Mathematics For Computing II

Elements of Computing II

FAKE NEWS DETECTION USING LINEAR CLASSIFICATION

Group Members:

Diya Prakash-CB.SC.U4AIE24111

Dondluru Keerthana-CB.SC.U4AIE24112

V.R. Sridevi- CB.SC.U4AIE24166

INTRODUCTION

- Fake news has become a significant societal issue, especially on social media.
- Traditional fact-checking is slow, while deep learning models require high computational power.
- This project presents a lightweight, mathematical approach using linear classification to detect fake news.

OBJECTIVE

Primary Goal:

Develop an efficient and scalable fake news detection model using linear classification techniques.

Key Objectives:

- Apply SVM, Logistic Regression, PCA to classify news articles.
- Improve detection accuracy using feature selection and dimensionality reduction.
- Implement a mathematical classification system.
- Train and test the model using real-world datasets.

RELEVANCE

This project is highly relevant due to the rise of fake news and the need for scalable, efficient solutions. Unlike deep learning models, this approach can work on low-resource systems and be deployed for real-time fact-checking.

Why Is It Important?

- Governments and organizations need automated tools to fight misinformation.
- Social media platforms struggle with manual fact-checking.
- A scalable approach using linear models can be integrated into real-world applications.

ADVANTAGES

1. **Fast and Efficient** – Works quickly and doesn't need high computing power.
2. **Easy to Train** – Requires less data and trains faster than complex AI models.
3. **Simple to Understand** – The decision-making process is clear and interpretable.
4. **Useful in Real Life** – Can help in stopping misinformation on social media and news platforms.
5. **Scalable** – Can be easily applied to large amounts of data, making it useful for big platforms.

LITERATURE REVIEW BASED ON PAPER 1

1. Methodology

• Compares seven ML models: SVM, Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, XGBoost, and Artificial Neural Networks.

• Uses TF-IDF and Word Embeddings for text feature extraction.

• Splits dataset into 80% training, 20% testing and evaluates models on accuracy, precision, recall, and F1-score.

2. Advantages

- High Accuracy (94%) – SVM with TF-IDF outperforms other models.
- Computationally Efficient – Faster and requires less computing power than deep learning.
- Scalable & Interpretable – Suitable for large-scale fake news detection.

LITERATURE REVIEW

3. Limitations & Disadvantages

- Limited to English datasets – Doesn't test multilingual news.
- No Image-Based Detection – Only works for text classification.
- Feature Engineering Required – Needs manual text processing (e.g., TF-IDF tuning).

4. Metrics Used

- Accuracy (94%) – Measures overall classification performance.
- Precision, Recall, F1-score – Evaluates fake vs. real news detection.

5. Dataset Used

📂 Kaggle Fake News Challenge, LIAR Dataset, BuzzFeed News.

6. Why This Paper?

- Best accuracy (94%) with a lightweight model suitable for real-world applications.
- Simple yet effective methodology, making it easy to implement and present.

LITERATURE REVIEW BASED ON PAPER 2

1. Methodology

The paper compares Logistic Regression (LR), Na¨ıve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Deep Neural Networks (DNN) for fake news detection. The process includes dataset collection, text preprocessing (stemming, stopword removal, NLP feature extraction), and model training/testing to classify news as real or fake.

2.Advantages • DNN achieves the highest accuracy (91%), outperforming other models. • Fast processing time compared to conventional ML techniques. • Effective feature extraction using NLP techniques enhances detection

LITERATURE REVIEW

3. Limitations & Disadvantages

- Limited to English datasets – Doesn’t test multilingual news.
- No Image-Based Detection – Only works for text classification.
- Feature Engineering Required – Needs manual text processing (e.g., TF-IDF tuning).

4. Metrics Used

- Accuracy (94%) – Measures overall classification performance.
- Precision, Recall, F1-score – Evaluates fake vs. real news detection.

5. Dataset Used

📂 Kaggle Fake News Challenge, LIAR Dataset, BuzzFeed News.

6. Why This Paper?

- Best accuracy (94%) with a lightweight model suitable for real-world applications.
- Simple yet effective methodology, making it easy to implement and present.

RESEARCH GAP

WHAT IS LACKING?

- Deep Learning models require large datasets & expensive computing power.
- Lack of interpretability in deep models.

WHY EXPLORE THIS APPROACH?

- Linear classification provides high accuracy with lower computational cost.
- Easier to deploy for real-time fake news detection.

PROBLEM STATEMENT

- The rapid spread of misinformation and fake news on online platforms, especially social media, has become a major societal issue
- Traditional fact-checking methods are slow and ineffective for handling large-scale information. Existing fake news detection methods rely on deep learning, which often requires extensive computational resources.
- This project proposes a mathematical approach using linear classification and feature reduction techniques to detect fake news more efficiently, even with limited computational power.

METHODOLOGY

The project follows a structured methodology that includes data collection, pre-processing, feature extraction, classification, and evaluation:

Step 1: Data Collection

- Collect news articles from Kaggle's Fake News Dataset, BuzzFeed News, or FactCheck.org.
- Extract headlines, content, and source credibility for feature extraction.

Step 2: Data Preprocessing

- Convert text into numerical representations using TF-IDF (Term Frequency-Inverse Document Frequency).
- Apply stopword removal, stemming, and tokenization to clean the text.

Step 3: Feature Extraction

- Represent news articles as vectors in a high-dimensional space.
- Use Principal Component Analysis (PCA) or Singular Value Decomposition (SVD) for feature reduction.

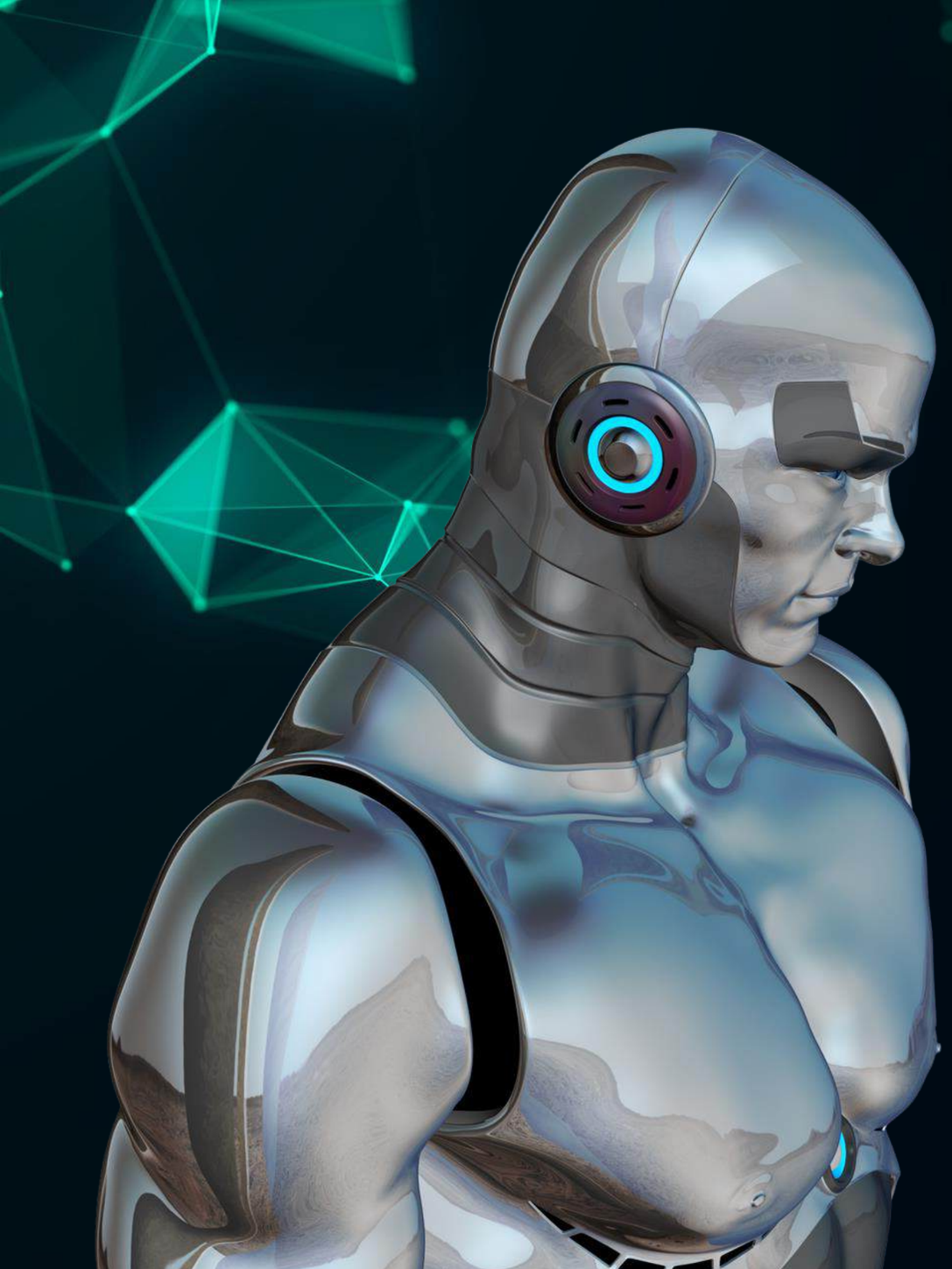
METHODOLOGY

Step 4: Classification Using Linear Models

- Implement Support Vector Machines (SVM) or Logistic Regression to classify articles.
- Optimize hyperparameters for better accuracy

Step 5: Model Evaluation

- Use accuracy, precision, recall, and F1-score to evaluate performance.
- Compare results with deep learning models to analyze efficiency.



NOVELTY

- Hybrid Approach – Integrates text + image analysis for better accuracy.
- Lightweight & Efficient – Uses SVM + TF-IDF, requiring low computational power compared to deep learning.
- Feature Reduction – PCA removes noise, improving accuracy and scalability.
- Scalable for Social Media – Can be extended for real-time fake news detection on platforms like Twitter and Facebook.
- Multi-Language Potential – Can be adapted beyond English using TF-IDF

HOW IS IT DIFFERENT FROM EXISTING SOLUTIONS?

Features	Existing Fake News Models	Proposed Model
Approach	Deep Learning (CNN, LSTM)	Linear Classification (SVM, PCA)
Data Requirement	Requires large labelled datasets	Works with small labelled datasets
Computational Cost	High (requires GPUs)	Low (runs on standard CPUs)
Interpretability	Low (black-box nature)	High (understandable model)
Scalability	Difficult for smaller applications	Easy to integrate in real-world systems

APPLICATIONS

This project has wide-ranging applications, including:

Social Media Platforms

- Can be integrated into Facebook, Twitter, and Instagram to filter fake news posts.

Journalism & News Agencies

- Helps organizations like BBC, Reuters, and The New York Times to verify news sources.

Government and Policy Making

- Assists law enforcement and fact-checking agencies in combating misinformation campaigns.

DATASET DETAILS

Title: LIAR data set

W**hat is it:** This is an media viral which gives news about the other media websites and posts.

Why its chosen: It shows the fake news of any one the posts in social media

source: Buzz feed website

Title: Fake news net dataset

W**hat is it:** Give a clear graph about the news published in the political region.

Why is it chosen: Get to know mainly about the content in the political conferences.

source: Kaggle website

PROPOSED TIMELINE

Research & Data Collection (Weeks 1-2)

- Review existing fake news detection methods.
- Gather datasets (Kaggle, BuzzFeed News, FactCheck.org).

Preprocessing & Feature Extraction (Weeks 3-4)

- Clean and tokenize text, apply TF-IDF.
- Perform dimensionality reduction using PCA/SVD.

Model Implementation (Weeks 5-6)

- Train and optimize SVM & Logistic Regression models.

Evaluation & Comparison(Weeks 7-8)

- Measure accuracy, precision, recall, and F1-score.
- Compare results with deep learning models.

Documentation & Final Presentation (Weeks 9-10)

- Compile findings, challenges, and future improvements.
- Prepare the final report and presentation.

CONCLUSION

This project presents an efficient and scalable fake news detection system using linear classification techniques. By leveraging mathematical modeling and computational methods, we develop a solution that is:

- Efficient and lightweight, requiring minimal computational resources.
- Scalable, making it deployable for real-time fake news detection.
- Mathematically robust, utilizing concepts from Linear Algebra (MFC) and Elements of Computing (EOC).

This project will significantly contribute to the ongoing research on automated misinformation detection, making online information more reliable and trustworthy.

THANK YOU!