department of Computer Science and Engineering

Project Synopsis

Group ID:SY0703

Project Title: "Fake News Detection System"

Group Members:

Enrollment Number	Roll No.	Name of student	Email Id	Contact Number
ADT23SOCB0396	18	Gaurav Gaikwad	gauravgaikwad.gg2707@gmail,.com	9209777074
ADT23SOCB0250	15	Atharva Bhuse	atharvabhuse2005@gm ail.com	9322821030
ADT23SOCB0415	19	Harsh Jagtap	harshjagtap0407@gmail.com	7498467733

Problem Statement:

In the era of digital communication, the rapid spread of misinformation poses significant threats to public opinion, political stability, and societal trust. Fake news, often shared on social media platforms and online news outlets, can mislead users and cause widespread panic or influence public decisions unfairly.

The goal of this project is to develop a machine learning-based system that can accurately classify news articles as either "Real" or "Fake" based on their textual content. This system will utilize Natural Language Processing (NLP) techniques to extract relevant features from the news content and apply classification algorithms to detect fake news in real time.

Abstract:

This project presents a machine learning-based **Fake News Detection System** that identifies and classifies news content as either real or fake using textual analysis. The system employs Natural Language Processing (NLP) techniques such as tokenization, stop-word removal, and stemming to clean and standardize text. TF-IDF (Term Frequency-Inverse Document Frequency) is used to convert the text into numerical feature vectors. A supervised learning algorithm—**Random Forest Classifier**—is trained on a labeled dataset comprising real and fake news articles to learn linguistic patterns associated with deceptive content.

Literature Survey:

Fake news detection has become a prominent area of research in recent years due to the explosive growth of misinformation across digital platforms. Researchers from various disciplines, including computer science, journalism, and data science, have proposed numerous approaches to automatically identify fake news using machine learning and natural language processing techniques. These studies aim to improve the accuracy, efficiency, and scalability of systems that can help prevent the spread of misleading or false information online.

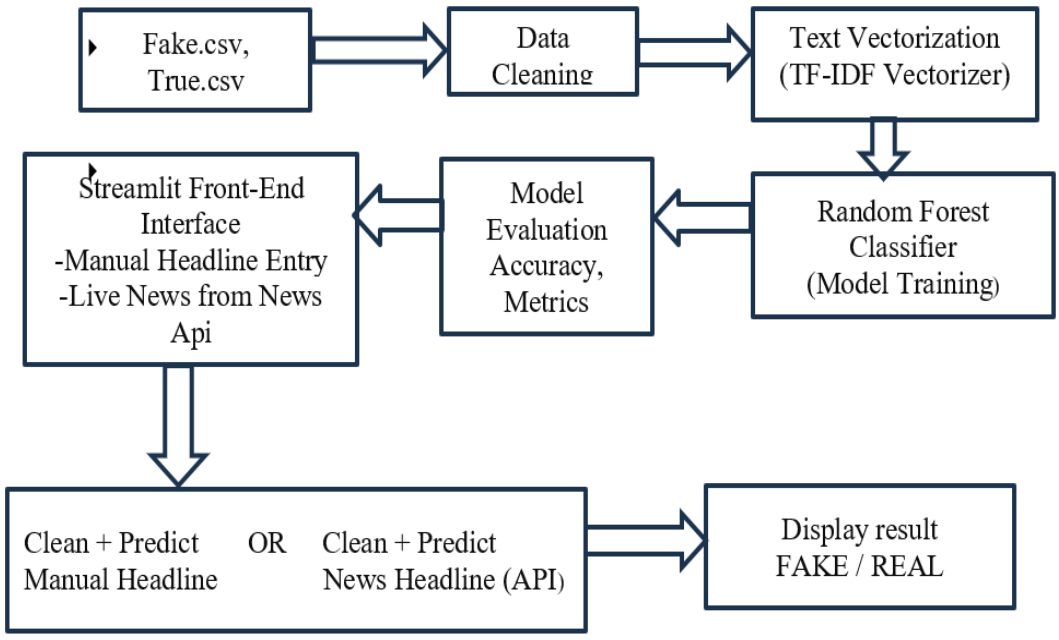
One of the early approaches to fake news detection involved the application of traditional machine learning models such as Support Vector Machines (SVM), Logistic Regression, and Naïve Bayes classifiers. In a 2018 study by (4) Ajao, Bhowmik, and Zargari, machine learning models were trained on Twitter datasets containing fake and real news tweets. Their research found that SVM outperformed the other models in terms of accuracy, especially when combined with text-based features like word counts and n-grams. These early efforts laid the groundwork for using algorithmic techniques to classify news content based on textual patterns.

Later studies focused on using more advanced feature extraction techniques like TF-IDF (Term Frequency-Inverse Document Frequency) to represent textual data. In 2020,

(5) Roy et al. demonstrated the effectiveness of using TF-IDF combined with a Random Forest classifier to detect fake news articles from Kaggle's Fake and True News dataset. Their method involved a comprehensive preprocessing pipeline including text cleaning, stopword removal, and vectorization. The study showed that even with relatively simple models, high accuracy could be achieved by focusing on quality feature engineering.

With the emergence of deep learning models, researchers began using Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, to handle the sequential nature of textual data. A major milestone came with the introduction of the LIAR dataset by (6) William Y. Wang in 2017, which consisted of over 12,000 labeled political statements. The study showed that LSTM models, capable of understanding context and temporal dependencies in sentences, provided better performance than traditional ML models when trained on larger datasets.

Proposed System (Block Diagram):



Conclusion:

The increasing spread of fake news across digital platforms poses a serious challenge to information reliability and public trust. This project proposes an effective and scalable fake news detection system using Natural Language Processing and Machine Learning techniques. By leveraging text preprocessing, feature extraction (like TF-IDF), and classification algorithms (such as Logistic Regression or Random Forest), the system can accurately differentiate between real and fake news articles.

The proposed solution not only aids in curbing misinformation but also serves as a foundation for building more advanced, real-time fake news detection systems. Future enhancements can include deep learning models, source verification, and multilingual support, making the system more robust and widely applicable.

References:

- Ahmed, H., Traore, I., & Saad, A. (2018). "Detecting Opinion Spam and Fake News Using Machine Learning: A Survey." *International Journal of Computer Applications*, 180(20), 1-9. doi:10.5120/ijca2018916447
- Wang, W. Y. (2017). "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection." *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 3181-3188.
- Ruchansky, N., Seo, S., & Liu, Y. (2017). "CSI: A Hybrid Deep Learning Approach for Fake News Detection." *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 797-806. doi:10.1145/3132847.3132893
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of NAACL-HLT 2019*, 4171-4186. doi:10.18653/v1/N19-1423
- Shu, K., Sliva, A., Wang, S., Liu, H., & Yu, S. (2019). "Fake News Detection on Social Media: A Data Mining Perspective." *ACM SIGKDD Explorations Newsletter*, 19(1), 22-36. doi:10.1145/3137597.3137600

Annexure I: Form A - Title Approval

Title of the Project: "Fake News Detection System"

Department Approval:

✓Title is relevant to current technology and aligned with course objectives.
✓Project scope is well defined and achievable within the timeline.

Approved by:

Guide Name: _____
Signature: _____
Date: _____

Annexure II: Form B - Market and Financial Feasibility

Market Feasibility

- **Target Market**: Social media platforms, news agencies, government bodies, educational institutions, and end-users.
- **Market Demand**: Increasing misinformation on digital platforms and regulatory pressure drive the demand for fake news detection systems.
- **Competitor Analysis**: Key competitors include **Factmata** and **Fake News Detector**, with the proposed system offering a real-time, customizable solution as a competitive advantage.

2. Financial Feasibility

- **Development Costs**: Includes cloud services, data collection, machine learning tools, and development team expenses.
- **Revenue Models**: Subscription-based service (SaaS), licensing to news agencies, and custom enterprise solutions.

Feasibility Verified by Guide:
Name: _____
Signature: _____
Date: _____

Annexure III: Literature Survey Papers

1. **"Fake News Detection on Social Media: A Data Mining Perspective"**
2. **Authors**: Shu et al. (2019)
3. **Summary**: Discusses content-based and network-based approaches for fake news detection on social media, emphasizing the need for real-time detection and context analysis.
4. **"Detecting Fake News on the Web: A Data Mining Perspective"**
5. **Authors**: Chang et al. (2018)
6. **Summary**: Overviews content-based, network analysis, and hybrid approaches for detecting fake news, with a focus on big data analytics.
7. **"LIAR: A Benchmark Dataset for Fake News Detection"**
8. **Author**: Wang (2017)
9. **Summary**: Introduces the LIAR dataset and compares classification techniques for fake news detection.
10.