Marwadi University
Marwadi Chandarana Group

NAAC A+

Major Project-II (01CE0807)

Review 1

(08/11/2025)

Explainable Multilingual Fake News Detection using Text + Image Features

Team ID: 8PRO_161

Devanshee Parmar (92200103267) (8TC5)
Janvi (92200103256) (8TC2)
Sudhriti Bari (92200103256) (8TC5)

Guided By

Prof. Monisha Mohan

Department of Computer Engineering,

Faculty of Engineering & Technology

Outline

1. ABSTRACT
2. INTRODUCTION
3. PROBLEM STATEMENT
4. LITERATURE REVIEW
5. OBJECTIVE
6. KEY TECHNOLOGIES INVOLVED
7. DATASET DESCRIPTION
8. SYSTEM ARCHITECTURE
9. METHODOLOGY
10. EVALUATION METRICS
11. RESULT ANALYSIS
12. MODEL EXPLAINABILITY
13. CHALLENGES & LIMITATIONS
14. FUTURE ENHANCEMENTS
15. CONCLUSION
16. REFERENCES

Abstract

- The rapid spread of fake news on digital platforms has increased public misinformation and reduced trust in media. This project presents an Explainable Fake News Detection model that classifies news articles as real or fake using text analysis. TF-IDF is used to convert news content into numerical features, and a Logistic Regression model performs the classification. To ensure transparency, LIME(Local Interpretable Model-Agnostic Explanations) is integrated to highlight the words influencing each prediction. The model, trained on True.csv and Fake.csv datasets, achieves 98.97% accuracy while remaining interpretable, making it a reliable solution for automated news verification.

Introduction

- Over the past few years, online platforms have become a primary source of news for most people. While this has made information easier to access, it has also increased the spread of misleading and fabricated news. Fake news often appears convincing, making it difficult for readers to identify what is real and what is not, which can mislead the public and create unnecessary confusion.

- Due to the large volume of information shared every day, manually verifying news is neither practical nor efficient. This project aims to build a machine learning-based system that can automatically classify news articles as real or fake by analyzing their text content. The focus is not only on accuracy, but also on providing clear explanations for the model's decisions to ensure transparency and trust.

Problem Statement

- **Challenge:** The rapid spread of fake news across digital and social media platforms has made it difficult to distinguish between real and fabricated information.
- **Limitations:** Existing detection methods focus mainly on accuracy but lack transparency, making it hard for users to trust automated predictions.
- **Gap:** There are limited explainable AI solutions that provide clear reasoning behind news classification, especially for text-based misinformation.
- **Need:** A reliable, interpretable, and data-driven system that can automatically detect fake news while explaining the features influencing its decisions, thereby improving user trust and understanding.

Literature Review

Marwadi University
Marwadi Chandarana Group
NAAC A+

RESEARCH PAPER (YEAR)	TECNOLOGY	DATASET	ACCURACY	ARCHITECTURE	KEY CHALLENGES	APPLICATION
r/Fakeddit: new large multimodal benchmark — provides large multimodal dataset for fine-grained FND and baseline results. (2019)	Dataset release, baseline DL models	Fakeddit (~1M posts)	Baselines: varies (binary up to ~87 reported for some models).	Multimodal dataset (text + image + metadata); baselines (text/image CNN + fusion)	Noisy distant labels; multimodal alignment.	Benchmarking multimodal models.
Entity-enhanced multimodal fusion — model that extracts visual entities to model text–image correlations (entity inconsistency, mutual enhancement, embedded image text). (2021)	Entity extraction, transformers, multimodal fusion	Fakeddit, other multimodal sets	Improves SOTA (paper reports significant gains vs baselines)	Entity-oriented multimodal alignment & fusion network (EMAF)	Reliable entity recognition in noisy images; cross-modal linking.	Social media moderation, fact-checking
Multimodal Fake News Detection (fine-grained study) — shows images help certain fake categories and reports multimodal vs unimodal performance. (2021 /2022)	CNNs, BERT for text + CNN image encoders	Fakeddit (full)	Multimodal CNN approach reported ~87% (task dependent).	Unimodal (BERT / CNN) and multimodal CNN fusion; fine-grained 6-way taxonomy	Fine-grained label noise; class imbalance.	Detection & taxonomy analysis.
ETMA: Efficient Transformer Multilevel Attention — hierarchical attention for each modality then joint attention. (2022)	Transformers, multi-level attention	Twitter, Pontes, other datasets	Reports SOTA improvements (dataset dependent)	Visual attention encoder + textual attention encoder + joint attention fusion	Overfitting small datasets; modality imbalance.	Real-time moderation, platform filtering
SEMI-FND: Stacked ensemble multimodal inference — stacked ensemble for faster, parameter-efficient multimodal FND. (2022)	Ensemble (BERT+ELECTRA + NasNet Mobile for images)	Twitter MediaEval, Weibo	~85.8% (Twitter), ~86.8% (Weibo) reported.	Stacked ensemble combining text transformers + CNN image model	Ensemble interpretability; latency vs accuracy tradeoffs.	Fast scanning, edge deployments.

Literature Review

Marwadi University
Marwadi Chandarana Group
NAAC A+

RESEARCH PAPER (YEAR)	TECNOLOGY	DATASET	ACCURACY	ARCHITECTURE	KEY CHALLENGES	APPLICATION
Progressive fusion multimodal model — progressive alignment & fusion of text & image representations to capture cross-modal cues. (2023)	Deep fusion networks, attention	Multiple social datasets	Shows improved F1/accuracy vs naive fusion (dataset dependent)	Progressive fusion layers aligning text-image features gradually	Cross-modal semantic gap; hallucination.	Social media monitoring.
Decision uncertainty for multimodal FND — uses uncertainty estimation to improve robustness and detect out-of-distribution samples. (2024)	Bayesian/uncertainty-aware DL, multimodal fusion	Social datasets (paper)	Improves reliability (metrics reported per dataset)	Uncertainty-weighted multimodal fusion with rejection/flagging	Calibrating uncertainty across modalities.	Trustworthy pipelines, triage systems.
Explainable multimodal fake posts detection (attention XAI) — attention-based feature extraction + attention maps for explaining predictions. (2023)	Attention mechanisms + XAI visualizations	Custom multimodal dataset (human & AI generated mixes)	Reported improvements; exact % dataset dependent	Feature extraction with attention; attention maps show text/image cues	Attention ≠ explanation; faithfulness issues.	Forensic analysis, user transparency.
Automated & interpretable detection (ensemble + ELA for images) — hybrid of classical ML for text + CNN for images + XAI. (2022)	Ensemble ML (NB, RF, DT), CNN for image ELA, XAI tools	Aggregated fake/real article corpora; CASIA for image forensics	Reported strong performance vs baselines (paper)	Hybrid pipeline: text ensemble + CNN image forgery detection + XAI explanation output	Generalizing ELA across formats; adversarial manipulation.	Newsrooms, forensic teams.
Survey: Multi-modal misinformation detection (comprehensive review) — taxonomy of datasets, features, fusion strategies & open challenges. (2022)	Survey (analysis)	N/A (survey)	N/A (survey)	Taxonomy: data collection → feature studies → model fusion & evaluation	Fragmented datasets; reproducibility.	Roadmap for researchers & practitioners.

Objective

- To design a machine-learning model capable of accurately detecting fake news.
- To convert text data into meaningful features using TF-IDF Vectorization.
- To train and evaluate a Logistic Regression classifier for text-based classification.
- To integrate LIMEfor explaining model predictions and improving transparency.
- To visualize model performance through graphs such as the confusion matrix and explainability plots.

Key Technologies Involved

- **Python** - Programming language used for implementation
- **Pandas & NumPy** - Data preprocessing and manipulation
- **Scikit-learn** - Machine learning model building and evaluation
- **TF-IDF Vectorizer** - Feature extraction from text data
- **Logistic Regression** - Classification model used
- **LIME(Local Interpretable Model-Agnostic Explanations)** - Model interpretability and explanation tool
- **Matplotlib** - Visualization of results and confusion matrix

Dataset Description

- The dataset used in this project is the Fake and Real News Dataset sourced from Kaggle.

- It contains two CSV files:
  True.csv – news articles verified as real.
  Fake.csv – news articles identified as false or misleading.

- Both datasets are combined into a single labeled dataset where:
  Label 0 = Real news
  Label 1 = Fake news

- Each file includes attributes such as title, text, subject, and date.

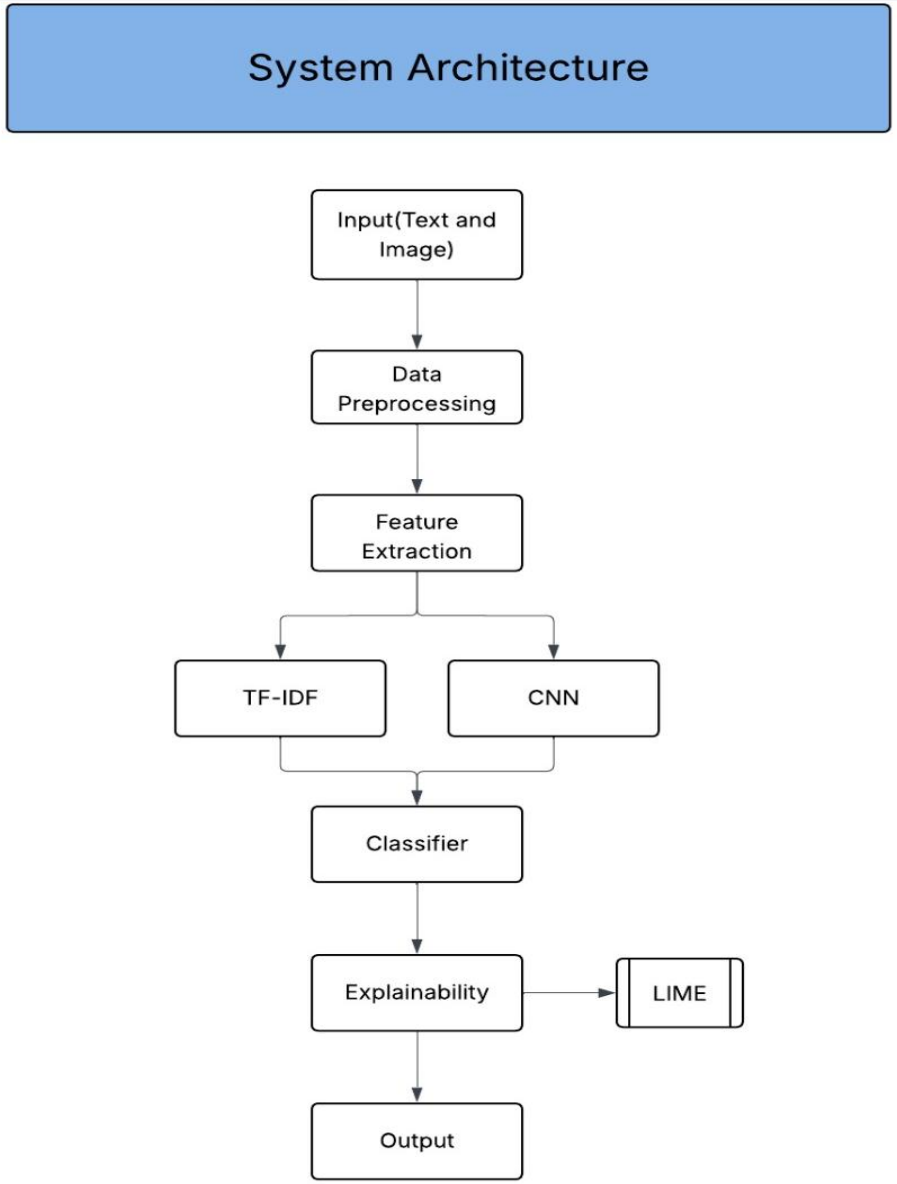
Dataset Description

- The data was cleaned and preprocessed through the following steps:
  - Converting all text to lowercase.
  - Removing URLs, special characters, and punctuation marks.
  - Eliminating extra spaces and standardizing text format.

- After merging both datasets, the final data contained approximately 23,481 articles (12,999 real and 10,482 fake), ensuring a balanced representation of real and fake news.

- Currently, the dataset includes English-language news articles only; multilingual data will be added in the next project phase.

System Architecture



Methodology

1. **Data Collection**: Import and combine real and fake news datasets.
2. **Preprocessing**: Clean the text data using regex to remove unwanted characters, links, and noise.
3. **Feature Engineering**: Apply TF-IDF Vectorizer to convert textual data into numerical form, capturing the relevance of each term.
4. **Model Training**: Use Logistic Regression to classify articles into “real” or “fake.”
5. **Evaluation**: Performance measured using **Accuracy, Precision, Recall, F1-score and Confusion matrix.**
6. **Explainability Layer**: Implement **LIME** to interpret model predictions and visualize influential features.
7. **Visualization & Output**: Generate graphs and explanation plots to analyze model behavior.

Evaluation Metrics

- Accuracy

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

- Precision

$$Precision = \frac{TP}{(TP + FP)}$$

- Recall

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

- F-1

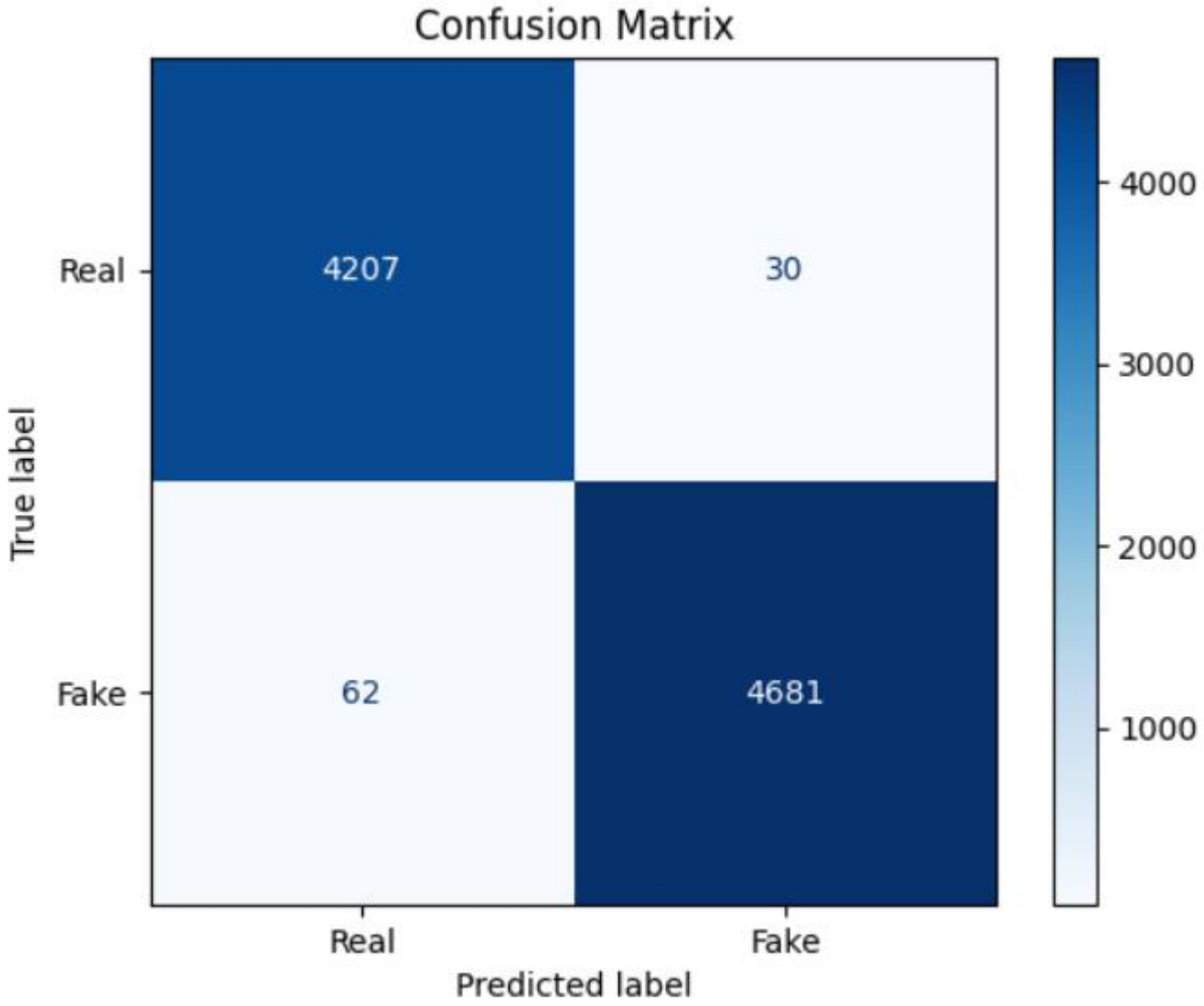
$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)}$$

- Confusion Matrix - Shows classification results in terms of TP,TN,FP,FN.

Evaluation Metrics

	Precision	Recall	F1 Score	Support
Real News (0)	0.99	0.99	0.99	4237
Fake News (1)	0.99	0.99	0.99	4737
Accuracy			0.99	8980
Macro Avg		0.99	0.99	8980
Weighted Avg		0.99	0.99	8980

Evaluation Metrics



Result Analysis

- The Logistic Regression model trained using TF-IDF features achieved an accuracy of 98.97%, demonstrating strong performance in distinguishing real and fake news.
- The confusion matrix indicates very few misclassifications, proving model reliability.
- Logistic Regression is lightweight, fast to train, and easy to interpret, making it more practical than heavy models like SVM or Deep Learning.
- The integration of LIME enhances trustworthiness by highlighting keywords and factors influencing each prediction.

Result Analysis

Example Predictions :

```
Text: brussels reuters us vice president mike pence said on monday he was disappointed the former white house national security adviser michae
Actual: 0 Predicted: 0
--------------------------------------------------------------------------------
Text: jakarta reuters thousands protested outside the us embassy in the indonesian capital on sunday against us president donald trump s decis
Actual: 0 Predicted: 0
--------------------------------------------------------------------------------
Text: as more time passes after the violent riots related to the white nationalist rally that took place yesterday in charlottesville virginia
Actual: 1 Predicted: 1
--------------------------------------------------------------------------------
Text: washington reuters us commerce secretary wilbur ross said on thursday he hopes to start the day countdown clock to launch a renegotiatio
Actual: 0 Predicted: 0
--------------------------------------------------------------------------------
Text: joe piscopo is hysterical he was on with neil cavuto and broke out with imitations of waters sanders and schiff and then you got maxine
Actual: 1 Predicted: 1
```

Model Explainability using LIME

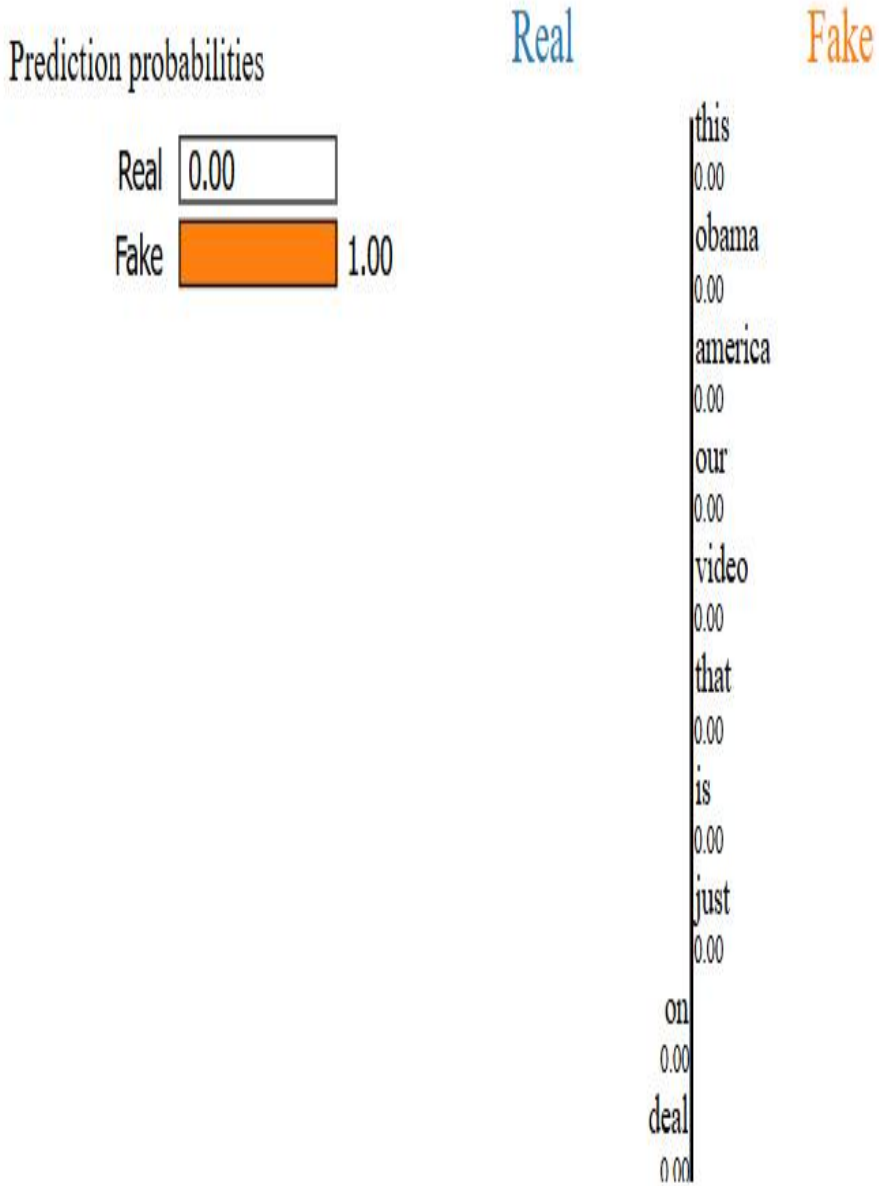
Why Explainability is Needed?

- Helps us understand why the model predicts news as Real or Fake.
- Builds trust in the system, especially for real-world or media use.

What LIME Does:

- Highlights specific words that influenced the prediction
- Shows whether those words pushed the model towards “Real” or “Fake”
- e.g. LIME highlights influential words contributing to predictions such as ‘fake’, ‘claims’, or ‘reportedly’.”

Model Explainability



Text with highlighted words

mark levin lays out obama s onesided unconstitutional iran nuke deal in this easy to understand video in the video below he points out the danger we as a nation face with this reckless president and the cowardly republican majority congress who refuses to stand up to him and his radical transformation of america obama is patting himself on the back for making what he seems to think is a great deal with our long time enemy iran the truth is this deal does nothing for america and it does everything to help iran get a nuclear bomb the goal of this deal was to dismantle iran s nuclear program completely it doesn t do that on that point alone it s a failed deal this deal allows iran to continue enriching uranium and it drops the sanctions that have been working well until this pointalso if the us decides to do a surprise inspection we have to give them days notice so what do we get out of this historic deal nothing we get an enemy with more money and more power to build a bombas mark levin points out in this video obama just planted the seeds for world war iii this president has done nothing but help our enemies and he should be convicted of treason for it there is a reason iranians are cheering it s because they know that america just bowed to them what obama has done today goes far beyond politics he put our lives and our children s lives at stakevia viralsneak

Challenges & Limitations

- The dataset is limited to English, restricting its applicability to multilingual environments.
- The model relies solely on text content and does not consider images, user behavior, or social media patterns.
- Logistic Regression may not capture deeper contextual meaning as effectively as modern transformer-based models.
- LIME explanations are local to individual predictions and do not provide a complete global understanding of the model.
- This phase currently supports English-language text only; multilingual expansion is planned in the next phase.

Future Enhancements

- Short-Term (Next Phase)
  - Implement deep learning models like LSTM, Bi-LSTM, or BERT for better context understanding.
  - Add advanced explainability methods such as SHAP or attention-based visualization.
- Mid-Term
  - Introduce multi-lingual support to detect fake news in regional languages.
  - Include image verification using computer vision to identify manipulated media.

Future Enhancements

- Long-Term

  - Integrate with social media APIs (Twitter, YouTube) for real-time fake news detection.
  - Develop a dashboard or browser plugin for live monitoring and explainable predictions.
  - Publish a research paper based on the project's findings and experimental results.

References

[1] Kaggle, “Fake and Real News Dataset,” [Online]. Available: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset . [Accessed: Nov. 8, 2025].

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[3] L. Buitinck, G. Louppe, M. Blondel, et al., “API Design for Machine Learning Software: Experiences from the scikit-learn project,” ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’ Explaining the Predictions of Any Classifier,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1135–1144, 2016.

[5] J. Ramos, “Using TF-IDF to Determine Word Relevance in Document Queries,” Proceedings of the First Instructional Conference on Machine Learning, 2003.

[6] C. M. Bishop, Pattern Recognition and Machine Learning, New York, NY: Springer, 2006.

[7] W. McKinney, “Data Structures for Statistical Computing in Python,” Proceedings of the 9th Python in Science Conference (SciPy), pp. 51–56, 2010.

[8] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.

[9] R. S. Pressman, Software Engineering: A Practitioner’s Approach, 8th ed., New York, NY: McGraw-Hill, 2014.

[10] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” SIGKDD Explorations, vol. 19, no. 1, pp. 22–36, 2017.

THANK YOU