MAVS: An Ensemble-Based Multi-Agent Framework for Fake News Detection

Dhruv Tyagi[1], Anurag Singh[1], and Hocine Cherifi[2]

[1] National Institute of Technology Delhi, India
211220018@nitdelhi.ac.in, anuragsg@nitdelhi.ac.in
[2] University de Bourgogne Europe, France
hocine.cherifi@u-bourgogne.fr

Abstract. The Multi-Agent Verification System (MAVS)Framework aims to improve fake news detection by leveraging a multi-agent system that enhances decision-making through multidimensional evaluation, mitigating adversarial attack vulnerabilities.
MAVS utilizes four specialized agents (a GNN model and Generative AI models for fact-checking, stance-checking, and sentiment analysis) each operating independently and in parallel. The final classification is determined through a weighted aggregation of the agents' outputs, optimized using Stochastic Gradient Descent (SGD)-based Logistic Regression to ensure optimal weight distribution.
MAVS achieves an accuracy of 97.6% and an F1 score of 98%. Under a Multi-Agent Reinforcement Learning (MARL) attack, the system's accuracy drops to 74.19% and the F1 score to 71%, while maintaining a precision of 100%. This highlights the framework's resilience and ability to maintain high precision despite adversarial conditions.
The proposed framework strengthens fake news detection by combining multiple verification strategies, reducing susceptibility to adversarial attacks. Future work includes refining agent interactions and exploring real-time deployment for broader applicability.

Keywords: Multi-Agent Systems · Fake News Detection · Graph Neural Networks · Fact Checking · Sentiment Analysis· Stance Analysis

1 Introduction

The spread of misinformation undermines public trust, distorts democratic processes, and poses risks to public health and safety [13]. Fake news detection is essential to protect public discourse, ensuring informed debates and strengthening public dialogue. It is also crucial for safeguarding democracy, as misinformation can manipulate public opinion and influence elections [13]. Events like the COVID-19 pandemic have demonstrated the real-world harm of unchecked false information, necessitating robust detection mechanisms [2, 14].

Traditional fact-checking methods are time-consuming and inefficient, prompting researchers to explore AI and machine learning-based solutions [16, 3, 15]. While NLP-based sentiment and stance analysis help analyze textual content,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

they fail to capture misinformation propagation patterns, making them susceptible to manipulation [2]. Deep learning approaches improved detection accuracy [11, 12] but they require large labeled datasets and struggle with propagation analysis. This led to the adoption of Graph Neural Networks (GNNs), which model relationships and dependencies within news propagation networks [12]. By analyzing dissemination patterns, user interactions, and source credibility, GNNs provide a more comprehensive detection framework. However, they remain vulnerable to adversarial attacks, where fraudsters manipulate graph structures to spread misinformation [21].

To enhance robustness, we propose MAVS (Multi-Agent Verification System), an ensemble-based framework that integrates GNNs with generative AI models for fake news detection. MAVS combines multiple AI agents specializing in fact-checking, stance detection, and sentiment analysis to analyze both news content and its propagation context. The fact-checking agent retrieves external evidence using generative AI, while the stance and sentiment analysis agents assess contextual consistency and emotional bias [9]. The final classification is determined through a weighted fusion of agent outputs, ensuring a multidimensional evaluation [12].

To validate our approach, we conduct extensive experiments on the Politifact (UPFD) dataset [7], evaluating accuracy, precision, recall, and F1-score, where the F1-score is the harmonic mean of precision and recall. It provides a balanced measure and helps to evaluate the model's robustness in identifying both fake and real news accurately. We also analyze model interpretability by examining agent contributions and weight distributions in decision-making.

Contributions of this research:

1. Evaluating adversarial vulnerabilities of the GNN.
2. Developing a multi-agent verification system integrating fact-checking, stance detection, and sentiment analysis.
3. Optimizing framework weights using machine learning.

By combining graph-based learning with AI-driven verification, MAVS offers a novel and effective approach to combating misinformation in the digital era.

2 Related Work

The rise of misinformation and advancements in AI have driven fake news detection techniques. Early approaches relied on rule-based systems and manual fact-checking, which were labor-intensive and lacked scalability. Machine learning models improved efficiency by classifying news based on writing style, sentiment, and credibility, but struggled with contextual understanding and adversarial attacks. Recent advancements incorporate NLP techniques, deep learning models, and GNNs to analyze both content and propagation patterns.

2.1 Fact-Checking Methods

Fact-checking plays a crucial role in fake news detection by verifying claims using external knowledge bases. These systems provide explainability but face

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

challenges in dynamic scenarios due to their reliance on up-to-date knowledge [20].

2.2 NLP-Based Methods

Early detection methods used lexical features like BoW and TF-IDF to analyze textual content [2]. However, these models lacked contextual understanding and were vulnerable to adversarial manipulation [23].

2.3 Deep Learning for Fake News Detection

Deep learning models such as RNNs and CNNs improved contextual comprehension by capturing sequential dependencies [1, 8]. However, they required extensive labeled datasets, lacked interpretability and often overlooked social interactions and propagation dynamics, making them susceptible to misinformation tactics [5].

2.4 Graph Neural Networks (GNNs)

GNNs effectively model social interactions and misinformation spread by analyzing propagation patterns [19]. However, GNNs face challenges such as high computational complexity, scalability issues, and adversarial vulnerabilities [22, 4, 6].They remain vulnerable to adversarial attacks where malicious actors manipulate the graph structure to evade detection [21]. Real-world attackers add new nodes and edges to alter misinformation spread patterns just like in MARL attack shown in **Fig. 1**, making detection less effective.

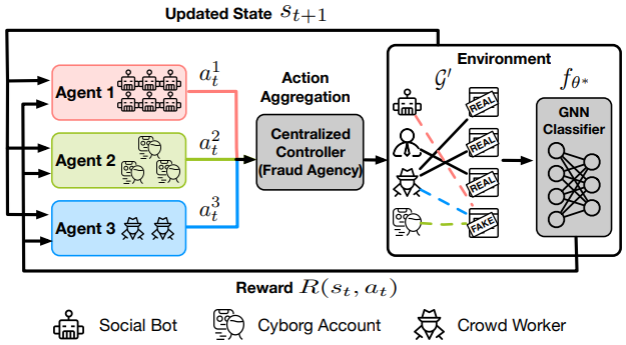

Fig. 1: MARL adversarial attack against GNN-based fake news detection, where manipulated edges and nodes degrade detection accuracy [21].

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

2.5 MAVS vs. Other Approaches

Table 1 presents a feature-wise comparison of detection methods. MAVS integrates GNNs with Fact, Stance, and Sentiment Checkers, offering a more robust and adaptable approach.

Table 1: Feature comparison across fake news detection models (✓: Yes, ✗: No, △ : *Partial*).

Feature	GNN	Fact Checker	Stance Checker	Sentiment Checker	MAVS
Propagation Analysis	✓	✗	✗	✗	✓
Textual Analysis	✗	✓	✓	✓	✓
Credibility Checking	✓	✓	✗	✗	✓
Context Understanding	✗	✓	✓	✗	✓
Stance Detection	✗	✗	✓	✗	✓
Sentiment Analysis	✗	✗	✗	✓	✓
Fact Verification	✗	✓	✗	✗	✓
Adversarial Resilience	✗	✗	✗	✗	△

Key Takeaways:

- **Propagation Analysis:** GNNs excel at misinformation spread modeling, which traditional methods lack.
- **Context Understanding:** MAVS integrates multiple sources for more comprehensive evaluations.
- **Resilience to Adversarial Attacks:** MAVS reduces attack vulnerability by combining independent agents.

MAVS combines the best aspects of fact-checking, NLP, deep learning, and GNNs to create a robust, multi-perspective fake news detection system.

3 Methodology

The section discusses the methodology of the research in detail.

3.1 Proposed Framework

To effectively detect fake news, the proposed **MAVS (Multi-Agent Verification System)** framework leverages multiple AI agents for analyzing both content and propagation patterns in parallel and independently. The methodology is designed to enhance the reliability of fake news detection by integrating different perspectives, including **network propagation, factual consistency, stance alignment, and sentiment analysis**.This section outlines the dataset used, key components of the framework their roles, and how they contribute to the final classification decision.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

3.2 Dataset

The dataset used in this study is the well-known **UPFD-Politifact** dataset [7], specifically curated for evaluating binary graph classification, graph anomaly detection, and fake news detection tasks. It is structured as a `PyTorch-Geometric` dataset object having field as shown in **Table 2**, enabling seamless integration with various GNN models.

The dataset consists of tree-structured graphs representing news propagation networks on Twitter. These graphs are constructed using fact-check information from *Politifact* and *Gossipcop*, originally extracted by FakeNewsNet[17]. The structure of each graph is as follows:

- **Root Node:** Represents the original news article.
- **Leaf Nodes:** Twitter users who retweeted the article.
- **Edges:** A directed edge exists from a user to the news node if they retweeted it. Two users are connected if one retweeted the news from the other.

The dataset includes three types of node features (used in our case):

- **SpaCy Features (300-dimensional):** Word2Vec embeddings generated using the `spaCy` library.
- **Profile Features (10-dimensional):** Extracted from Twitter user profiles, capturing metadata like follower count and verification status.
- **Content Features (310-dimensional):** Combines a 300-dimensional user comment word2vec embedding with the 10-dimensional profile feature.

Table 2: Dataset Columns and Descriptions

Column Name	Description	Example Value
id	Unique identifier	"politifact4190"
news_url	Source URL	`http://www.c.gov/doc.pdf`
title	Headline	"Budget and Economic Outlook"
tweet_ids	Related tweet IDs	"1102113056 1102113348 ..."
label	Real (0) or Fake (1)	1

3.3 Key Components

As shown in the architecture of proposed MAVS in **Fig. 2** the key components are Graphical Neural Network (GNN), Fact-Checker, Stance-Checker and Sentiment-Checker. The detailed description of these components are as follows,

1. **Graphical Neural Network (GNN):-** The GNN Model operates on hierarchical tree-structured graphs as shown in **Fig. 3**
   The GNN captures the structure and influence patterns within this retweet graph, helping identify propagation trends [21]. The model leverages **Graph**

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

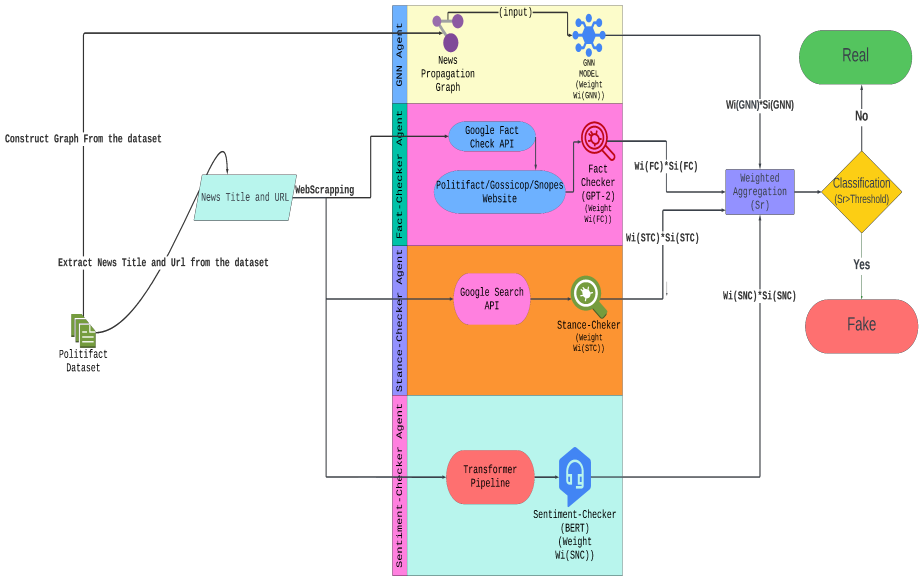

Fig. 2: The architecture of MAVS framework, integrating GNN and AI agents (Fact-Checker, Stance-Checker, Sentiment-Checker) for fake news detection.

Attention Networks (GAT) for learning propagation patterns. GNN consists of three stacked GATConv layers followed by a global max pooling operation, which aggregates graph-level embeddings. The final output is obtained through a fully connected readout layer with a sigmoid activation for binary classification.

2. **Fact-Checker**: The fact-checker agent uses **Algorithm 1** which ensures that the information aligns with verified facts from trusted sources, helping to distinguish misinformation from truthful content[20]. The fact-checking process involves analyzing claims and assigning a weighted score based on their verdicts. Given a claim's verdict V_i, The final weighted score S_{weighted} is computed as the average of all S_i values. If S_{weighted} is negative, the statement is considered more likely to be true; otherwise, it is likely false.
3. **Stance-Checker**: The stance-checker agent determines how well your content is aligned from different perspectives using a variety of predefined stance detection algorithms. Whether the news concurs or would seem to refute identified affiliations and views, more easily detecting bias. Basically, using **Algorithm 2** it evaluates the stance of the article relative to a claim or premise, classifying it as *supports*, *contradicts*, or *neutral*[9].

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

Algorithm 1 Fact-Checking Model Using GPT-2 and API

Input : S: Statement to verify, API_Key: API key for fact-checking.
Output: S_{weighted}: Final weighted score, Generated_Text: Explanation from GPT-2.
Step 1
Initialize tokenizer and text generator: Tokenizer $\leftarrow$ GPT2Tokenizer('gpt2'), Generator $\leftarrow$ pipeline('text-generation', model = 'gpt2')
Step 2
Fetch results: $F \leftarrow$ API_Request(S, API_Key) **if** *status_code* $\neq 200$ **then**
  return Error
Extract claims: $C \leftarrow F$['claims']
Step 3
for *each claim* $C_i \in C$ **do**
  Retrieve verdict V_i and compute score: $S_i = \begin{cases} -1, & V_i \in \{\text{"true"}, \text{"mostly true"}, \text{"half true"}\} \\ 1, & V_i \in \{\text{"false"}, \text{"mostly false"}, \text{"pants on fire"}\} \\ 0, & \text{otherwise} \end{cases}$ Accumulate: $S \leftarrow S + S_i$
Step 4
$S_{\text{weighted}} \leftarrow \frac{S}{|C|}$
Step 5
Construct prompt: Prompt $\leftarrow$ "Given the statement 'S' and the fact-check results:" Generate explanation: Generated_Text $\leftarrow$ Generator(Prompt)
return S_{weighted}, Generated_Text

4. **Sentiment-Checker**: The sentiment-checker agent leverages generative AI models to assess the emotional tone of the content retweeted by users. Using **Algorithm 3** it categorizes the sentiment as positive, negative, or neutral, helping in identifying emotional manipulation or polarizing content.[23]. Each agent generates a score, which is aggregated to classify the news as real or fake.

Algorithm 2 Stance-Checking Model Using BART

Input : T: News title, URL: Article URL.
Output: L_{final}: Final stance label, S_{final}: Stance score.
Step 1
Initialize stance detection model: Model $\leftarrow$ pipeline("zero-shot-classification", model = "bart")
Step 2
Extract article content and summary: $C \leftarrow$ ExtractText(URL), $S \leftarrow C[:300]$
Step 3
Compute stance of title w.r.t. content: $S_T \leftarrow w_i \cdot p(L_i \mid T, S)$
Step 4
Perform Google search for related URLs.
foreach *related URL* i **do**
  Extract content H_i (first 200 words), compute stance score: $S_{R_i} \leftarrow w_i \cdot p(L_i \mid T, H_i)$ Append to stance list.
Compute average stance score: $S_R = \frac{1}{n} \sum_{i=1}^{n} S_{R_i}$
Step 5
Compute weighted final stance score: $S_{\text{final}} = 0.3 \cdot S_T + 0.7 \cdot S_R$
Step 6
if $0.7 \cdot S_C > 0.3 \cdot S_T$ **then**
  $L_{\text{final}} \leftarrow L_C$
else
  $L_{\text{final}} \leftarrow L_T$
Step 7
if L_{final} *is "supports"* **then**
  $S_{\text{adjusted}} \leftarrow -S_{\text{final}}$
else if L_{final} *is "neutral"* **then**
  $S_{\text{adjusted}} \leftarrow 0$
else
  $S_{\text{adjusted}} \leftarrow S_{\text{final}}$
return $L_{\text{final}}, S_{\text{final}}$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

Algorithm 3 Sentiment Score Computation for News Articles

Input : T: News title, URL: Article URL.
Output: S_{final}: Sentiment score, Sentiment Label: Positive, Neutral, or Negative.
Step 1
Initialize model: Model $\leftarrow$ pipeline("sentiment-analysis", model = "bert-multilingual")
Step 2
Compute sentiment score for title: $(L_T, S_T) \leftarrow \text{Model}(T)$ **if** L_T *is "4 stars" or "5 stars"* **then**
  $S_T \leftarrow -S_T$
else if L_T *is "3 stars"* **then**
  $S_T \leftarrow 0$
Step 3
Extract and summarize article content: $C \leftarrow \text{ExtractText(URL)}$, $S \leftarrow C[:200]$ Compute sentiment score: $(L_C, S_C) \leftarrow \text{Model}(S)$ **if** L_C *is "4 stars" or "5 stars"* **then**
  $S_C \leftarrow -S_C$
else if L_C *is "3 stars"* **then**
  $S_C \leftarrow 0$
Step 4
Compute weighted sentiment score: $S_{\text{final}} = 0.3 \cdot S_T + 0.7 \cdot S_C$
Step 5
if $S_{final} \geq 0$ **then**
  Assign label as Positive.
else
  Assign label as Negative.
return S_{final}, Sentiment Label

3.4 Final Decision Making

The final classification of a news article is determined using a logistic regression model trained with Stochastic Gradient Descent (SGD) using **Algorithm 4**. Each agent provides a score, which is used as input features for classification. The feature vector for news item n_i is represented in Equation (1) as:

$$X_i = [S_{i,\text{GNN}}, S_{i,\text{FC}}, S_{i,\text{STC}}, S_{i,\text{SNC}}] \tag{1}$$

where $S_{i,\text{GNN}}, S_{i,\text{FC}}, S_{i,\text{STC}}, S_{i,\text{SNC}}$ are the scores from the GNN, Fact-Checker, Stance-Checker, and Sentiment-Checker, respectively.

The final classification decision is obtained using a logistic regression model trained with Stochastic Gradient Descent (SGD). The model learns weight coefficients w_1, w_2, w_3, w_4, which were initially **randomly initialized** and subsequently optimized during training. These weights determine the contribution of each agent's output to the final classification.

The MAVS score, denoted as S_r, is computed as:

$$S_{\text{r}} = w_1 S_{i,\text{GNN}} + w_2 S_{i,\text{FC}} + w_3 S_{i,\text{STC}} + w_4 S_{i,\text{SNC}} \tag{2}$$

- $S_{i,\text{GNN}}$ - Score from the GNN model
- $S_{i,\text{FC}}$ - Score from the Fact-Checking Agent
- $S_{i,\text{STC}}$ - Score from the Stance-Checking Agent
- $S_{i,\text{SNC}}$ - Score from the Sentiment-Checking Agent

The final score S_r in Equation (2) represents the **weighted fusion** of these agent outputs and is passed through a sigmoid activation function to determine the probability of the news being real or fake.

Applying the sigmoid activation function as shown in Equation (3):

$$P(\text{Real News}) = \frac{1}{1 + e^{-S_{\text{r}}}} \tag{3}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

The classification decision is made as follows:

$$\text{Classify } n_i = \begin{cases} \text{Fake}, & \text{if } P \geq 0.5, \\ \text{Real}, & \text{otherwise.} \end{cases} \tag{4}$$

The threshold $P \geq 0.5$ for classification used in Equation (4) was chosen experimentally, aligning with the standard practice for sigmoid-based binary classification. Since the label encoding assigns 0 to real news and 1 to fake news, a lower sigmoid output (closer to 0) indicates a stronger belief in news being real, whereas a higher value indicates fake. Thus, the 0.5 cutoff serves as a balanced midpoint for binary decision-making.

Algorithm 4 MAVS-Based Fake News Classification Using SGD Logistic Regression

Input : Feature matrix X containing agent scores, Binary labels y (0 = Fake, 1 = Real).
Output: Trained SGD Logistic Regression Model, Classification result for new instances.
Step 1
Construct feature vectors: $X = [S_{i,\text{GNN}}, S_{i,\text{FC}}, S_{i,\text{STC}}, S_{i,\text{SNC}}]$ Assign labels and split dataset: $(X_{\text{train}}, y_{\text{train}}), (X_{\text{test}}, y_{\text{test}})$
Step 2
Initialize and train SGD Logistic Regression Model: model = SGDClassifier(loss = 'log_loss', max_iter = 1000, tol = $1e^{-3}$) model.fit($X_{\text{train}}, y_{\text{train}}$)
Step 3
Predict and compute accuracy: y_{pred} = model.predict(X_{test}), accuracy = $\frac{\text{Correct Predictions}}{\text{Total Predictions}}$ Extract feature weights: $w_1, w_2, w_3, w_4 =$
else
return Fake News (0)

The proposed MAVS framework presents a computationally efficient and scalable approach to fake news detection by integrating multi-agent verification strategies. In the following section, we detail the experimental setup, computational performance analysis, and evaluation metrics to assess the effectiveness of MAVS in real-world applications.

4 Experimental Setup

4.1 System Configuration and Data Processing

All experiments were conducted on an **Intel Core i5 system (16GB RAM) running Ubuntu 20.04 LTS**. The software stack included **Python 3.8+**, **PyTorch 1.10+**, and **Torch Geometric 2.0+** for machine learning, **Hugging Face Transformers** for NLP, **NetworkX** for graph processing, **Selenium** and **BeautifulSoup** for web scraping, and **Stochastic Gradient Descent (SGD)** for optimization.This section details the data preparation, processing pipeline, evaluation metrics, and final classification algorithm (**Algorithm 5**) used in the MAVS framework for fake news detection.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

UPFD Politifact dataset [7], containing **314 news propagation graphs** (157 fake), was used to train a **Graph Neural Network (GNN)** for fake news detection. The dataset was split into **70% training, 10% validation, and 20% testing**, ensuring a balanced distribution of **115 fake** and **105 real** graphs as shown in **Fig. ??**.

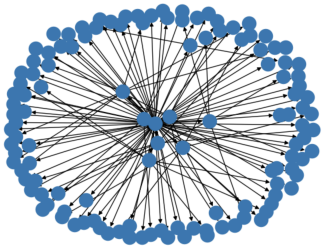

Fig. 3: Hierarchical Tree-Structured Graph Used in the GNN Model

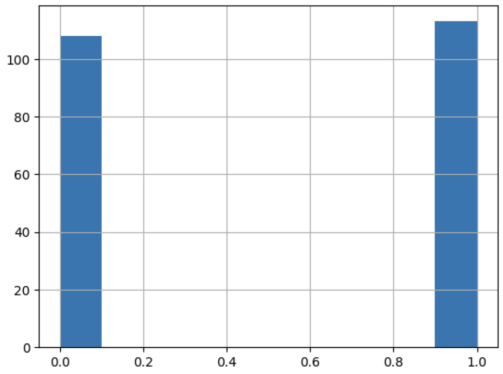

Fig. 4: Dataset labeling for MAVS.

The **MAVS framework** integrates GNNs with AI agents for fact-checking, stance detection, and sentiment analysis. News propagation was modeled using **NetworkX**, with **SpaCy Word2Vec embeddings** as input. A three-layer **GATConv** architecture (**310 input, 128 hidden, 128 output**) was optimized with **Adam (learning rate: 0.01, activation: LeakyReLU)**. To evaluate adversarial robustness, a **Multi-Agent Reinforcement Learning (MARL)** attack [21] was applied, modifying propagation graphs via **bot-driven adversarial edges**. Additionally, the **PolitiFact++ dataset** [18] was used to test the impact of **Human-written Fake news (HF)** and **LLM-generated Fake news (MF)**. Final classification was performed using **SGD**, where the outputs of all AI agents were weighted and fused, as outlined in **Algorithm 4**.

4.2 Evaluation Metrics

The performance of the MAVS framework was evaluated using standard classification metrics:

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Precision

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

Recall

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

F1-Score

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Fig.5 represents the structure of confusion matrix. The formulas for the evaluation metrics are given by Equations (5), (6), (7), and (8).

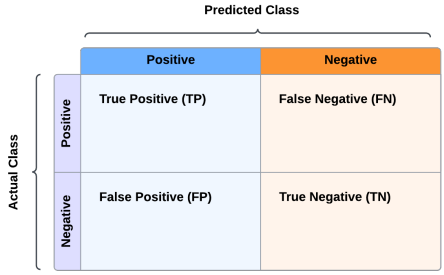

Fig. 5: Confusion Matrix

The Accuracy and F1-score were compared across all baseline models: HGFND, UPFD-SAGE (GraphSAGE), UPFD-GAT, LSTM, BERT, HiSS, TextCNN, FactAgent with Expert Workflow, CNN, and RoBERTa with RoBERTa-base features, alongside our MAVS framework. Additionally, he impact of adversarial attacks was analyzed specifically for BERT, RoBERTa, GraphSAGE, MAVS by using a **4-fold cross-validation**, evaluating their performance under attack scenarios.

5 Results and Analysis

This section presents the evaluation of our MAVS framework, including the following mentioned topics

- Accuracy and F1-score comparisons with baseline models
- Post-attack performance analysis
- An ablation study regarding different AI Agents.

5.1 Accuracy and F1-score Comparison of Baseline Models with MAVS

To evaluate the effectiveness of our MAVS framework, we compared its Accuracy and F1-score against several baseline models: HGFND, UPFD-SAGE (GraphSAGE), UPFD-GAT, LSTM, BERT, HiSS, TextCNN, FactAgent with Expert Workflow, CNN, and RoBERTa with RoBERTa-base features.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

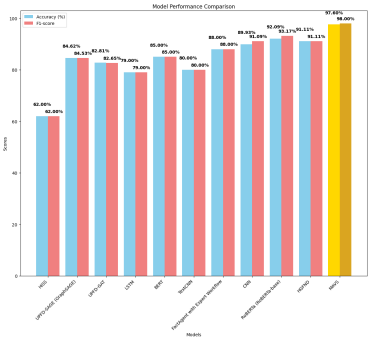

Fig. 6: Accuracy and F1-score comparison of baseline models and MAVS.

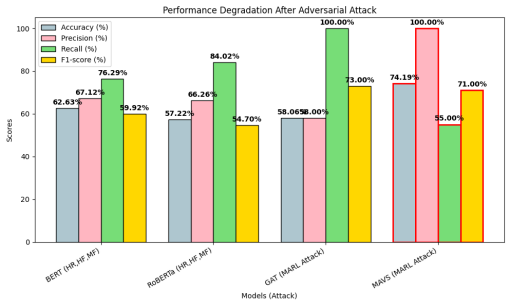

Fig. 7: Performance comparison after adversarial attack for BERT, RoBERTa, GraphSAGE, and MAVS.

The bar chart in **Fig. 6** visualizes the performance comparison of different models. MAVS achieves the highest accuracy (97.60%) and F1-score (98.00), significantly outperforming other models. The closest competitor, RoBERTa (RoBERTa-base), attains an accuracy of 92.09% and an F1-score of 93.17%, indicating that MAVS leverages a more effective fusion of features to enhance predictive performance.

Models such as CNN [1], HGFND [10], and FactAgent with Expert Workflow [11] show relatively strong performance, achieving accuracy and F1-scores above 88%. However, traditional models like LSTM (79.00%) and TextCNN (80.00%) fall behind, demonstrating their limitations in capturing the complex relationships in fake news detection.HiSS was the worst performer with least accuracy (62.00%).

5.2 Performance Comparison After Adversarial Attack

Notably, GraphSAGE and MAVS were subjected to **MARL** attack as implemented by Wang et al. [21], where bot agents manipulate the news propagation graph by injecting misleading nodes and edges to mimic legitimate user behavior. This coordinated manipulation alters the graph structure to degrade detection performance. In contrast, BERT and RoBERTa were evaluated on adversarially perturbed versions of the **PolitiFact**++ dataset.

The results in **Fig. 7** reveal that adversarial attacks significantly degrade the performance of all models, but the extent of the degradation varies. BERT and RoBERTa experience a substantial drop, particularly in Recall and F1-score, indicating that they misclassify a significant portion of adversarially perturbed samples. GAT, despite maintaining a high Recall of 100%, suffers from a considerable Precision drop, suggesting that it incorrectly predicts a large number of false positives.

MAVS, however, demonstrates better robustness with an Accuracy of 74.19% and a balanced F1-score of 71%, outperforming the other models. The Precision

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

score of 100% for MAVS suggests that it avoids false positives, but its lower Recall (55%) indicates some difficulty in capturing all real instances. This performance stability highlights MAVS's resilience to adversarial attacks, likely due to its fact-checking mechanisms.

Notably, one of the major findings is shown in the confusion matrices in **Fig. 8** that reveals that after the adversarial attack, MAVS still maintains a balance between true positives (20) and false negatives (16) due to its fact-checking agent, indicating **partial resilience to adversarial interference**. Conversely, GNNs demonstrate a severe decline, completely failing to classify "Fake News" instances, as all predictions default to "Real." This highlights GNNs' heightened vulnerability to adversarial perturbations and the advantage of MAVS's architecture in handling such challenges.

5.3 Ablation Study

To analyze the contributions of individual components in the MAVS framework, we perform an ablation study on the GNN, Sentiment Checker, Stance Checker, and Fact Checker.

The results indicate that the **GNN** achieves the highest precision (94.29%), ensuring minimal false positives, while maintaining a balanced F1-score of 92.96%, demonstrating its ability to generalize well. The **Fact Checker** exhibits the highest recall (97.22%), making it the most effective at detecting fake news instances. The **Sentiment Checker** and **Stance Checker**, while weaker in precision, provide valuable complementary information. Their recall values indicate a tendency to detect more fake news instances, which is crucial in adversarial settings where attackers attempt to manipulate narratives.

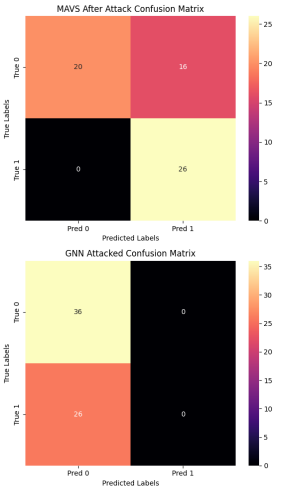

Fig. 8: Confusion Matrices for MAVS and GNN After Attacks

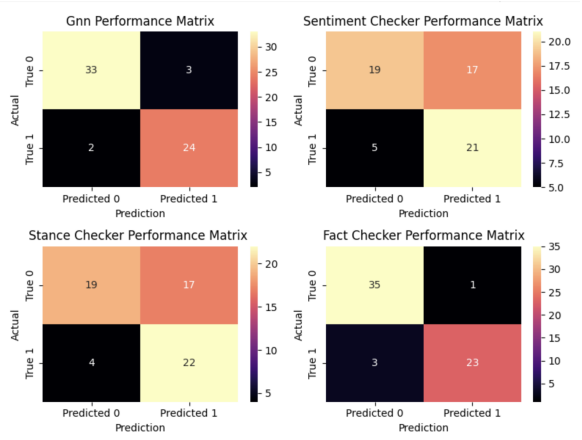

Fig. 9: Confusion matrices for GNN, Sentiment Checker, Stance Checker, and Fact Checker.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

The confusion matrices in **Fig. 9** provide additional insights, illustrating that the Fact Checker exhibits strong classification ability, misclassifying only a few instances. In contrast, the Sentiment and Stance Checkers show higher false positives, underscoring the need for their weighted contribution in MAVS.

6 Conclusion and Future Work

This research presents **MAVS (Multi-Agent Verification System)**, an ensemble-based framework integrating **Graph Neural Networks (GNNs)** with **generative AI models** for fact-checking, stance detection, and sentiment analysis. By combining **news propagation modeling** with **text-based verification**, MAVS provides a robust and multidimensional approach to fake news detection.

Using a **weighted aggregation mechanism** optimized via **SGD-based logistic regression**, MAVS achieves **97.6% accuracy** and **98% F1-score**, outperforming **RoBERTa, CNNs, and GAT-based models**. The framework demonstrates **partial resilience to adversarial attacks**, with the fact-checking agent maintaining **74.19% accuracy** under adversarial conditions and ensuring **100% precision**. The fusion of **graph-based structural analysis, linguistic verification, and factual retrieval** enables MAVS to be a **scalable, interpretable, and resilient** misinformation detection system.

Future work will enhance MAVS by enabling real-time adaptive weighting using RL, strengthening multilingual and adversarial defenses, and integrating RAG-based APIs for instant fact-checking. Additionally, MAVS could be used in simulations to study misinformation's impact on public opinion dynamics.

References

1. Alghamdi, J., Lin, Y., Luo, S.: A comparative study of machine learning and deep learning techniques for fake news detection. Information **13**, 576 (2022). https://doi.org/10.3390/info13120576
2. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of Economic Perspectives **31**, 211–236 (2017)
3. Arquam, M., Singh, A., Cherifi, H.: Impact of seasonal conditions on vector-borne epidemiological dynamics. IEEE Access **8**, 94510–94525 (2020). https://doi.org/10.1109/ACCESS.2020.2995650
4. Cherifi, H.: Complex networks and their applications (2014)
5. Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of twitter users. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. pp. 192–199 (10 2011). https://doi.org/10.1109/PASSAT/SocialCom.2011.34
6. Diop, I.M., Cherifi, C., Diallo, C., Cherifi, H.: Revealing the component structure of the world air transportation network. Applied Network Science **6**, 1–50 (2021)
7. Dou, Y., Shu, K., Xia, C., Yu, P.S., Sun, L.: User preference-aware fake news detection. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20

8. Drif, A., Zerrad, H.E., Cherifi, H.: Ensvae: Ensemble variational autoencoders for recommendations. IEEE Access **8**, 188335–188351 (2020)
9. Hardalov, M., Hristov, T., Nakov, P., Koychev, I.: Few-shot stance detection for political claims. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 7759–7771 (2022)
10. Jeong, U., Ding, K., Cheng, L., Guo, R., Shu, K., Liu, H.: Nothing stands alone: Relational fake news detection with hypergraph neural networks (2022)
11. Li, X., Zhang, Y., Malthouse, E.C.: Large language model agent for fake news detection (2024)
12. Monti, F., Frasca, F., Eynard, D., Mannion, D., Bronstein, M.M.: Fake news detection on social media using geometric deep learning. In: Proceedings of the National Academy of Sciences. vol. 116, pp. 25738–25743 (2019)
13. Pennycook, G., Rand, D.G.: The psychology of fake news. Trends in Cognitive Sciences **25**, 388–402 (2021). https://doi.org/https://doi.org/10.1016/j.tics.2021.02.007
14. Qureshi, K., Malick, R., Sabih, M., Cherifi, H.: Complex network and source inspired covid-19 fake news classification on twitter. IEEE Access **9**, 1–1 (2021). https://doi.org/10.1109/ACCESS.2021.3119404
15. Qureshi, K.A., Malick, R.A.S., Sabih, M., Cherifi, H.: Deception detection on social media: A source-based perspective. Knowledge-Based Systems **256**, 109649 (2022). https://doi.org/https://doi.org/10.1016/j.knosys.2022.109649, `https://www.sciencedirect.com/science/article/pii/S0950705122008346`
16. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., Liu, Y.: Combating fake news: A survey on identification and mitigation techniques. ACM Transactions on Intelligent Systems and Technology (TIST) **10**, 1–42 (2019)
17. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286 (2018)
18. Su, J., Zhuo, T.Y., Mansurov, J., Wang, D., Nakov, P.: Fake news detectors are biased against texts generated by large language models (2023)
19. Su, X., Xue, S., Liu, F., Wu, J., Yang, J., Zhou, C., Hu, W., Paris, C., Nepal, S., Jin, D., Sheng, Q., Yu, P.: A comprehensive survey on community detection with deep learning. IEEE Transactions on Neural Networks and Learning Systems **PP**, 1–21 (2022). https://doi.org/10.1109/TNNLS.2021.3137396
20. Thorne, J., Chen, M., Myrianthous, G., Pu, J., Wang, X., Vlachos, A.: Fact extraction and verification (fever). In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 809–819 (2018)
21. Wang, H., Dou, Y., Chen, C.H., Sun, L., Yu, P.S., Shu, K.: Attacking fake news detectors via manipulating news social engagement. In: Proceedings of the ACM Web Conference 2023 (WWW '23). pp. 3978–3986 (2023)
22. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications (2021)
23. Zhu, J., Peng, X., Wang, L.: Sentiment analysis meets fake news detection: A deep-learning approach. Expert Systems with Applications **167**, 114171 (2021)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97557-8_20