Truthlens: Hybrid Context Aware And Evidence-Based Fact Checking System For Online News

[1]Dr. G. Madhavi, [2]Sikhakolli Pradyumna, [3]Murapaka D V S S S Vinay

[1]Associate Professor and HOD, [2]CSE Student, [3]CSE Student
[1]Computer Science and Engineering Department,
[1]UCEN JNTUK, Narasaraopet, India
[1]madhavi.researchinfo@gmail.com, [2]pradyusikhakolli12345@gmail.com,
[3]murapakavinay@gmail.com

***Abstract*—This paper presents TruthLens, a hybrid context aware and evidence-based fact checking system designed to detect misinformation in online news and social media content. With the rapid growth of digital platforms, misinformation spreads quickly, creating challenges for individuals and society [11][19]. Traditional detection systems rely mainly on static datasets and fail to adapt to real-time scenarios [3][13]. The proposed system integrates transformer-based deep learning models such as BERT for contextual understanding [5] and RoBERTa-based Natural Language Inference (NLI) for reasoning over retrieved evidence [6][10]. Unlike conventional approaches, TruthLens incorporates source-aware analysis and real-time evidence retrieval to improve prediction reliability [8]. The system supports multiple input formats including text, URLs, and images using OCR techniques, and also includes multilingual capabilities through translation mechanisms [9]. It generates explainable outputs with confidence scores, enhancing transparency and user trust. Experimental results show that transformer-based models outperform traditional methods, achieving high accuracy and improved real-world applicability [1][7].**

***Index Terms*—Misinformation Detection, BERT, RoBERTa, Natural Language Inference, Deep Learning, Fake News Detection**

I. INTRODUCTION

The rise of digital media and online platforms has changed the way information is produced and used. While these platforms have improved accessibility and speed of communication, they have also enabled the widespread dissemination of misleading and false information [11]. This phenomenon has become a major concern, as misinformation can influence public opinion, distort factual understanding, and in some cases, lead to serious societal consequences [19].

Traditional misinformation detection techniques largely rely on supervised learning models trained on pre-labeled datasets [3][14]. Although these models achieve good performance in controlled environments, they often struggle when applied to real-world scenarios where new and evolving claims continuously emerge [13]. Moreover, misinformation is no longer limited to plain text; it appears in various formats such as social media posts, news articles, and images containing embedded text, making detection more complex [12].

To address these type of challenges, this work proposes a hybrid and context aware misinformation detection system that integrates deep learning with evidence-based verification. Instead of relying solely on classification results, the system actively gathers supporting information from external sources and evaluates claims through reasoning [8]. It employs transformer-based models such as BERT for understanding contextual patterns in text [5], along with a RoBERTa-based Natural Language Inference (NLI) model to validate evidence [6].

A key strength of the proposed system is to process multiple input types. It can analyze textual content, extract information from URLs, and interpret text embedded within images using Optical Character Recognition (OCR) [9]. Additionally, the system incorporates multilingual capabilities through translation mechanisms, allowing it to handle non-English inputs effectively [9].

By combining real-time data retrieval with advanced deep learning techniques and reasoning mechanisms, the proposed approach improves both accuracy and interpretability [1]. The system also provides confidence scores and explanations for its predictions, enhancing transparency and user trust. Overall, this work aims to contribute toward building a practical solution for misinformation detection in modern digital environments.

II. LITERATURE REVIEW

Misinformation detection has gained significant attention due to the increasing influence of digital platforms [11]. Early approaches depend on traditional machine learning algorithms such as Logistic Regression, Random Forests, and Support Vector Machines, which used handcrafted features for classification [14].

With advancements in deep learning, models like CNN, LSTM, and GRU improved performance by capturing contextual and sequential patterns in text [2][4]. More recently, transformer-based models such as BERT have achieved superior results due to their ability to generate context-aware representations [5].

However, most existing models rely on static datasets and lack adaptability to real-time scenarios [13]. To overcome this limitation, evidence-based verification approaches have been introduced, where claims are validated using external sources [8].

Natural Language Inference (NLI) models, particularly RoBERTa, are widely used to determine whether evidence supports or contradicts a claim [6][10]. Additionally, OCR-based techniques have been used to extract text from images, enabling multimodal analysis [9].

Despite these advancements, there is still a need for systems that combine deep learning, real-time evidence retrieval, and multimodal processing in a unified framework. The proposed system addresses these gaps effectively.

III. PROPOSED METHODOLOGY

A. System Overview

The system was developed using a Python-based environment with libraries such as PyTorch, HuggingFace Transformers, and Scikit-learn. Additional tools were used for web data extraction and OCR processing. The BERT model was fine-tuned on the ISOT dataset for classification [5], while a pre-trained RoBERTa-based NLI model was used for reasoning [6]. The implementation was tested in both local and cloud environments (e.g., Google Colab) to ensure efficient execution. Tokenization and batch processing were applied to maintain computational efficiency [14].

B. Dataset Description

The ISOT Fake News Dataset was utilized as the primary dataset for training and evaluation [24]. It comprises labeled news articles categorized as real or fake, facilitating the learning of meaningful textual patterns by the model. Its well-structured format and public availability make it an appropriate and reliable resource for research purposes.

TABLE I: Dataset Statistics

Attribute	Value
Dataset Name	ISOT Fake News Dataset
Total Samples	~44,000
Real News Samples	~21,000
Fake News Samples	~23,000
Data Type	Text (News Articles)
Language	English
Train/Test Split	80% / 20%

C. Data Preprocessing and Tokenization

The input data undergoes preprocessing steps to remove noise such as special characters, HTML tags, and unnecessary spaces [14]. The cleaned text is then normalized to maintain consistency. For model training, the processed text is converted into tokens using the BERT tokenizer [5]. This process generates input IDs and attention masks required by the model. Padding and truncation are applied to ensure uniform input length, and labels are encoded into binary form for classification.

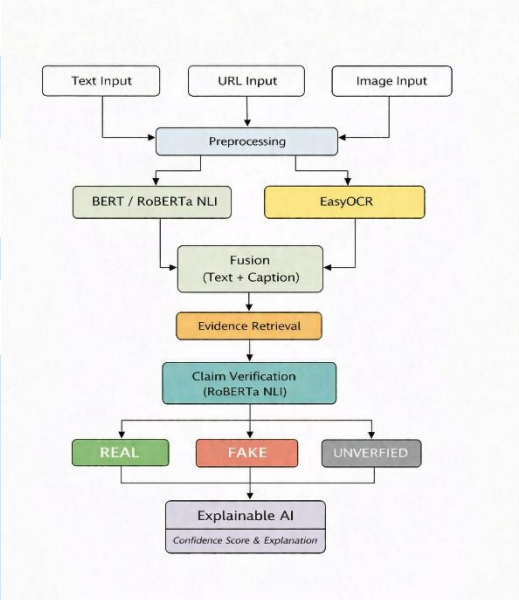

Fig. 1: Proposed architecture for TruthLens

D. Input Handling and Claim Extraction

The system accepts text, URLs, and images as input. Text is processed directly, while content from URLs is extracted using web scraping techniques. A claim extraction step identifies the most relevant statement from the input. For images, OCR is used to extract embedded text, which is then processed similarly to textual input [9].

E. Multilingual Processing

To handle non-English inputs, the system includes a language detection step followed by translation into English [9].

F. Evidence Retrieval

The system retrieves relevant information from the web to verify claims. A query is generated and search APIs are used to collect evidence [8].

G. NLI-Based Reasoning

A pre-trained RoBERTa NLI model classifies each claim–evidence pair into:

- Entailment (supports the claim)
- Contradiction (opposes the claim)
- Neutral (insufficient information) [6][10]

H. Decision Mechanism

Final classification is based on threshold values:

- High contradiction → FAKE
- High entailment → REAL
- Otherwise → UNVERIFIED

IV. Experimental Results and Discussion

The performance of the proposed system was evaluated using multiple machine learning and deep learning models trained on the ISOT dataset [24].

A. Evaluation Metrics

A confusion matrix consists of:

- TP (True Positive) → Correctly predicted positive
- TN (True Negative) → Correctly predicted negative
- FP (False Positive) → Incorrectly predicted positive
- FN (False Negative) → Incorrectly predicted negative

Assume:

Word Embeddings	Model	Accuracy	Precision	Recall	F1-Score
without word embeddings	Logistic Regression	0.98	0.98	0.97	0.98
without word embeddings	CNN	0.99	0.99	0.99	0.99
without word embeddings	LSTM	0.98	0.98	0.98	0.98
without word embeddings	GRU	0.98	0.98	0.99	0.98
with automatic word embeddings	BERT	0.999	0.999	0.999	0.999

- xi = actual label
- xˆi = predicted label
- Binary classification:
  - 1 → Positive (Fake)
  - 0 → Negative (Real)

True Positive (TP) :

$$TP = \sum_{i=1}^{M} 1(x_i = 1 \wedge \hat{x}_i = 1) \quad (1)$$

True Negative (TN):

$$TN = \sum_{i=1}^{M} 1(x_i = 0 \wedge \hat{x}_i = 0) \quad (2)$$

False Positive (FP):

$$FP = \sum_{i=1}^{M} 1(x_i = 0 \wedge \hat{x}_i = 1) \quad (3)$$

False Negative (FN):

$$FN = \sum_{i=1}^{M} 1(x_i = 1 \wedge \hat{x}_i = 0) \quad (4)$$

Accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision:

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall:

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

F1-Score:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \text{ [14]} \quad (8)$$

B. Confusion Matrix Analysis

The confusion matrices show that BERT achieves the highest accuracy with minimal misclassification [5][7].

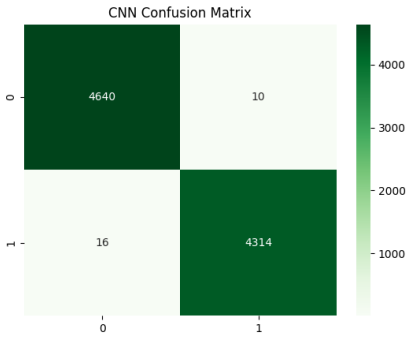

Fig. 2: CNN Confusion Matrix

TABLE II: Model Performance Comparison

C. Model Performance Comparison

Deep learning models outperform traditional approaches, with BERT achieving the best performance [5][4].

D. Quantitative Results

The results show that BERT performs better than other models, highlighting how important

context understanding is in detecting misinformation [5].

E. Discussion

The superior performance of transformer-based models highlights the importance of contextual representation in misinformation detection tasks [1][5].

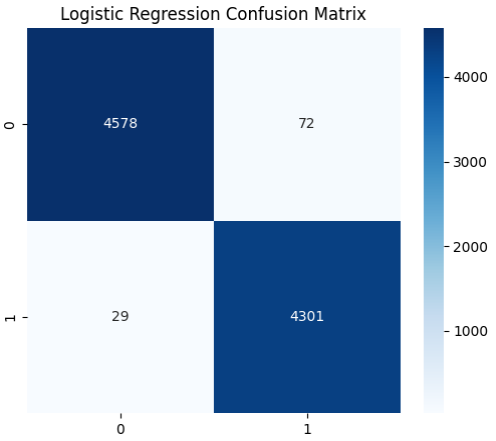

Fig. 3: Logistic Regression Confusion Matrix

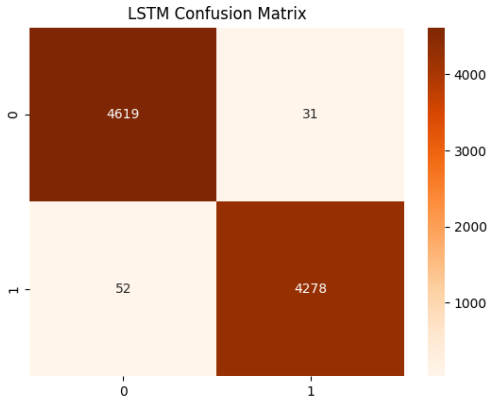

Fig. 4: LSTM Confusion Matrix

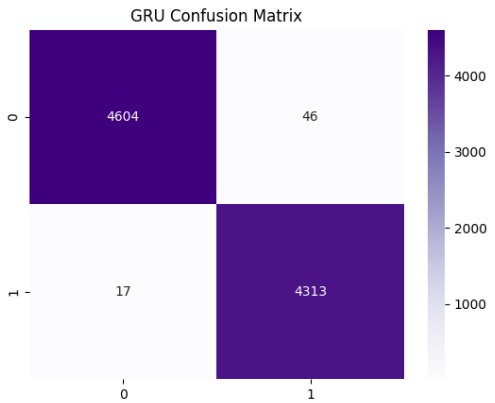

Fig. 5: GRU Confusion Matrix

F. Real-World Applicability

The system supports:

- Handling unseen claims
- Explainable predictions
- Real-time verification [8]

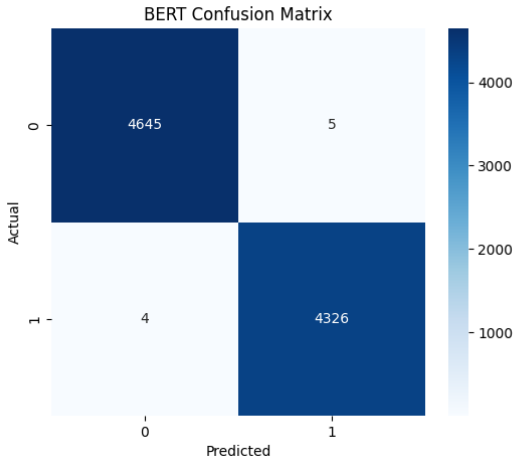

Fig. 6: BERT Confusion Matrix

V. Conclusion and Future Work

This work presents a multimodal misinformation detection system that combines deep learning with evidence-based verification to assess the reliability of online information. The proposed approach integrates BERT for text understanding [5] and a RoBERTa-based NLI model for reasoning [6], enabling the system to move beyond traditional classification and incorporate real-time validation using external evidence [8].

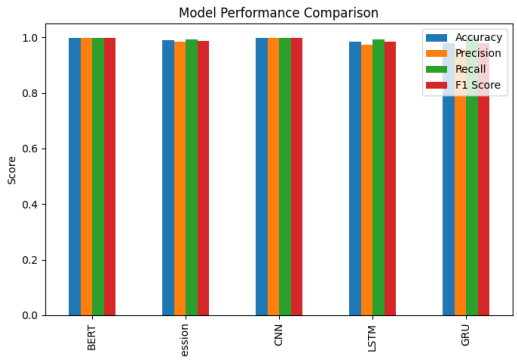

Fig. 7: Model Performance Comparison

The system was evaluated using multiple models, including Logistic Regression, CNN, LSTM, GRU, and BERT, on the ISOT dataset [24]. The results show that transformer-based models, particularly BERT, achieve the best performance across all evaluation metrics [5][7]. More importantly, the inclusion of realtime evidence retrieval and reasoning allows the system to handle new and unseen claims while providing explainable outcomes, which improves both reliability and user trust.

Another key aspect of this work is the design of a flexible pipeline capable of handling different input formats such as text, URLs, and images. The use of OCR for extracting text from images and translation for non-English inputs

extends the system's usability in practical scenarios [9]. Additionally, the generation of confidence scores and explanations makes the system more transparent and easier to interpret.

However, certain limitations still exist. The overall performance depends on the relevance and quality of the retrieved evidence, which may vary across different queries. In some cases, incomplete or neutral evidence can affect the final decision. Moreover, the use of translation for multilingual inputs may introduce minor inconsistencies, and evaluating reasoning performance in dynamic real-world conditions remains a challenge [9].

Future improvements can focus on enhancing the reasoning component by fine-tuning models like RoBERTa on domain-specific fact-checking datasets [6][8]. Improving evidence retrieval through more advanced search techniques or structured knowledge sources can further increase accuracy. The system can also be extended to include deeper analysis of visual content beyond text extraction.

In addition, deploying the system as a real-time application, such as a browser extension or monitoring tool, can improve its practical impact. Incorporating continuous learning and larger datasets will help the system adapt to evolving misinformation patterns. With these enhancements, the proposed framework has the potential to serve as a reliable solution for detecting misleading information and supporting trustworthy information sharing.

REFERENCES

[1] V.-I. Ilie et al., "Context-Aware Misinformation Detection: Benchmark of Deep Learning Architectures Using Word Embeddings," IEEE Access, vol. 9, pp. 162144–162160, 2021.

[2] O. Ajao, D. Bhowmik, and S. Zargari, "Fake news identification on Twitter with hybrid CNN and RNN models," in Proc. 9th Int. Conf. Social Media Soc., Jul. 2018, pp. 226–230.

[3] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," in Proc. ASIST Annual Meeting, 2015.

[4] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "FNDNet—A deep convolutional neural network for fake news detection," Cognitive Systems Research, vol. 61, pp. 32–44, 2020.

[5] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019.

[6] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[7] S. Kula, M. Choras, and R. Kozik, "Application of the BERT-based ´ architecture in fake news detection," in Proc. CISIS, 2020, pp. 239–249.

[8] J. Thorne et al., "FEVER: A large-scale dataset for fact extraction and verification," in Proc. NAACL-HLT, 2018.

[9] H. Schuster, M. Gupta, R. Shah, and M. Lewis, "Cross-lingual fact checking," in Proc. EMNLP, 2019.

[10] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," in Proc. EMNLP, 2019.

[11] A. Gelfert, "Fake news: A definition," Informal Logic, vol. 38, no. 1, pp. 84–117, 2018.

[12] B. Ghanem, P. Rosso, and F. Rangel, "An emotional analysis of false information in social media and news articles," ACM Transactions on Internet Technology, vol. 20, no. 2, 2020.

[13] G. Gravanis et al., "Behind the cues: A benchmarking study for fake news detection," Expert Systems with Applications, vol. 128, pp. 201–213, 2019.

[14] J. Hartmann et al., "Comparing automated text classification methods," International Journal of Research in Marketing, vol. 36, no. 1, pp. 20–38, 2019.

[15] S. Helmstetter and H. Paulheim, "Weakly supervised learning for fake news detection on Twitter," in Proc. IEEE/ACM ASONAM, 2018.

[16] A. Choudhary and A. Arora, "Linguistic feature-based learning model for fake news detection," Expert Systems with Applications, vol. 169, 2021.

[17] V. Feldman, R. Frostig, and M. Hardt, "The advantages of multiple classes for reducing overfitting," in Proc. ICML, 2019.

[18] A. E. A. Gautam, "Fake news detection using XLNet with topic distributions," in Proc. AAAI, 2021.

[19] J. Hua and R. Shaw, "Corona virus (COVID-19) 'infodemic' and emerging issues through a data lens," International Journal of Environmental Research and Public Health, vol. 17, no. 7, 2020.

[20] P. Bojanowski et al., "Enriching word vectors with subword information," Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017.

[21] M. Hardalov, I. Koychev, and P. Nakov, "In search of credible news," in Proc. AI: Methodology, Systems, Applications, 2016.

[22] T. Mikolov et al., "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[23] K. Higgins, "Post-truth: A guide for the perplexed," Nature, vol. 540, 2016.

[24] ISOT Fake News Dataset – University of Victoria, Canada