

Fake News detection using NLP and Ensemble Machine Learning

Abhishek Tate

Department of Computer Science

Dr. D. Y. Patil Arts, Commerce & Science College, Pimpri, Pune, Maharashtra, India

abtate2002@gmail.com

Shreyas Walhekar

Department of Computer Science

Dr. D. Y. Patil Arts, Commerce & Science College, Pimpri, Pune, Maharashtra, India

shreyas4walhekar@gmail.com

ABSTRACT

The development of the internet has resulted in an increased number of online news platforms, which has raised issues related to the rapid spread of fake news. Misinformation may affect people's opinions and beliefs and cause instability in various sectors such as the economy, politics, and social welfare. Manual verification and conventional techniques are inefficient approaches to fake news detection.

The combination of artificial intelligence and natural language processing offers a highly efficient approach to detecting fake news. The latter involves techniques that extract meaningful information, linguistic characteristics, and contextual details from news sources and statements. Artificial intelligence improves the detection process through the development of machine learning algorithms that learn from huge volumes of news data and classify it into true and fake.

The present research explores an AI-based fake news detection technique that utilizes natural language processing and machine learning algorithms. It describes the operational methodology of the system in terms of data preparation, extraction of TF-IDF features, and classification of news using different models, including Logistic Regression, XGBoost, LightGBM, and Voting Ensemble. In addition, the paper analyzes the performance indicators, benefits, drawbacks, and future potential of the proposed system.

KEYWORDS

Fake News Detection, Artificial Intelligence (AI), Natural Language Processing (NLP), Machine Learning, Text Classification, Misinformation Detection, TF-IDF, Ensemble Learning, Data Mining, Cybersecurity

OBJECTIVE OF THE STUDY

In order to achieve the above objectives, the central purpose of this paper will be to come up with an intelligent tool capable of detecting fake news through analyzing text content. Because of the deceptive nature and manipulation involved in the creation of fake news, the main approach will involve using Natural Language Processing techniques in the search for suspicious language and semantic anomalies in text.

Besides, another purpose of this research is to analyze various classification algorithms in a bid to discover which method produces accurate classification results. In addition, efforts will be directed towards enhancing the accuracy rate of the detection process without increasing the number of false detections.

Moreover, there is also the need for exploring linguistic and contextual characteristics that are normally employed in the creation of fake news. In addition, there will be an effort to exploit such characteristics in improving the accuracy of the models in classification.

Lastly, one other objective of this research is to create a reliable model capable of helping people to distinguish between authentic and fake news.

INTRODUCTION

As the usage of internet is growing fast, information dissemination has been amplified immensely. On the other hand, the issue of spreading fake news and misinformation is becoming a big challenge. Fake news consists of misleading and false information, which is passed to people under the guise of genuine news. Such articles look very convincing; hence, it becomes difficult to ascertain their authenticity.

Conventionally, human verification methods and rule-based approaches have been used for fake news detection. However, since such strategies do not deal with vast amounts of information efficiently, there has been the need for advanced methodologies for fake news detection such as Artificial Intelligence and Natural Language Processing techniques.

Natural Language Processing allows machines to learn, comprehend and process language. With this technology, the system examines the text in each article, analyzes linguistic patterns within the text and detects if there is any misguiding information present. AI models play a vital role in learning from the vast dataset for effective detection of fake news.

LITERATURE REVIEW

Many researchers have worked on fake news detection using AI and NLP techniques.

- **Oshikawa et al. (2020)** presented a comprehensive survey on fake news detection using NLP techniques. Their study highlighted the importance of linguistic features and advanced deep learning models for identifying misleading content.
- **Shu et al. (2017)** proposed a data mining approach for fake news detection by analyzing both textual content and social context. The study showed that combining multiple features improves detection accuracy.
- **Nasir et al. (2021)** developed a hybrid deep learning model using CNN and RNN architectures for fake news detection. Their model achieved higher accuracy by capturing both local and sequential text patterns.
- **Prachi et al. (2022)** focused on detecting fake news using NLP-based preprocessing and feature extraction techniques. Their study compared machine learning and deep learning models and showed improved performance with advanced models.
- **Alshuwaier (2025)** presented a comprehensive review of fake news detection techniques using machine learning and deep learning. The study highlighted challenges such as dataset imbalance and generalization issues.

These studies indicate that AI and NLP significantly improve fake news detection accuracy. However, challenges such as data imbalance, evolving misinformation patterns, and lack of generalization across different domains still exist.

METHODOLOGY

Data Collection

Data is collected from various sources such as:

- Fake news datasets (e.g., LIAR Dataset, FakeNewsNet, ISOT Dataset)
- News article repositories and online news websites
- Social media posts and real-world news samples
- Fact-checking websites and verified news sources

Data Preprocessing

Before analysis, the data is cleaned and prepared:

- Removal of stop words
- Tokenization (splitting text into words)
- Stemming and lemmatization
- Removal of special characters

Feature Extraction using NLP

Important features are extracted using NLP techniques:

- Term Frequency-Inverse Document Frequency (TF-IDF)

- Bag of Words (BoW)
- N-grams (unigrams, bigrams)
- Detection of misleading or biased keywords

Model Implementation

Various AI models are used:

- Machine Learning Models:
 - Logistic Regression
 - Support Vector Machine (SVM)
 - Random Forest
 - XGBoost
- Ensemble Learning Models:
 - Voting Classifier
 - LightGBM

System Workflow

1. Input news article / claim
2. Preprocessing using NLP
3. Feature extraction
4. Model classification
5. Output: Real or Fake News

RESULTS AND DISCUSSION

Accuracy Enhancement

Artificial intelligence-based approaches provide better accuracy than conventional approaches. Machine learning and ensemble models deliver a high degree of accuracy in detecting fake news, often reaching more than 90%.

Real-Time Identification

The framework is capable of identifying fake news in real-time by processing input news articles or claims, enabling individuals to avoid misinformation.

Lower Rate of False Positives

NLP technology lowers the rate of false positives in genuine news stories.

Limitations and Challenges

- High computation costs
- Necessity for big data sets
- Constantly changing trends in misinformation
- Privacy-related issues with data

Future Prospects

Future frameworks may include:

- Fake news detection in real-time on social media websites
- Multilingual detection of fake news
- Integration with cloud-based verification systems
- Use of sophisticated AI algorithms such as GPT and transformers

CONCLUSION

Fake news detection through NLP and AI is a valuable technique that can solve present-day problems in digital media. It allows us to detect misinformation and protect users from its negative effects. The application of NLP technology makes it easier to analyze textual data, whereas AI algorithms increase accuracy and flexibility.

Despite the fact that there may be some difficulties like excessive computational complexity and changing misinformation, constant development in AI and machine learning will make it possible to create even more efficient systems to detect fake news.

REFERENCES

- [1] Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A Survey on Natural Language Processing for Fake News Detection.
- [2] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective.
- [3] Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection using a hybrid CNN-RNN based deep learning approach.

- [4] Prachi, Habibullah, Rafi, M., Alam, S., & Khan, M. (2022). Detection of Fake News Using Machine Learning and Natural Language Processing Algorithms.
- [5] Alshuwaier, F. A. (2025). Fake News Detection Using Machine Learning and Deep Learning Algorithms: A Comprehensive Review and Future Perspectives.