Disinformation Detection: Integrating Predictive and Generative AI Model

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

Yiheng Yuan [1] Luran Zhang [1] Jade Zhou [1] Dr. Ali Arsanjani [2]

[1]UC San Diego [2]Google

Introduction

Misinformation and disinformation are major challenges in the digital age, eroding public trust and distorting decision-making.

- **Misinformation**: False information spread unintentionally.
- **Disinformation**: Deliberately misleading content designed to deceive.

To address this challenge, we have developed a combined predictive and generative AI system, which is designed to detect and curb the spread of misleading content. By leveraging the capabilities of generative AI, the model can detect and mitigate the spread of false or misleading content.

Scan to visit our project website

Dataset and Preprocessing

- **LIAR PLUS dataset - Politics**
  We utilize the LIAR-PLUS dataset for predictive modeling. Web scraping and vector embedding techniques store textual data for AI-based analysis.

ID	Label	Statement	Subjects	Speaker	Job Title	State	Party	Context	Justification
2635.json	False	When did the decline of coal start?	Energy, History	Dwayne Bohac	State Representative	Texas	Republican	A floor speech	Surovell said the decline of coal "started when ...
10540.json	Half-True	Hillary Clinton agrees with McCain on Iraq	Foreign Policy	Barack Obama	President	Illinois	Democrat	Denver	Obama said he would have voted against the ame...
324.json	Mostly-True	Health care reform makes Medicare worse	Health Care	Blog Posting	-	-	None	A news release	The release may have a point that Mikulskis co...
1123.json	False	Economic turnaround started in 2009	Economy, Jobs	Charlie Crist	-	Florida	Democrat	CNN Interview	Crist said that the economic "turnaround start...
9028.json	Half-True	Chicago Bears have had most QBs since 2000	Education	Robin Vos	Assembly Speaker	Wisconsin	Republican	Online Opinion Piece	But Vos specifically used the word "hired," wh...

Table 1. Sample Dataset from LIAR-PLUS

- **Politifact** Fact-checking website <https://www.politifact.com/>

Predictive Model Approach

The analysis focuses on developing models to assess the veracity of statements based on various factuality factors. The approach employs **Random Forest Classifier** to classify text by factual accuracy, incorporating natural language processing (NLP) methods to quantify elements that suggest factual reliability.

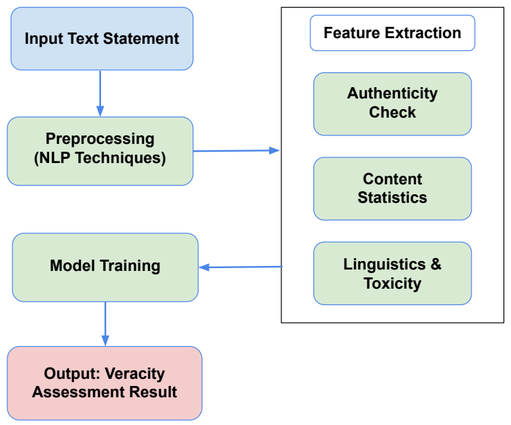

Figure 1. Predictive Model Flow Chart

- **Authenticity**: Cross-references *Liar Plus* statements with *PolitiFact* claims using TF-IDF vectorization and cosine similarity (≥ 0.8) to flag high-similarity statements.
- **Content Statistics**: Analyzes syntactic and semantic structure via POS, NER, and relationship extraction. Constructs content graphs to assess semantic similarity with known sources.
- **Linguistics & Toxicity**: Extracts linguistic features (punctuation, adjectives, modals, complex words) and aggregates content metrics (token/entity counts, sentiment, readability) to detect references to studies or data.

Generative AI Model Approach

- Google Gemini Model for nuanced language analysis
- Evaluating Factuality Factors - Biases Factuality Factor, Context Veracity Factor, and Information Utility Factor.
- Fractal Chain-of-Thought (FCoT) prompting to improve consistency
- Function calling for contextual verification
- Search engine scrapping to enhance the generative AI model's ability to identify misinformation (SerpAPI)

Hybrid Model Implementation

Our framework combines both predictive and generative models:

1. Predictive Model provides structured analysis
2. Generative AI refines contextual understanding
3. Final truthfulness score merges both outputs

.

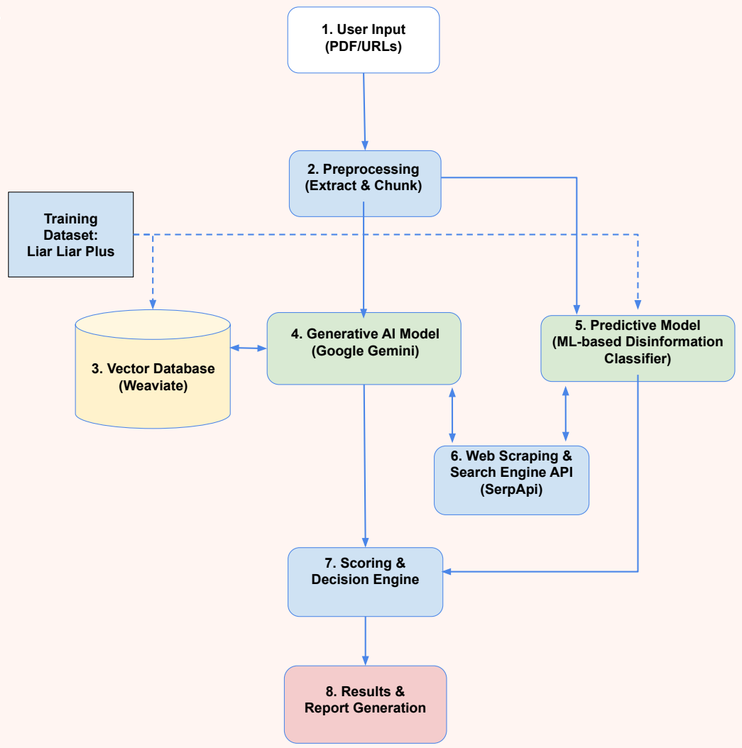

Figure 2. Flow Chart

Results & Evaluation

- Predictive model achieves high accuracy using structured data

Factuality Factor	Dataset used for training	Accuracy
Authenticity	Politifact and Liar Plus Dataset	0.39
Linguistic based	Liar Plus Dataset	0.21
Content Statistic	Liar Plus Dataset	0.21

Table 2. Pred Model Factuality Factor Accuracy

- Comparison: Normal CoT vs FCoT

Factuality Factors	Normal CoT (1 to 6)	FCoT (1 to 6)	Comparison
Biases	4	4.5	FCoT is more critical of missing counterarguments and detects subtle biases better.
Context Veracity	4	3.75	FCoT penalizes unverified claims more heavily and is stricter on context shifts.
Information Utility	4.25	4.5	FCoT values completeness, considering historical context like past recalls to improve article usefulness.
Average	4.08	4.25	The article selected is factual, and FCoT provided better precision in evaluating it.

Table 3. FCoT vs Normal CoT Experiment Results

- Graphical Representation of CoT vs FCoT

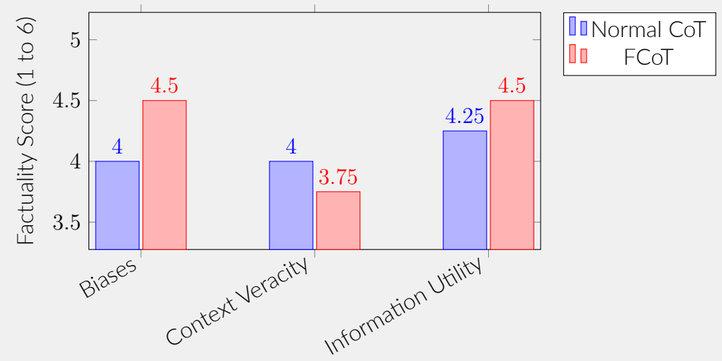

Figure 3. Comparison of Normal CoT vs FCoT in Factuality Scores

- Model Vs. Actual PolitiFact Rankings Mean Absolute Error is 1.00

Future Directions

To enhance our misinformation detection system, we aim to:

- **Expand Factuality Factors:** Increase from 6 to 12 for a more detailed veracity assessment.
- **Improve Data Retrieval:** Enhance WeaviateDB ranking for more accurate and relevant content matching.
- **Integrate Real-Time Sources:** Incorporate live fact-checking databases and authoritative news APIs.
- **Optimize FCoT Prompting:** Refine Fractal Chain-of-Thought (FCoT) to improve model consistency and reliability.
- **Scale for Deployment:** Optimize performance for integration into social media platforms and fact-checking systems.

These improvements will strengthen our system's accuracy, usability, and scalability.