MisInfoCheckXAI: Mis-Information Detection Framework using linguistic features and An Explainable AI

Makhan Kumbhkar[1], Shashi Kant Gupta[2]

[1*]PostDoc Fellow, Lincoln University College, Malaysia
kumbhkar010385@gmail.com, ORCID: 0000-0001-9241-5331

[2*]Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia,
raj2008enator@gmail.com , ORCID: 0000-0001-6587-5607

Keywords: Misinformation, SMOTE, TF-IDF, Linguistic Features

ABSTRACT

The rapid spread of misleading online information requires detection systems that are both accurate and interpretable. This research present MisInfoCheckXAI**,** an explainable Mis- information detection framework based on linguistic features including Disclosure, Pragmatic, and Integration using Logistic Regression. The model combines TF-IDF representation, SMOTE for data balancing, and SHAP-based explainable AI to provide transparent decision insights. Experimental results achieved an accuracy of 0.8562 with strong macro precision, recall, and F1-score, demonstrating superior classification performance on the machine learning and deep learning approaches, the proposed MisInfoCheckXAI**,** framework provides a computationally efficient and interpretable solution for trustworthy Mis-information detection using Explainable AI.

Introduction

Fake content or news denotes to information that is fabricated, manipulated, or presented in a misleading manner to influence users or public perception [6-7]. With the rapid growth of digital platforms and social media, the spread of misinformation has become a major global challenge affecting public trust, healthcare decisions, political communication, and social stability. As a result, automatic misinformation detection has emerged as an important research area within Natural Language Processing (NLP), machine learning and deep learning, where linguistic based patterns,

sentiment cues, and textual features are analyzed to distinguish between real and fake information [2-3]. Earlier research primarily relied on traditional machine learning models such as Naïve Bayes, Support Vector Machines, Decision Trees, Random Forest and Logistic Regression using lexical features like Bag-of-Words and TF-IDF [4-6]. More recently, deep learning approaches including LSTM, CNN, and transformer-based architectures have achieved higher predictive performance by capturing contextual relationships in text[8-9] However, many of these models operate as black-box systems, making their decision process difficult to interpret and reducing transparency and trust in real-world applications where explainability is essential.

To address these challenges, this study proposes MisInfoCheckXAI, an explainable Mis-Information detection framework that integrates Disclosure, Pragmatic, and Integration (DIP) linguistic feature groups with an interpretable Logistic Regression classifier. The MisInfoCheckXAI, framework combines TF-IDF semantic representation, sentiment-based pragmatic analysis, and disclosure-based metadata features, while SHAP-based grouped explainability provides transparent insights into model decisions. By balancing prediction performance with interpretability and computational efficiency, the proposed approach offers a practical and trustworthy alternative to complex deep learning-based Mis-Information detection systems.

Related work

Rapid growth of social media and web platforms has increased the spread of Mis-Information detection and misleading content, creating a need for automated detection systems [6-7]. Early research focused on rule-based and manual fact-checking approaches**,** which were interpretable but lacked scalability and performance. Traditional machine learning models such as Naïve Bayes, SVM, Decision Tree, Random Forest, and Logistic Regression improved detection accuracy using lexical features like Bag-of-Words and TF-IDF [1, 6, and 7]. Researchers introduced pragmatic linguistic features including sentiment, subjectivity, and writing style to capture emotional and deceptive language patterns. Recent studies adopted deep learning models such as LSTM, CNN, and transformer-based architectures (e.g., BERT) for better contextual understanding, but these models

are computationally expensive and often behave as black-box systems [8]. To improve transparency, Explainable AI (XAI) techniques such as SHAP have been explored to interpret model decisions and feature contributions [9-10]. Combining interpretable machine learning with linguistic features such as Disclosure, Pragmatic, and Integration (DIP) provides a balanced approach that maintains accuracy while improving explainability and practical usability.

Reference	**Machine Learning / Deep Learning Used**	**Linguistic Features Used**	**Explainable AI (XAI) Included**
Makhan Kumbhkar et al. (2024)	Yes	No	No
Li Zhang (2025)	Yes	No	No
Omar Bashaddadh et al. (2025)	Yes	Yes	No
N. Rai et al. (2022)	Yes	Yes	No
A. B. Athira et al. (2023)	Yes	Yes	Yes
V. U. Gongane et al. (2024)	Yes	Yes	Yes
MisInfoCheckXAI	Yes	Yes (DIP Features)	Yes

METHODOLOGY

4.1 Data Collection and Pre-processing

The dataset used in this study collected from Politifact and contains labelled statements categorized as real or Mis-Information along with metadata such as author name, date, and source URL. Data cleaning performed such as removing duplicate, incomplete records, normalizing text to lowercase, eliminating URLs and punctuation, and applying tokenization and stop-word removal. These steps ensured consistent and noise-free input for feature extraction and model training.

4.2 DIP Feature Extraction

The proposed MisInfoCheckXAI**,** framework extracts three complementary linguistic feature groups such as Disclosure features capture source credibility using attributes such as author availability and domain length, Pragmatic features analyse linguistic behaviour through sentiment polarity, subjectivity score, and exclamation usage to identify emotional or exaggerated writing styles and Integration features represent semantic text patterns using TF-IDF with n-gram configuration, enabling the model to learn contextual word importance. All feature groups are combined to form a unified representation of each statement.

4.3 Model Training

A Logistic Regression classifier was selected due to its interpretability and computational efficiency. The combined DIP feature divided into training and testing sets to maintain class distribution. SMOTE was applied during training to address class imbalance and improve model generalization before generating predictions on unseen data.

4.4 Explainable AI using SHAP

To enhance transparency, SHAP (SHapley Additive exPlanations) was integrated to interpret model behaviour at both feature and group levels. Individual feature analysis highlights the influence of sentiment scores, domain properties, and important TF-IDF terms on predictions. Additionally, grouped SHAP values aggregate Disclosure, Pragmatic, and Integration features, providing clear insights into how different linguistic dimensions contribute to Mis-Information detection decisions.

RESULTS AND DISCUSSION

The proposed MisInfoCheckXAI**,** framework based on Logistic Regression model attained an accuracy of 0.8562 with macro precision, recall, and F1-score of 0.8472**,** 0.8664**,** and 0.8563**,** demonstrating effective Mis-Information classification using linguistic features including Disclosure, Pragmatic, and Integration. Confusion matrix investigation shows solid performance in identifying Mis-information

and real content, while some Mis-Information instances were misclassified due to dataset imbalance and dominance of semantic patterns.

Explainable AI using grouped SHAP analysis revealed that Integration features (TF-IDF) contributed the most to prediction decisions, followed by Pragmatic and Disclosure features. These findings confirm that combining DIP feature groups improves transparency, computational efficiency, and interpretability while maintaining strong predictive performance for real-world fake content detection tasks.

CONCLUSION AND FUTURE WORK

The proposed MisInfoCheckXAI**,** framework integrates linguistic features such as Disclosure, Pragmatic, and Integration (DIP) with Logistic Regression and SHAP-based explainable AI for apparent fake content detection. The model attained strong performance while preserving interpretability and computational efficiency. Grouped SHAP analysis showed that Integration features such as TF-IDF contributed the highest influence on prediction of Mis-Information which supported by Pragmatic and Disclosure cues features. Our MisInfoCheckXAI**,** framework compare to other traditional model provides a practical alternative to complex NLP, machine learning and deep learning models for real-world deployment. Future work will focus on improving Mis-Information recall, integrating transformer-based models, and extending explainability for large-scale real-time applications.

REFERENCES

1. Baptista, J. P., & Gradim, A. (2022). *A working definition of fake news*. Encyclopedia.
2. Bashaddadh, Omar, et al. "Machine learning and deep learning approaches for fake news detection: A systematic review of techniques, challenges, and advancements." IEEE Access (2025).Molina, M. D., et al. (2021). *"Fake news" is not simply false information*. American Behavioral Scientist.
3. Prachi, N. N., et al. (2022). *Detection of fake news using machine learning and NLP algorithms*. Journal of Advances in Information Technology.

4. Zhang, Li. "Features extraction based on Naive Bayes algorithm and TF-IDF for news classification." PLoS One 20.7 (2025): e0327347.
5. Choudhary, A., & Arora, A. (2021). *Linguistic feature-based learning model for fake news detection*. Expert Systems with Applications.
6. Kumbhkar, Makhan, Shraddha Masih, and Savita Kolhe. "LSSS-System: Fake Content Detection Using Linguistic Features." *2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG)*. IEEE, 2024.
7. Kumbhkar, Makhan, Shraddha Masih, and Seema Kamble. "An Expert System for Detecting Fake Content Using Machine Learning and Deep Learning Model Through Existing Dataset." *International Conference on Data Science and Big Data Analysis*. Singapore: Springer Nature Singapore, 2024.
8. Rai, N., et al. (2022). *Fake news classification using transformer-based enhanced LSTM and BERT*. IJCCE.
9. Athira, A. B., et al. (2023). *Explainable AI applied to fake news detection: A survey*. Engineering Applications of AI.
10. Gongane, V. U., et al. (2024). *Explainable AI techniques for fake news detection*. Journal of Computational Social Science.