International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Semantic Paradigms for Deceptive Content Detection in Digital Media

Diksha Durgapal[1], Yogita Parmar[2], Pooja Bhaliya[3]

[1]ITM SLS Baroda University, Vadodara, Gujarat diksha
[2]ITM SLS Baroda University, Vadodara, Gujarat
[3]ITM SLS Baroda University, Vadodara, Gujarat
[1]durgapal@itmbu.ac.in, [2]yogita.parmar@itmbu.ac.in,
[3]poojabhaliya.cse@itmbu.ac.in

Abstract

The spread of deceptive content through digital media platforms has become a persistent concern, particularly due to its potential impact on public trust and informed decision-making. Earlier detection systems relied largely on surface-level textual features and were often ineffective when deceptive content adopted credible linguistic forms. This review examines recent research efforts that apply semantic analysis techniques to deceptive content detection between 2020 and 2025. The selected studies are analyzed with respect to their modeling approaches, datasets, and evaluation strategies. The review shows that transformer-based semantic models and evidence-aware verification frameworks have improved detection performance; however, limitations related to generalization, explainability, and deployability remain. These observations motivate the need for more robust and practically applicable semantic detection systems.

Keywords: Semantic analysis, deceptive content detection, misinformation, transformer models, fact verification, explainable artificial intelligence, digital media

1. Introduction

1.1 Background

Digital media platforms such as social networking sites, online news portals, discussion forums, and messaging applications have revolutionized information dissemination. While these platforms facilitate rapid access to information, they have also enabled the widespread circulation of deceptive content. Such content often exploits semantic ambiguity, emotional framing, and selective presentation of facts to mislead audiences.

Deceptive content poses serious risks to public trust, democratic processes, and public health. Manual content verification is insufficient due to the massive scale of online information generation, motivating the development of automated detection systems.

1.2 Limitations of Traditional Detection Approaches

Early automated systems primarily relied on lexical cues, stylistic features, and shallow machine learning algorithms. Although effective against overtly false content, these approaches fail when deceptive information closely resembles legitimate news in tone and structure. As deceptive narratives become more sophisticated, detection systems must move beyond word-level patterns and incorporate deeper semantic understanding.

Recent advances in Natural Language Processing (NLP), particularly the emergence of transformer-based architectures, have enabled models to capture contextual relationships and semantic meaning at a much deeper level. These developments have significantly reshaped research in deceptive content detection.

1.3 Motivation for Semantic Analysis

Semantic analysis focuses on understanding meaning, intent, and contextual relationships rather than relying solely on surface-level features. By modeling how information is expressed and how it relates to real-world facts, semantic approaches offer improved robustness against linguistically complex deception. Techniques such as contextual embeddings, evidence-based reasoning, and intent modeling have shown promise in addressing the shortcomings of traditional methods.

However, despite these advances, semantic deception detection remains an open research challenge due to issues related to generalization, explainability, and computational efficiency.

1.4 Objectives of the Review

The primary objective of this review is to examine and synthesize recent research on semantic analysis techniques for deceptive content detection in digital media. Specifically, the review aims to analyze how semantic modeling approaches—such as transformer-based representations, evidence-aware verification frameworks, explainable detection models, and intent-driven semantic learning—have been applied to address the limitations of traditional detection systems. In addition, this work seeks to compare representative studies based on their methodologies, datasets, and evaluation strategies, highlighting both their strengths and limitations. By identifying unresolved challenges related to generalization, explainability, multilingual robustness, and real-world deployability, the review provides a structured understanding of current research gaps and outlines future directions for developing robust, trustworthy, and scalable semantic deception detection systems.

2. Related Work

Research on deceptive content detection has evolved significantly over the last decade, moving from surface-level text analysis to deep semantic understanding. Recent studies increasingly focus on capturing contextual meaning, factual consistency, and deceptive intent using advanced semantic modeling techniques. The reviewed literature can be grouped based on techniques, algorithms, and application domains.

2.1 Technique-Based Categorization

2.1.1 Semantic Text Classification Using Transformers

Transformer-based language models have become the dominant technique for deceptive content detection due to their ability to learn contextual semantic representations. Studies such as Patwa et al. (2020), Kaliyar et al. (2022), and Shaar et al. (2022) fine-tune pre-trained transformer architectures to classify deceptive content directly from textual input. These approaches rely on attention mechanisms to model word relationships across entire documents, enabling improved contextual understanding compared to traditional neural networks.

While transformer fine-tuning yields high performance on benchmark datasets, these models largely treat deception detection as a classification task and do not verify the factual correctness of claims.

2.1.2 Evidence-Aware Semantic Verification

To address the limitation of classification-only approaches, verification-based methods integrate external evidence into the detection process. HoVer (Jiang et al., 2020) introduces a dataset and framework that enables semantic reasoning across multiple evidence documents. Hassan et al. (2020) further propose a pipeline that connects fact verification with fake news detection, emphasizing claim–evidence consistency.

These approaches enhance semantic reliability by grounding predictions in factual evidence. However, they introduce additional system complexity and increased computational cost.

2.1.3 Explainable Semantic Models

Explainability has emerged as a crucial requirement for trustworthy deceptive content detection. Atanasova et al. (2021) propose transformer-based models that generate natural-language explanations for fact-checking decisions. EX-FEVER (Schuster et al., 2024) further advances explainability by providing multi-hop reasoning paths and evidence-backed explanations.

Although these models improve transparency, explanation faithfulness and computational efficiency remain open challenges.

2.2 Algorithm-Based Categorization

From an algorithmic perspective, existing research can be divided into four major categories:

- **Transformer Fine-Tuning Models:** Rely on contextual embeddings for classification (Patwa et al., 2020; Kaliyar et al., 2022; Shaar et al., 2022).
- **Verification Pipelines:** Perform semantic verification before final classification (Jiang et al., 2020; Hassan et al., 2020).
- **Multi-Hop Graph Reasoning Models:** Represent claims and evidence as graphs to support complex reasoning (Jiang et al., 2020; Schuster et al., 2024).
- **Intent–Semantic Joint Learning Models:** Incorporate intent modeling alongside semantic representation (Liu et al., 2025).

This categorization highlights a gradual shift from text-only classification toward reasoning-driven and intent-aware semantic frameworks.

2.3 Application-Oriented Classification

Semantic deception detection has been applied across various domains. Health misinformation detection, particularly during the COVID-19 pandemic, is addressed in Patwa et al. (2020). Political misinformation and public discourse manipulation are explored in Hassan et al. (2020) and Liu et al. (2025). Multilingual deceptive content detection, especially for low-resource languages, is examined in Chakravarthi et al. (2023). Fact-checking and explainable verification systems are central to Jiang et al. (2020) and Schuster et al. (2024).

3. Comparative Analysis

This section presents a comparative analysis of selected research works to highlight differences in methodology, semantic depth, datasets, and practical limitations.

3.1 Overview of Comparative Studies

A structured comparison of recent approaches highlights the trade-offs between semantic depth and computational efficiency. Table 1 synthesizes key studies reviewed in this paper, categorizing them by methodology, primary focus, strengths, and limitations. The comparison illustrates a clear progression from classification-oriented models to more complex reasoning and intent-aware frameworks.

Table 1: Summary of Comparative Studies in Semantic Deception Detection

Study	Methodology	Key Focus	Key Strength	Limitation
Patwa et al. (2020)	Transformer Fine-Tuning	Health (COVID-19)	High accuracy on specific domain	Poor cross-domain generalization
Jiang et al. (2020)	Multi-hop Verification	Fact Extraction	Reasoning across multiple docs	High computational overhead
Hassan et al. (2020)	Verification Pipeline	Pol/News verification	Grounded in evidence	Latency issues
Atanasova et al. (2021)	Explainable Semantic	Fact-checking Explanations	Natural language justification	Faithfulness of explanation

Kaliyar et al. (2022)	BERT-based Classification	Rumour Detection	Strong semantic representation	Lack of explicit reasoning
Schuster et al. (2024)	Explainable Verification	Multi-hop Reasoning	Transparent reasoning paths	Complexity in retrieval
Liu et al. (2025)	Intent-Semantic Learning	Deceptive Intent	Models persuasive intent	Dependency on intent labels

3.2 Critical Discussion

The analysis of Table 1 reveals that while transformer-based semantic models (Patwa et al., 2020; Kaliyar et al., 2022; Shaar et al., 2022) consistently outperform traditional machine learning approaches on benchmark datasets, they rely heavily on implicit semantics. Their strength lies in capturing contextual relationships within text, which helps distinguish deceptive content that closely resembles legitimate information, but they often struggle with factual inconsistencies that require external knowledge.

Evidence-aware and verification-based approaches (Jiang et al., 2020; Hassan et al., 2020) address this by incorporating factual grounding, improving reliability in high-stakes domains. However, this comes at the cost of computational complexity. Multi-hop reasoning models further improve semantic understanding but require extensive resources and structured evidence, limiting their scalability in real-time applications.

Explainable models (Atanasova et al., 2021; Schuster et al., 2024) add an essential layer of transparency, yet faithful explanation generation remains an open challenge. Finally, intent-aware frameworks (Liu et al., 2025) represent a promising direction for detecting manipulative content, though they currently face hurdles regarding data availability. In summary, no single approach currently addresses all aspects of detection; classification models excel in efficiency, while verification models provide necessary reliability at the expense of speed.

4. Research Challenges and Gaps

Despite notable progress in semantic analysis–based deceptive content detection, a critical examination of existing literature reveals several unresolved challenges. These can be broadly categorized into semantic depth, generalization capabilities, system trustworthiness, and computational feasibility.

4.1 Semantic and Logical Limitations

A fundamental gap in current research is the **separation between semantic detection and factual verification**. Most transformer-based approaches (Patwa et al., 2020; Kaliyar et al., 2022) rely on implicit contextual embeddings, which capture linguistic patterns but do not explicitly reason about logical consistency, entailment, or contradiction. As a result, these models often misclassify content that is factually incorrect but linguistically coherent. Although verification-based pipelines exist (Jiang et al.,

2020), they are rarely fully integrated into end-to-end detection frameworks, leaving systems vulnerable to well-written but false narratives.

- **Gap:** Lack of integrated frameworks that combine explicit logical reasoning with implicit semantic classification.

4.2 Generalization and Adaptability

Current models exhibit poor **cross-domain and temporal generalization**. Studies show that performance degrades significantly when models trained on specific topics (e.g., politics) are applied to new domains (e.g., health) or evolving narratives (Shaar et al., 2022). Furthermore, the research landscape is heavily skewed toward **English-language content**, leaving a significant gap in detecting deceptive content in low-resource and code-mixed languages (Chakravarthi et al., 2023).

- **Gap:** Inability of static models to adapt to dynamic misinformation trends and diverse linguistic environments.

4.3 Trustworthiness and Evaluation

There is a critical disconnect between **model performance metrics and real-world reliability**. Standard metrics like accuracy and F1-score fail to capture semantic robustness or confidence calibration. Additionally, while explainability is a growing focus, current methods often rely on post-hoc visualizations (e.g., attention maps) that do not necessarily reflect the model's actual decision-making process (Atanasova et al., 2021).

- **Gap:** Absence of semantic reliability metrics and "reasoning-aligned" explanations that build user trust.

4.4 Computational Efficiency vs. Semantic Depth

Advanced techniques such as multi-hop reasoning (Schuster et al., 2024) significantly improve detection accuracy but introduce prohibitive **computational overhead**. The requirement for extensive memory and multiple evidence retrieval steps makes these models unsuitable for real-time applications on social media platforms where latency is critical.

- **Gap:** Trade-off between deep semantic verification and the scalability required for real-time deployment.

4.5 Summary of Challenges

Table 2 summarizes the critical research gaps identified in this review and suggests potential directions for future investigation.

Table 2: Strategic Summary of Research Gaps and Future Directions

Challenge Category	Critical Gap	Future Research Direction
Reasoning	Models classify style rather than verifying facts.	Integration of Neuro-Symbolic AI for explicit logical reasoning.
Generalization	Models fail on new topics and languages.	Few-shot learning and cross-lingual transfer learning.
Trust	Explanations are often unfaithful to model logic.	Chain-of-Thought (CoT) prompting and reasoning-aligned generation.
Efficiency	High accuracy models are too slow for real-time use.	Knowledge distillation to compress reasoning models for edge deployment.

5. Conclusion and Future Work

Based on the studies reviewed in this paper, it is clear that semantic analysis has become central to recent deceptive content detection research. Transformer-based models have enabled improved contextual understanding, while evidence-aware frameworks contribute to greater factual reliability. Nevertheless, the reviewed literature also reveals persistent challenges. Many approaches rely on implicit semantic representations and struggle to generalize across domains, languages, and evolving narratives. In addition, the computational demands of verification-based models raise concerns regarding real-world deployment.

Future research should therefore focus on developing semantic models that balance reasoning capability with efficiency. Greater attention is needed toward multilingual datasets, realistic evaluation settings, and explanation methods that accurately reflect model decision processes. Incorporating human feedback into detection pipelines may further improve robustness. Addressing these issues will be essential for advancing deceptive content detection systems from experimental settings toward practical deployment.

References

1. Atanasova, S., et al. (2021). Generating fluent fact-checking explanations. arXiv preprint arXiv:2112.06924.
2. Chakravarthi, R., et al. (2023). Fake news detection in Dravidian languages. ACL Anthology. DOI: 10.18653/v1/2025.dravidianlangtech-1.29
3. Hassan, N., et al. (2020). Connecting fact verification and fake news detection. arXiv preprint arXiv:2010.05202.
4. Jiang, Y., et al. (2020). HoVer: A dataset for many-hop fact extraction and claim verification. Findings of EMNLP. DOI: 10.18653/v1/2020.findings-emnlp.309
5. Kaliyar, R. K., et al. (2022). Evaluating BERT-based models for rumour and fake news detection. arXiv preprint arXiv:2203.07731.
6. Liu, Z., et al. (2025). Intent–semantic joint learning for fake news detection..
7. Patwa, A., et al. (2020). Two-stage transformer model for COVID-19 fake news detection. NLP4IF. Link: https://aclanthology.org/2020.nlp4if-1.1/
8. Schuster, S., et al. (2024). EX-FEVER: Multi-hop explainable fact verification. Findings of ACL. DOI: 10.18653/v1/2024.findings-acl.556
9. Shaar, F., et al. (2022). Transformer-based fake news detection. CLEF CheckThat!. arXiv:2307.00610v2
10. Wang, J., et al. (2024). Knowledge-guided semantic fake news detection. Electronics. DOI: https://doi.org/10.48550/arXiv.2404.01336