Journal of Computer Science

Original Research Paper

Mitigating the Evidence-Related Factors in Automated Fact-Checking

[1]Aruna Shankar, [1]Narayanan Kulathuramaiyer, [1]Johari Bin Abdullah and [2]Muthukumaran Pakkirisamy

[1]*Faculty of Computer Science and Information Technology, University of Malaysia, Sarawak, Malaysia*
[2]*Department of General Education, American University of Phnom Penh, Phnom Penh, Cambodia*

Article history
Received: 12-10-2024
Revised: 18-11-2024
Accepted: 16-12-2024

Corresponding Author:
Aruna Shankar
Faculty of Computer Science
and Information Technology,
University of Malaysia,
Sarawak, Malaysia
Email: 21010101@siswa.unimas.my

Abstract: The rapid proliferation of digital misinformation highlights the urgent need for robust automated fact-checking systems that can accurately distinguish truth from falsehood. A persistent challenge for these systems is the occurrence of false positives, where truthful information is incorrectly flagged as misleading due to limitations in evidence assessment, including insufficient evidence, logical inconsistencies, and conflicting information. This research introduces a novel two-phase approach to address these issues. In Phase 1, relationships between claims and evidence are modeled using a graph-based mechanism to identify evidence-related shortcomings that contribute to false positives. Phase 2 enhances evidence quality by integrating domain-specific knowledge, employing pretrained language models such as BERT, RoBERTa, and BioBERT across diverse datasets like FEVER, LIAR-Plus, HoVER, and PubMed. Our findings demonstrate that addressing these evidence-related factors significantly reduces false positives, resulting in more accurate fact-checking. These results underscore the effectiveness of our enhanced evidence assessment method, providing valuable insights for developing reliable fact-checking systems adaptable across multiple domains. This research lays a foundation for future innovations in misinformation mitigation, fostering a more trustworthy digital information landscape.

Keywords: Fact-Checking, False Positives, Evidence-Related Factors, Misinformation, Insufficient Evidence, Claim-Evidence Mapping

Introduction

The massive increase in digital content production presents a significant challenge in maintaining accuracy and truthfulness. In these circumstances, automated fact-checking systems have become essential to upholding the credibility of public discourse (Thorne and Vlachos, 2018). Initially, these systems relied on basic techniques, primarily matching claims against verified facts Lee *et al.* (2023). However, as misinformation has grown more complex, the demand for sophisticated approaches has intensified Schlichtkrull *et al.* (2023).

Advanced techniques are now crucial for addressing the complexities of false narratives and intricate deceptions Vladika *et al.* (2023). Despite these advancements, even the most sophisticated fact-checking systems face significant limitations. The multifaceted nature of evidence can lead these systems to incorrectly flag accurate information as false Martín *et al.* (2022). Such errors often stem from various evidence-related challenges, including insufficient support, discrepancies between sources, and logical contradictions Zhu *et al.* (2021). Recognizing and addressing these factors is essential; fact-checking systems must be designed with a heightened awareness of these challenges and equipped with mechanisms to mitigate their impact.

The challenge, therefore, lies not only in detecting falsehoods but also in discerning the nuances that signify truth, especially when evidence is incomplete or open to interpretation. A crucial element of this nuanced analysis is understanding the complex interplay between a claim and its supporting or refuting evidence. This requires a mechanism capable of mapping and visualizing these relationships, moving beyond mere evidence retrieval to critically assess the relevance, coherence, and overall contribution of the evidence to veracity assessment. Integrating such a mapping mechanism into fact-checking systems can help reduce false positives, ensuring that genuinely misleading content is flagged while credible information is preserved.

Science Publications

Previous studies by Atanasova *et al*. (2019; 2020; 2022) have shown that fact-checking systems can produce misleading results due to inaccuracies, ambiguities, and biases in claims. While some research has examined the role of evidence in fact-checking, existing systems often fail to thoroughly analyze the evidence-related factors necessary for accurate verification Barik *et al*. (2022); Barve *et al.* (2022). Notably, there is a gap in research on developing robust mapping mechanisms that connect claims to their supporting or refuting evidence. Addressing this gap is essential for understanding cases where evidence may be insufficient, conflicting, or open to interpretation. Our study focuses on this critical need by developing and integrating a novel mapping mechanism within an automated fact-checking system.

We aim to identify key evidence-related factors that contribute to inaccurate predictions, such as insufficient support for claims, conflicting information from various domains, and issues with logical coherence. Our mapping mechanism will enable us to visually represent and analyze these factors, revealing potential inconsistencies, evaluating source quality, and assessing the overall strength and coherence of the evidence supporting a claim.

We closely examine the performance of our mapping mechanism by focusing on instances where the fact-checking model produces inaccurate predictions. The goal is to understand the nature and causes of these inaccuracies, particularly how they relate to the evidence-related factors identified through the mapping process. Our analysis takes a two-pronged approach. First, we systematically introduce specific evidence-related factors into the fact-checking model and observe their individual and combined effects on performance, enabling us to determine which factors are most strongly linked to inaccurate predictions. Second, we refine the mapping mechanism by iteratively improving the quality and representation of the initially identified evidence. This may involve incorporating source reliability metrics, enhancing the visualization of logical connections, or developing more nuanced categorizations of evidence types.

This systematic process helps us assess the incremental effect of each factor on fact-checking accuracy, ensuring that the model's effectiveness improves without altering the fundamental meaning of the claims being evaluated. To rigorously test our approach, we experiment with established language models such as BERT Devlin *et al*. (2019), RoBERTa Liu *et al.* (2019), and BioBERT Lee *et al.* (2020). These models are evaluated by strategically excluding and including specific evidence-related factors, using comprehensive datasets such as FEVER Thorne *et al*. (2018), HoVER Jiang *et al*. (2020a), LIAR-PLUS Alhindi *et al*. (2018) and PubMed Dernoncourt and Lee (2017).

We assess the effectiveness of our mapping mechanism through a two-part evaluation. First, we use standard metrics, including accuracy, precision, recall, and F1-score, to measure the model's overall performance. Second, we conduct expert evaluations, focusing on instances where evidence-related factors could lead to misleading fact-checking results. This human-in-the-loop approach provides qualitative insights into the model's strengths and weaknesses, particularly in navigating complex evidence landscapes.

Our findings demonstrate that the proposed mapping mechanism effectively identifies and highlights misleading evidence-related factors, resulting in a more nuanced and accurate assessment of claim veracity. This research not only advances the field methodologically but also offers practical insights for improving the precision and reliability of automated fact-checking systems. By promoting rigorous benchmarks and continuous improvement, we aim to contribute to a digital environment where truth prevails, fostering a more informed global community. To aid in understanding the key terms and acronyms used throughout this study, a list of abbreviations is provided in Table (1).

Related Work

The rapid expansion of digital content and the rise of unreliable and false narratives present a significant challenge in maintaining information integrity. This underscores the urgent need for effective automated fact-checking systems. While technological advancements are vital for the development of these systems, it is equally essential to examine their underlying structures, especially regarding evidence evaluation. This section reviews current fact-checking approaches, focusing on the intricate relationship between evidence-related factors and the prediction of inaccuracies.

Table 1: List of abbreviations and their descriptions

Abbreviation	Description
BERT	Bidirectional encoder representations from transformers
RoBERTa	Robustly optimized BERT
BioBERT	BERT model pre-trained on large-scale biomedical corpora for biomedical text mining
FEVER	Fact extraction and verification dataset, used for fact-checking tasks
LIAR	A dataset consisting of political statements and their truthfulness labels
HoVER	A dataset for many-hop fact extraction and claim verification
PubMed	A dataset consisting of biomedical research articles used for scientific claim verification

A substantial body of research Bekoulis *et al*. (2021); Kruengkrai *et al.* (2021); Ostrowski *et al.* (2021); Sarrouti *et al.* (2021); Sathe *et al.* (2020); Thorne and Vlachos (2018) have explored the role of evidence in validating or refuting claims. Numerous studies by Augenstein *et al*. (2019); and Popat *et al.* (2018a) reveal that automated systems frequently struggle with the complexity of evidence, particularly in terms of adequacy, relevance, and source credibility. The reliability of these systems is strongly influenced by their ability to assess evidence contextually, evaluate its credibility, and integrate information from multiple sources.

Several studies Augenstein *et al.* (2019); Rashkin *et al.* (2017); Thorne (2021); Walter *et al.* (2020); Yang *et al.* (2024) have highlighted the impact of evidence presentation on fact-checking outcomes, with variations in evidence leading to inconsistent results across different systems. These disparities underscore the challenges of creating consistent frameworks for evidence evaluation that are both robust and adaptable to diverse informational scenarios. Additionally, some research by Das *et al*. (2023); Freeze *et al*. (2021); and Pathak (2022) has emphasized the limitations of current models in capturing and utilizing subtle evidential nuances for well-informed decision-making.

Numerous investigations by Barve *et al*. (2022); Bekoulis *et al*. (2021); Oh *et al.* (2022); Conroy *et al.* (2015); Hassan *et al.* (2017); Smeros *et al.* (2021) have underscored the need for adaptable methods that can analyze various forms of evidence, including user-generated content, expert opinions and statistical data. These methods are critical in advancing fact-checking by ensuring a comprehensive evaluation of complex claims. Collectively, these studies Atanasova *et al*. (2019); Barve *et al*. (2022); Miranda *et al.* (2019); and Samarinas *et al*. (2021) provide foundational insights into how evidence assessment informs automated fact-checking and suggest directions for future research in this area. They highlight the complexities of evidence evaluation and emphasize its critical role in improving prediction accuracy.

A recurring challenge identified in prior research by Pennycook *et al*. (2020); and Thorne *et al.* (2018) is the issue of insufficient evidence, which can lead to false negatives (where false claims are classified as true) or false positives (where true claims are misclassified as false). Insufficient evidence limits a system's ability to verify claims, resulting in either conservative classifications or incorrect rejections Choi and Ferrara (2024); Forstmeier *et al.* (2017); Pennycook *et al.* (2020); Rosso *et al*. (2020). Furthermore, when evidence is ambiguous or contradictory, automated systems may struggle to identify the most credible sources or effectively synthesize information, complicating accurate classification Aly *et al.* (2021); Azevedo (2018); Hassan *et al.* 2017; Huynh and Papotti (2018); Singh *et al*. (2021).

Another critical issue is the logical coherence of evidence. Some fact-checking systems may overly rely on pattern recognition or superficial characteristics, failing to fully evaluate the logical framework of the evidence. This can lead to inaccurate classifications, particularly when evidence appears to support a claim initially but does not substantiate it upon closer analysis Ciampaglia *et al*. (2015); Gencheva *et al.* (2019); Huynh and Papotti (2018); Kotonya and Toni (2024); Yang *et al*. (2022); Yao *et al*. (2023); Zhang and El-Gohary (2017). Cross-domain inconsistencies also pose significant challenges, as fact-checking systems must assess the truthfulness of statements based on information from unrelated domains. Systems must navigate conflicting data and assess the trustworthiness of each source.

Research studies by Gencheva *et al*. (2019); Kao and Yen (2024); Li *et al*. (2022); Sathe *et al*. (2020); and Tsai (2023) have explored methods for improving the identification and resolution of domain-specific inconsistencies. One approach involves hybrid models that combine rule-based methods with statistical learning to more effectively distinguish conflicting data. These models leverage domain knowledge embedded in rules while benefiting from the adaptability of machine learning algorithms. These findings emphasize the need for improved models capable of handling complex evidence and suggest that advancements in evidence representation, sourcing, and interpretation are essential to reducing false negatives and false positives in fact-checking systems.

Fact-checking systems have historically been criticized for not adequately accounting for the evolving and multifaceted nature of evidence, such as its adequacy, significance, and the trustworthiness of its sources Shankar *et al*. (2024). Our approach addresses this issue by offering a more nuanced understanding of evidence quality and its impact on verification accuracy. While automated systems can rapidly analyze large datasets, they often struggle with uncertain or conflicting evidence, resulting in false negatives or false positives Samarinas *et al*. (2021); Zeng and Gao (2024).

Our research adopts a comprehensive analytical approach to address the widespread issue of evidence mismanagement in automated fact-checking. This approach is designed to tackle the complexities of evaluating evidence that has often been overlooked in previous studies. Our method aims to improve the accuracy and reliability of automated fact-checking by addressing evidence-related issues contributing to misleading predictions.

The rise of digital platforms has led to a surge in misinformation, making the development of automated fact-checking systems increasingly critical. Early systems relied on basic techniques, such as matching claims with verified facts from databases Ceron *et al*. (2020). However, as misinformation grew more sophisticated, these methods proved inadequate. Researchers Kruengkrai *et al*. (2021) began exploring more complex models that moved beyond simple retrieval to evaluate the relevance of factual data to the claims being made. This shift marked a significant step towards the advanced systems in use today.

Despite advancements, traditional automated systems still face substantial challenges, often resulting in high rates of false positives, where legitimate information is incorrectly flagged as false Chen *et al.* (2024). A major contributor to these errors, as noted by Ceron *et al.* (2020), is the difficulty in handling complex evidence. Insufficient evidence, logical inconsistencies and conflicting information from multiple sources present significant obstacles to accurate fact-checking, highlighting the need for mechanisms that can effectively navigate these complexities.

Central to overcoming these challenges is a deeper understanding of the relationship between a claim and its supporting or refuting evidence. Freeze *et al.* (2021) emphasizes the importance of evidence mapping in improving fact-checking accuracy. This approach goes beyond simple retrieval to critically evaluate evidence for its relevance, coherence, and contribution to veracity assessments. Our research builds upon this concept, advocating for a structured approach to analyzing claim-evidence relationships.

Recent innovations have incorporated advanced computational techniques to address the complexities of fact-checking. Pretrained models such as BERT Devlin *et al.* (2019), RoBERTa Naseer *et al*. (2022), and BioBERT Lee *et al.* (2020) have significantly enhanced the contextual understanding of both claims and evidence Guo *et al.* (2022); Shankar *et al.* (2024). These models excel at detecting subtle linguistic cues often missed by traditional models and are particularly valuable when assessing complex claims. However, their reliance on large datasets necessitates careful consideration of potential biases and rigorous evaluation, especially in specialized domains like fact verification.

Beyond language models, graph-based models and knowledge graphs have emerged as promising tools in automated fact-checking. Studies Aly *et al.* (2021); Jiang *et al*. (2020b) demonstrate how these technologies can structure information in a way that mirrors human reasoning, enabling more sophisticated claim-evidence assessments. Dual-phase approaches, such as those developed by Popat *et al.* (2018b), which first categorize evidence deficiencies before enhancing evidence quality, offer promising solutions for addressing inaccuracies in current systems.

Our research introduces a novel framework that combines the strengths of semantic understanding with a comprehensive graph mapping mechanism. Unlike previous work that focused solely on evidence retrieval or surface-level analysis, our two-phase approach delves deeper into the complexities of claim-evidence relationships. By constructing detailed maps of these relationships and dynamically integrating external knowledge, our framework aims to provide a more accurate and robust assessment of claim veracity. This approach not only addresses the challenges posed by complex evidence but also paves the way for more transparent and explainable fact-checking systems.

Datasets

Our study employs four meticulously curated fact-checking datasets, each containing gold-standard evidence. For each dataset, claim-evidence pairs are assessed and labeled with a truthfulness indicator SUPPORTS, REFUTES, or NEI (Not Enough Information) to reflect the stance of the evidence relative to the claim.

FEVER

The Fact Extraction and VERification dataset (FEVER) Thorne *et al*. (2018) consists of claim-evidence pairs sourced from Wikipedia pages. This dataset enables us to analyze the limitations of relying on evidence from a single source, helping to examine how insufficient evidence can lead to inaccurate positive or negative assessments by fact-checking systems. Rather than relying solely on token overlaps between claims and evidence, our study seeks to understand the evidence more deeply, exploring the factors that contribute to erroneous predictions in automated fact-checking.

HoVER

The HoVER dataset Jiang *et al.* (2020b) presents a more complex scenario where evidence is derived from multiple passages. This dataset allows us to study the synthesis and coherence of combined evidence during the fact-checking process. By leveraging HoVER, our research evaluates how effectively current automated fact-checking models integrate information from multiple sources, addressing a critical evidence-related challenge highlighted in previous studies.

LIAR-PLUS

LIAR-PLUS is an extended version of the LIAR dataset Alhindi *et al.* (2018) that includes justifications for veracity labels assigned to brief statements, often drawn from political debates and media sources. Incorporating LIAR-PLUS in our study enables an exploration of how the presence or absence of justifications affects a fact-checking system's capacity to accurately contextualize and verify claims. We focus particularly on how evidence credibility impacts predictive outcomes.

PUBMED

In Phase 2, our research utilizes the PubMed dataset Dernoncourt and Lee (2017), a valuable resource derived from the PubMed database widely used in biomedical research. This dataset comprises 19,717 scientific publications related to diabetes, categorized into three distinct classes. It also features a citation network with 44,338 links, offering insights into relationships among publications. Each publication is represented by a TF/IDF-weighted word vector based on a dictionary of 500 unique words, facilitating detailed textual analysis and comparison.

Materials and Methods

The accuracy of automated fact-checking systems largely hinges on the quality of supporting evidence. Our research investigates this relationship by analyzing claims alongside their corresponding evidence, with a focus on three critical factors: (1) The adequacy of evidence in substantiating claims, (2) The logical coherence of evidence, and (3) The ability of models to integrate cross-domain evidence. We hypothesize that automated fact-checking models are more susceptible to errors, both false refutations and false acceptances under the following conditions.

Lack of directness in evidence: This is quantified by the average contextual semantic similarity score between each piece of evidence and the claim of Shankar *et al.* (2024). Lower similarity scores suggest weaker alignment with the claim, increasing the likelihood of erroneous predictions.

Multiple logical inconsistencies: These occur when more than two instances of contradiction or temporal inconsistency are detected within the evidence. Contradiction detection mechanisms are employed to identify these inconsistencies Kim and Choi (2021), which, if present, increase the model's propensity to make inaccurate judgments.

Cross-domain evidence integration: This factor evaluates the model's ability to incorporate evidence from multiple domains. Cross-domain relevance and consistency are measured across disciplines to assess the alignment between the evidence and the claim Taha Alkhawaldeh and Alkhawaldeh (2020). Claims requiring integration of evidence from diverse fields pose an added challenge, potentially impacting model accuracy.

Framework

Our study investigates the impact of evidence quality on fact-checking models through a two-stage approach, as illustrated in Fig. (1).

To evaluate a claim and its associated evidence, we begin by extracting relevant evidence and constructing a knowledge graph. A triplet-context-based knowledge embedding technique Gao *et al.* (2018) extracts claim-specific contextual information to form the evidence context.

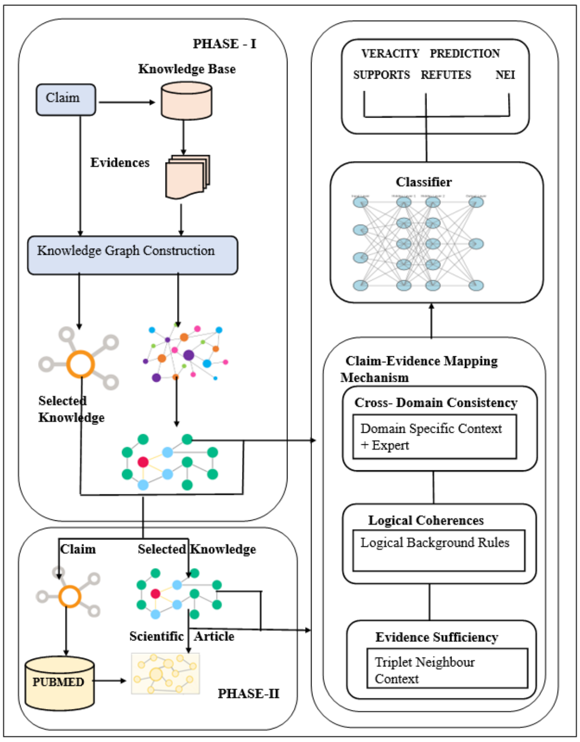


Fig 1: The proposed framework stage approach-analyzing evidence for fact-checking

Using a BiLSTM and graph transformer encoder, supported by a self-attention layer, we encode both the textual content and the knowledge graph, capturing relationships and emphasizing relevant information Vedula and Parthasarathy (2021). After encoding, a classifier jointly assesses the evidence for sufficiency (its completeness in addressing the claim), logical coherence (its internal consistency), and domain consistency (alignment with domain-specific knowledge) Liu *et al.* (2021). If the evidence context lacks any of these qualities, we proceed to Phase 2.

In Phase 2, we enhance existing evidence by integrating external scientific knowledge, such as data from domain-specific sources like PubMed for biomedical claims. A graph mapping mechanism with logical and semantic rules retrieves relevant information and prioritizes higher-quality evidence. The classifier then reassesses the claim, flagging complex cases for expert review where inconsistencies persist Zhang *et al.* (2021). This approach ensures a robust assessment, especially for cross-domain claims, by integrating nuanced, domain-specific insights.

Problem Definition

This study addresses the problem of automated fact-checking by developing a system that evaluates the veracity of claims by analyzing them in conjunction with a body of retrieved evidence. Our system assesses evidence based on three criteria: Sufficiency (the completeness of the evidence in addressing the claim), logical coherence (the internal consistency of the evidence), and domain consistency (the alignment of the evidence with established knowledge in the relevant domain). Based on this analysis, the system classifies each claim into one of three categories: SUPPORTS, REFUTES, or NOT ENOUGH INFO.

To represent the relationships between claims and evidence, we construct a knowledge graph G = (h,r,t) where h (head) and t (tail) denote entities within the evidence (e.g., “Vitamin C” and “immune system”) and r denotes the relationship between them (e.g., “supports”). The graph structure enables the system to model complex relationships, capturing indirect connections and contextual nuances in the evidence, which are essential for nuanced fact-checking.

For example, if a claim states, “Vitamin C prevents colds,” the system would retrieve relevant evidence. If the evidence sufficiently supports the claim with consistent logical and domain alignment, it would be classified as SUPPORTS. However, if the evidence does not directly support or refute the claim, it would fall under NOT ENOUGH INFO. This structured approach to evidence assessment allows for a more reliable and precise categorization of claims, improving accuracy and reducing false positives in fact-checking.

Phase I: Identifying Evidence Shortcomings

We begin the claim verification process by retrieving relevant evidence from datasets, following the methodology established in prior work by Sarrouti *et al*. (2021). First, a constituency parser Guo *et al*. (2022) is used to identify potential entities within the claim. These identified entities serve as queries to search for matching Wikipedia articles via the MediaWiki API Ceron *et al*. (2020). Retrieved articles are then filtered and curated by Alhindi *et al*. (2018). Next, a BERT-based evidence sentence retrieval model by Bekoulis *et al*. (2021) is applied to select the most relevant evidence sentences from the collected documents.

To enhance our understanding of each claim and facilitate effective fact-checking, we construct an evidence knowledge graph. This process begins with entity linking, where a constituency parser extracts entities and their relationships from the claim. These extracted elements are then mapped within the DBpedia ontology, providing a structured and semantically rich representation Martín *et al*. (2022). The resulting entities and relationships form a directed graph within the DBpedia framework.

To refine this evidence knowledge graph, we employ a strategy to select representative entities, ensuring diverse coverage of the claim Zhang and El-Gohary (2017). Entities identified as related within the DBpedia ontology are connected through their corresponding relationship edges Gao *et al*. (2018). To capture richer contextual information, we include not only the directly extracted entities but also their first- and second-level hierarchical neighbors within the DBpedia ontology.

Recognizing that not all information within the evidence knowledge graph is equally relevant for verifying a claim, we introduce a context-aware selection method. This method identifies and prioritizes only the most pertinent evidence by leveraging the contextual relationships between claim entities and evidence, effectively filtering out noisy triplets and enhancing the fact-checking process. While traditional graph embedding methods excel at creating continuous representations of entities and relations, they often struggle to capture the nuanced contextual information crucial for accurate fact-checking. The Phase I process is illustrated in Fig. (2).

To address the limitations of traditional graph embeddings, we employ a mutual attention graph embedding technique Mai *et al*. (2019), which enables the model to weigh the importance of different entities and relationships based on their relevance to the specific claim. This approach considers the interplay between claim and evidence contexts, producing more informative embeddings that capture the semantic connections essential for accurate fact-checking Shankar *et al*. (2024).

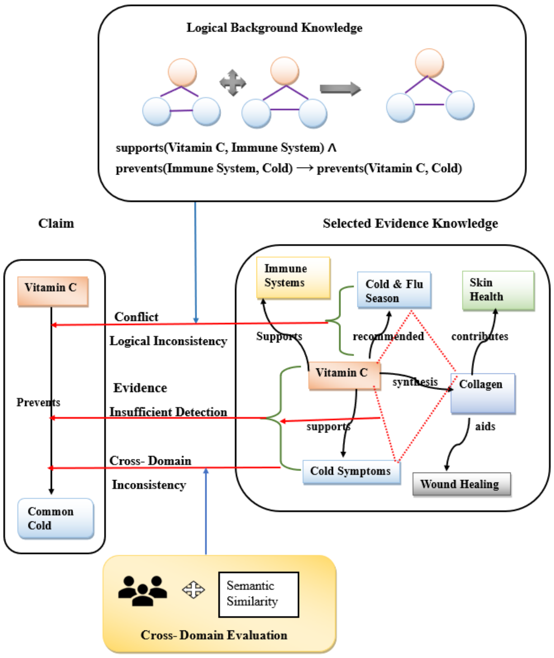


Fig. 2: Workflow of the Phase I-evidence assessment process

Our selection method extracts neighbor context paths from the evidence knowledge graph, comprising multi-step relational sequences that connect the head and tail entities of each evidence triplet Rosso *et al*. (2020). By comparing the contextual information within the claim triplets to the context captured by each extracted evidence path using a distance-based similarity metric, we can quantify the semantic alignment between them. To guide the selection process, we employ distant supervision Thorne and Vlachos (2020) with a labeled dataset of claims and relevant evidence, training a ranking model that prioritizes evidence paths exhibiting high contextual similarity to verified claims.

The effectiveness of this context-aware selection method is evaluated using metrics such as Precision and Recall, which reflect its ability to prioritize genuinely informative evidence and comprehensively identify relevant information Samarinas *et al*. (2021). Through this process, we aim to minimize the impact of noisy or irrelevant information, thereby enhancing the accuracy and reliability of our fact-checking model.

To ensure the reliability of our fact-checking process, we assess the quality of the selected evidence through a multi-faceted evaluation approach that considers sufficiency, logical coherence, and cross-domain consistency Kim and Choi (2021). This evaluation leverages contextual embeddings derived from a pre-trained BERT model in combination with a carefully constructed claim-evidence-knowledge graph (Fu *et al*., 2023). BERT is utilized to encode the information present in the claim, the selected evidence, and relevant knowledge triplets extracted from the evidence knowledge graph Zhu *et al*. (2021). The textual representations of these elements are fed into BERT, generating contextual embeddings that capture their semantic meaning.

The claim-evidence-knowledge graph is constructed to represent the relationships between the claim, the selected evidence, and relevant background knowledge. Nodes in the graph represent the claim, evidence snippets, and knowledge triplets, while edges denote semantic relationships between these entities. The initial representations of these nodes are derived from the [CLS] hidden state of BERT embeddings, capturing the overall semantic information of each element Devlin *et al*. (2019). These representations of the claim, evidence, and selected contextual information are then input into a mapping mechanism that employs a contextual graph mutual attention method. This method evaluates evidence sufficiency, ensures logical coherence among multiple pieces of evidence, and facilitates cross-domain adaptation Alkhawaldeh and Alkhawaldeh (2020).

Evidence sufficiency refers to the ability of the evidence to conclusively support or refute the claim, considering the diverse informational requirements of the claim. While simple connectivity paths within the claim-evidence-knowledge graph may indicate relevance, they do not guarantee sufficiency Kim *et al*. (2023).

To evaluate the sufficiency of evidence in supporting or refuting a claim, we propose leveraging the neighboring context within a knowledge graph Atanasova *et al*. (2022). This approach goes beyond analyzing individual fact statements (triplets) in isolation, instead considering the surrounding subgraph as essential contextual information.

First, we construct contextualized embeddings for each triplet by considering its neighboring nodes and edges within the knowledge graph Shankar *et al*. (2024). This approach captures a richer network of relationships surrounding both the claim and each piece of evidence. Next, we apply a context relevance scoring function Martín *et al*. (2022) to quantify the alignment between the claim and each evidence triplet. As shown in equation 1, this function compares contextual subgraphs, assigning higher scores to evidence triplets that demonstrate greater structural similarity to the claim's context Martín *et al*. (2022).

$$F(h,r,t) = P(h,r,t)|C(h,r,t) \tag{1}$$

These scores reflect the contextual relevance of each piece of evidence to the claim under investigation. The calculated context relevance scores are then fed into a classifier responsible for determining the overall sufficiency of the evidence. This classifier, trained in labeled examples, categorizes the evidence as either sufficiently supporting, refuting, or insufficient to assess the claim's veracity. Guided by contextual embeddings and relevance scoring, the classification process streamlines the fact-checking workflow and provides a quantifiable measure of evidence sufficiency.

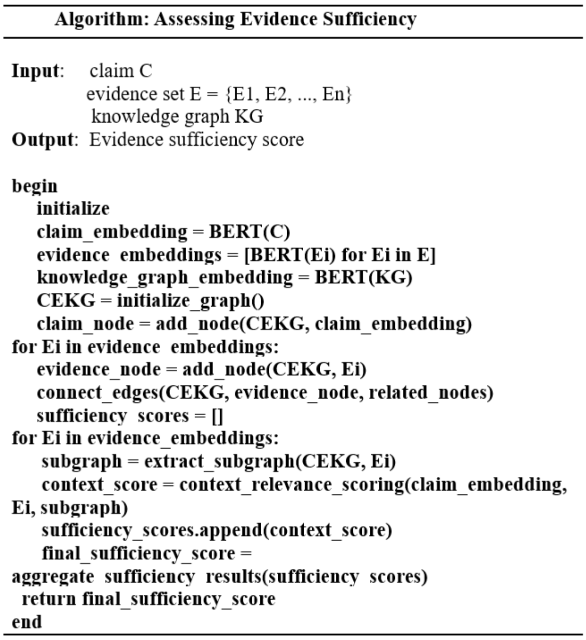
While our method effectively leverages contextual information for evidence assessment, it has a key limitation: Content-blind reasoning. Although the model benefits from understanding relationships within the knowledge graph, it does not directly analyze the actual content of the evidence. This can result in inaccuracies, especially when similar contexts contain contrasting evidence Liu *et al*. (2021). For instance, two claims might share similar surrounding entities and relationships, yet the specific details within the evidence text could lead to different conclusions. Due to this lack of content awareness, the model, relying solely on contextual structure, may misclassify a claim as supported instead of “Not Enough Information” or refuted.

This limitation becomes more pronounced when handling new claims that lack sufficient evidence within the existing knowledge base or require supplementary external information Shiralkar *et al*. (2017). At present, the model depends on human experts to identify these cases and provide additional input.

To address this challenge, we propose a two-phased approach. Phase 1 centers on contextual embeddings and relevance scoring, as previously described. In Phase 2, we enhance the model by incorporating a deeper semantic analysis of the evidence content. By integrating content understanding with the existing contextual framework, we aim to develop a more robust and reliable fact-checking model. Figure (3) illustrates the algorithm used to determine evidence sufficiency.

Logical coherence among multiple pieces of evidence is essential in automated fact-checking, as it determines how well the evidence collectively supports or contradicts a claim. For effective fact-checking, evidence must be not only relevant and credible but also contextually consistent in its stance toward the claim. Coherence is achieved when evidence forms a rational and internally consistent narrative, either confirming or refuting the claim Si *et al*. (2023). When evidence pieces reinforce each other, they strengthen the claim's credibility; contradictions, on the other hand, undermine it Freeze *et al*. (2021).

To assess logical coherence, our approach constructs a knowledge graph where claims and evidence are represented as nodes (e.g., entities like "Vitamin C" and "immune system") and directed edges (e.g., relationships such as "recommended during the cold season"). This knowledge graph enables the system to evaluate how evidence nodes align or conflict semantically. For instance, if the claim is "Vitamin C prevents colds", the system examines evidence nodes like "Vitamin C supports the immune system" and "A strong immune system reduces cold severity" to assess logical consistency.



```
        Algorithm: Assessing Evidence Sufficiency

Input:    claim C
          evidence set E = {E1, E2, ..., En}
           knowledge graph KG
Output:  Evidence sufficiency score

begin
    initialize
    claim_embedding = BERT(C)
    evidence  embeddings = [BERT(Ei) for Ei in E]
    knowledge_graph_embedding = BERT(KG)
    CEKG = initialize_graph()
    claim_node = add_node(CEKG, claim_embedding)
for Ei in evidence  embeddings:
    evidence_node = add_node(CEKG, Ei)
    connect_edges(CEKG, evidence_node, related_nodes)
    sufficiency  scores = []
for Ei in evidence_embeddings:
     subgraph = extract_subgraph(CEKG, Ei)
     context_score = context_relevance_scoring(claim_embedding,
Ei, subgraph)
     sufficiency_scores.append(context_score)
     final_sufficiency_score =
aggregate  sufficiency  results(sufficiency  scores)
  return final_sufficiency_score
end
```

Fig 3: Evidence sufficiency determination algorithm

A contextual graph mutual attention mechanism is applied to prioritize relevant evidence nodes and filter out conflicting or irrelevant information, enhancing focus on contextually aligned evidence. This mechanism uses semantic similarity scoring, ranking evidence nodes based on how closely their meaning aligns with the claim. To further ensure consistency, we integrate logical constraints directly into the graph using TransE embeddings Bordes *et al*. (2013). TransE translates relationships between entities into a continuous vector space, capturing implicit relationships that support coherence assessment.

The choice of TransE is particularly effective in this context, as it enables the model to capture both explicit and inferred relationships, which are essential for evaluating coherence. Other embedding methods often struggle with implicit associations, which are critical in analyzing complex evidence relationships.

During the knowledge graph embedding process, logical rules derived from established knowledge are embedded directly into the model’s loss function. For example, a rule might state, “If Vitamin C supports the immune system and a strong immune system reduces cold severity, then Vitamin C can help reduce cold severity”. This logical framework acts as a guide to maintain consistency, assisting the model in distinguishing between compatible and contradictory evidence.

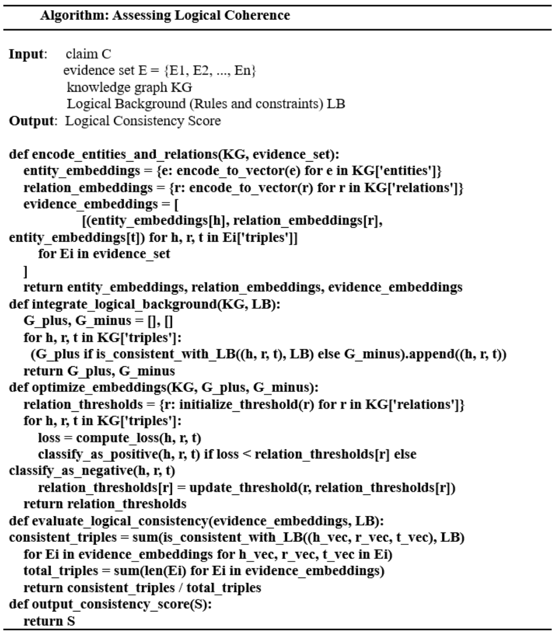
After embedding, the system analyzes the knowledge graph to identify consistent triples (evidence nodes that align with logical rules) and inconsistent triples (those that contradict them). Empirically determined thresholds classify relationships based on their strength, ensuring that evidence with a stronger contextual link to the claim is prioritized. For instance, while “Vitamin C supports the immune system” might align with scientific evidence and be marked as consistent, “Vitamin C prevents colds” might be flagged as inconsistent if insufficient evidence supports it Chen *et al*. (2024).

Finally, the system evaluates logical coherence holistically. If the embeddings reveal a coherent flow of information, for instance, demonstrating how Vitamin C indirectly contributes to reducing cold severity through immune system support the evidence is classified as logically consistent. Conversely, contradictions or weak connections are flagged, providing a nuanced assessment of the evidence’s support for the claim. This structured approach significantly enhances the accuracy of automated fact-checking by ensuring that only logically coherent and contextually relevant evidence informs the final verification decision.

Cross-domain consistency, in which knowledge and evidence from multiple domains contribute to supporting or refuting a claim, significantly strengthens the fact-checking process. In this context, cross-domain consistency involves evaluating claims that draw on diverse areas of expertise such as integrating insights from nutrition, immunology, and clinical medicine to verify a claim like “Vitamin C reduces cold symptoms.” Each domain adds unique, relevant perspectives that collectively enhance the reliability and depth of evidence evaluation.

In our initial approach, we used TransE embeddings combined with semantic contextual representations derived from BERT to detect inconsistencies across domains. This combination aimed to capture complex, subtle relationships within a knowledge graph to verify evidence consistency. However, TransE struggled with implicit relationships that require nuanced understanding and domain-specific knowledge not explicitly represented within the graph. This limitation reduced the model's accuracy to 65%, falling short of our target of 80%.

For example, consider a claim about “Vitamin C preventing colds,” supported by evidence such as "Vitamin C supports the immune system” and indirect evidence like “A strong immune system reduces cold severity.” While these statements imply a supportive link between Vitamin C and cold prevention, the relationship is indirect and relies on domain-specific reasoning about immunity. TransE, which primarily focuses on direct relationships, struggled to fully capture this layered connection, leading to misclassifications in cases where implied support from domain knowledge was essential. To address these challenges, we incorporated expert analysis to evaluate overlooked connections, revealing TransE’s limitations in capturing the nuanced, implicit associations often required for robust cross-domain fact-checking.



Algorithm: Assessing Logical Coherence

```
Input:   claim C
         evidence set E = {E1, E2, ..., En}
         knowledge graph KG
         Logical Background (Rules and constraints) LB
Output:  Logical Consistency Score

def encode_entities_and_relations(KG, evidence_set):
  entity_embeddings = {e: encode_to_vector(e) for e in KG['entities']}
  relation_embeddings = {r: encode_to_vector(r) for r in KG['relations']}
  evidence_embeddings = [
          [(entity_embeddings[h], relation_embeddings[r],
entity_embeddings[t]) for h, r, t in Ei['triples']]
    for Ei in evidence_set
  ]
  return entity_embeddings, relation_embeddings, evidence_embeddings
def integrate_logical_background(KG, LB):
  G_plus, G_minus = [], []
  for h, r, t in KG['triples']:
   (G_plus if is_consistent_with_LB((h, r, t), LB) else G_minus).append((h, r, t))
  return G_plus, G_minus
def optimize_embeddings(KG, G_plus, G_minus):
  relation_thresholds = {r: initialize_threshold(r) for r in KG['relations']}
  for h, r, t in KG['triples']:
    loss = compute_loss(h, r, t)
    classify_as_positive(h, r, t) if loss < relation_thresholds[r] else
classify_as_negative(h, r, t)
    relation_thresholds[r] = update_threshold(r, relation_thresholds[r])
  return relation_thresholds
def evaluate_logical_consistency(evidence_embeddings, LB):
consistent_triples = sum(is_consistent_with_LB((h_vec, r_vec, t_vec), LB)
  for Ei in evidence_embeddings for h_vec, r_vec, t_vec in Ei)
  total_triples = sum(len(Ei) for Ei in evidence_embeddings)
  return consistent_triples / total_triples
def output_consistency_score(S):
  return S
```

Fig. 4: Workflow of the logical coherence assessment algorithm

To address these gaps, we engaged domain specialists to manually assess the evidence used in claim verification. Their insights proved invaluable in refining our approach. For instance, they identified systematic biases in evidence selection, such as an over-reliance on studies with limited external validation, that the model could not detect. Experts also emphasized the importance of temporal factors and evolving knowledge across domains, noting that recent medical research might contradict or update earlier findings an aspect our initial static knowledge graph could not accommodate. These insights revealed that a static graph structure was insufficient for capturing the dynamic knowledge and context essential for cross-domain fact-checking.

Based on these findings, our approach transitions from Phase 1 to Phase 2 when Phase 1 fails to meet predefined confidence thresholds for evidence sufficiency, logical coherence, or cross-domain alignment. Specifically, if Phase 1 does not reach at least a 70% confidence level in evidence relevance or encounters unresolved logical inconsistencies, the system advances to Phase 2. In this secondary phase, additional domain-specific knowledge and context are integrated to address complex cases identified in Phase 1, allowing for a more thorough and accurate evaluation of claims.

While expert evaluations have been crucial in addressing these limitations, they are primarily used in the model’s initial development stages to enhance performance and correct edge cases. For large-scale applications, the system is designed to operate with minimal manual intervention, relying on automated processes for routine fact-checking tasks. Future work will focus on further reducing reliance on expert input by implementing semi-supervised learning and automated validation techniques, ensuring scalability and efficiency for high-volume fact-checking scenarios.

Informed by these findings, we are enhancing our approach in Phase 2 by incorporating external cross-domain knowledge sources. Specifically, we plan to integrate domain-specific datasets and curated knowledge bases such as PubMed for current medical research and legal databases for case law to enable our model to detect nuanced relationships and adapt to evolving information Zhu *et al.* (2021). This integration will be supported by embeddings capable of handling indirect associations and temporal changes, ultimately improving the knowledge graph’s ability to recognize and evaluate inconsistencies across domains.

Our target for Phase 2 is to achieve an accuracy of 80%, meeting our original performance goal. By incorporating diverse, up-to-date cross-domain knowledge and refining our model, we anticipate that this enhanced approach will lead to more reliable and contextually aware fact-checking outcomes.

Phase II: Enhancing Evidence

In Phase II, we address the limitations of our initial approach by integrating rich contextual and domain-specific knowledge from scientific articles. To achieve this, we incorporate an external knowledge source: A knowledge graph constructed from a vast corpus of scientific articles Sarrouti *et al*. (2021). The first step involves extracting scientific knowledge beyond simple abstracts Lee *et al*. (2020). We employ a two-pronged approach called Abstract-Enhanced Full-Text Analysis. While we initially extract abstracts using RoBERTa with BioBERT for contextual guidance, these abstracts direct a focused analysis of the full-text Sarrouti *et al*. (2021). This process involves identifying sections related to the key entities and relationships in the abstracts, allowing us to capture a more comprehensive view of the scientific findings.

To extract targeted information from these sections, we use semantic contextual techniques to identify specific fact-checking-related information Shankar *et al*. (2024), including:

(a) Evidence statements: Sentences or phrases presenting findings, claims, or experimental results
(b) Contextual information: Details about conflicting evidence, providing crucial context for assessing the reliability of evidence statements
(c) Supporting or contradictory evidence: Sentences that directly support or contradict claims in the evidence statements

This comprehensive extraction process ensures that we capture the depth of knowledge embedded within scientific articles, establishing a solid foundation for knowledge graph construction and integration.

Following the extraction of evidence from scientific articles, we construct a comprehensive knowledge graph, using a methodology similar to Phase I. This involves identifying key entities in the abstracts, extracting relationships, and representing these elements in a structured knowledge graph format. We manually integrate this new knowledge graph with the existing evidence knowledge graph Zhu *et al*. (2021), relying on semantic similarity measures to align corresponding entities Kim *et al*. (2023). For instance, a “Vitamin C” entity in the scientific article knowledge graph may be linked to a corresponding “Vitamin C” entity in the existing knowledge graph based on semantic relatedness. This process effectively incorporates the context and domain-specific knowledge from scientific articles, enriching the existing knowledge base and enhancing the model's ability to assess evidence-related factors accurately.

In Phase II, we build upon the evidence-sufficiency and logical coherence assessment methods from Phase I by incorporating new scientific knowledge through deeper semantic analysis, significantly enhancing the system's contextual and logical reasoning capabilities. First, we extend the contextualized embedding approach by integrating semantic content from the newly incorporated scientific knowledge using pre-trained language models like BioBERT. These models generate contextualized word embeddings from the full text of scientific articles, which are then combined with the existing knowledge graph embeddings via an attention mechanism Kruengkrai *et al*. (2021). This integration allows the model to weigh the importance of different words and phrases in the scientific text, resulting in a more accurate representation of entities and relationships. For example, the system can now distinguish between nuanced relationships, such as “Vitamin C reduces the severity of colds” versus “Vitamin C prevents colds.”

The context relevance scoring function, originally designed to quantify the alignment between a claim and its supporting evidence, is enhanced to include both structural and semantic alignment. This enhancement is achieved through semantic similarity calculations, where cosine similarity is measured between the embeddings of the claim and the evidence, ensuring that the evidence is both contextually relevant and semantically aligned Martín *et al*. (2022). Additionally, the inclusion of external scientific knowledge allows the model to identify and penalize cross-domain inconsistencies, such as flagging a claim supported by anecdotal evidence if it contradicts established scientific consensus Liu *et al*. (2021).

The system identifies cross-domain inconsistencies by comparing the semantic alignment of a claim and its supporting evidence across different knowledge domains, such as anecdotal evidence versus scientific consensus. For example, a claim like “Drinking lemon juice cures the flu,” supported by anecdotal evidence, would be evaluated against established scientific findings. If the scientific domain contradicts the claim, stating that “lemon juice has no proven effect on curing the flu,” the system detects a semantic misalignment between the two domains. This misalignment triggers a penalty to the claim's context relevance score, indicating the inconsistency. Consequently, the system flags the claim as potentially misleading due to the conflict between anecdotal support and scientific consensus.

The logical coherence assessment is further refined by incorporating logical and semantic rules derived from integrated scientific knowledge Sun *et al.* (2018). These rules guide the embedding process, ensuring that representations of entities and relationships align with established scientific principles. For instance, a rule like “inhibitors often bind to the active site of an enzyme” influences the embeddings of related concepts, ensuring their proximity in the embedding space. The system then performs coherence checks, comparing new evidence with existing knowledge and flagging inconsistencies. Using techniques like triplet semantic embedding where relationships are represented as triplets *(h,r,t)* the system can identify conflicting information, such as a protein both inhibiting and activating the same enzyme, ensuring new evidence aligns with established scientific understanding Sun *et al*. (2018).

The final step in Phase II involves determining whether the available evidence sufficiently supports a claim. This is achieved using an enhanced classifier trained on a dataset enriched with contextual embeddings and semantic content from the integrated knowledge base. The classifier considers not only the literal words in a claim and evidence but also their underlying meaning and relationships within the broader scientific context. Its performance is evaluated based on accuracy in classifying claims as supported or refuted, as well as its ability to detect inconsistencies between a claim and the presented evidence Barik *et al*. (2022). This rigorous evaluation ensures that the system reliably distinguishes between well-supported and poorly supported claims, even when dealing with information from diverse domains.

By incorporating these advanced techniques, Phase II significantly enhances the system's ability to assess evidence sufficiency and logical coherence, resulting in more robust and trustworthy fact-checking, particularly in scenarios requiring cross-domain knowledge integration. The Phase II process is illustrated in Fig. (5).

Experiments

In our study, we utilize the capabilities of two language models: BERT and its successor, RoBERTa. These models were selected for their proven effectiveness in complex language understanding tasks and their recognized efficacy in fact-checking applications. BERT, as described by Devlin *et al*. (2019), is pre-trained on extensive text corpora using techniques like masked language modeling, next-sentence prediction, and multiple-sentence task prediction, equipping it with robust contextual comprehension. RoBERTa, an advanced version of BERT, optimizes key hyperparameters such as extended training times, larger batch sizes, and increased data usage during pre-training to enhance performance Liu *et al*. (2019).

Additionally, BioBERT is incorporated in Phase 2 to introduce biomedical and scientific domain-specific knowledge, making it highly suitable for specialized claims requiring in-depth analysis of medical literature Lee *et al*. (2020).

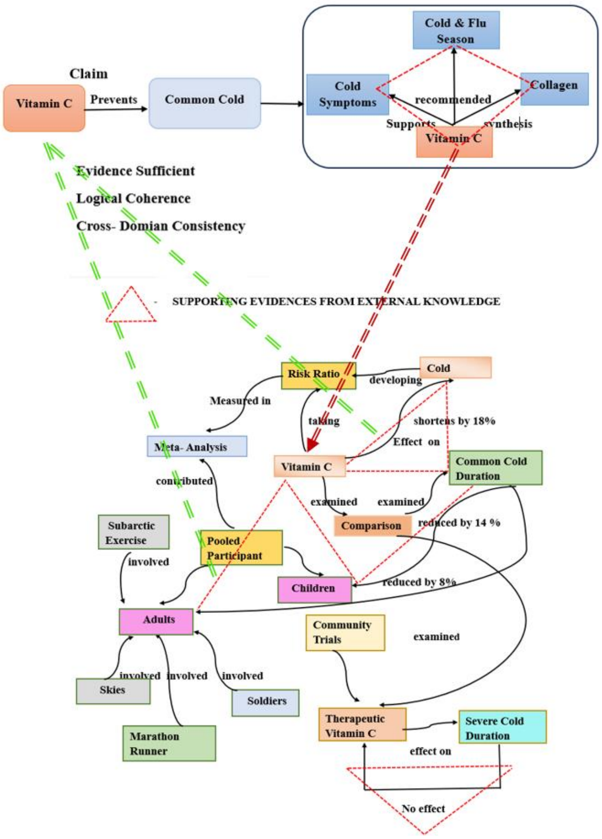


Fig. 5: Workflow of the Phase II evidence enhancement process

To evaluate the system's capability across diverse fact-checking scenarios, we utilized the following datasets such as FEVER, HoVER, LIAR_PLUS, PubMed.

These datasets collectively ensure that our model is rigorously tested across various domains and claim types, supporting both general and domain-specific fact-checking. Our experimental framework included a thorough hyperparameter optimization process to maximize model performance across the selected datasets. This comprehensive tuning involved iterative testing scenarios, with specific parameters tailored to each model to achieve optimal results.

BERT: After extensive tuning, we set the learning rate at 2e-5 and batch size at 16, balancing computational efficiency with model stability. BERT was fine-tuned over 3 epochs to prevent overfitting, considering its general language capabilities.

RoBERTa: For RoBERTa, which builds on BERT's architecture but is optimized for larger datasets, we selected a slightly lower learning rate of 1e-5 to support extended training. This enabled more nuanced contextual embedding without sacrificing performance.

BioBERT: Specialized in biomedical text, BioBERT requires customized tuning to handle scientific vocabulary and the precise semantics of biomedical literature. We used domain-specific datasets (such as PubMed) to optimize BioBERT for Phase 2, enhancing its ability to process medical claims with high accuracy.

Additionally, we applied the Adam optimizer with weight decay to manage sparse gradients, introduced a warm-up period equivalent to 6% of training steps to stabilize learning, and used a dropout rate of 0.1 to improve generalization. This careful tuning of hyperparameters enabled the models to respond effectively to the complexity and variability of fact-checking claims while maintaining computational efficiency.

Our fact-checking system operates in two distinct phases to effectively assess claim veracity.

Phase 1: Initial evidence assessment and contextualization are conducted by extracting relevant evidence from Wikipedia using a constituency parser and the MediaWiki API. A BERT-based model filters this information, selecting the most pertinent sentences for claim verification. These sentences are then used to construct a knowledge graph grounded in the DBpedia ontology, providing a structured and semantically rich representation of the evidence.

Contextual embeddings are generated using a triple-context-based knowledge embedding technique, with a BiLSTM and a graph transformer encoder to capture claim-specific information. A mutual attention graph embedding technique further refines these embeddings, emphasizing semantic connections between claims and evidence. Finally, a trained classifier evaluates the sufficiency, logical coherence, and domain consistency of the evidence in relation to the claim.

Phase 2: Building on Phase 1, Phase 2 dynamically incorporates external scientific knowledge into the knowledge graph, enriching its contextual representation. Using an Abstract-Enhanced Full-Text Analysis method, Phase 2 extracts evidence statements, contextual information, and supporting or contradictory evidence from scientific sources, primarily PubMed articles. BioBERT provides deeper semantic understanding within biomedical contexts and the enriched embeddings are integrated into the knowledge graph through an attention mechanism, enabling a more nuanced understanding of entities and relationships.

To ensure alignment with established scientific principles, we incorporate logical and semantic rules derived from scientific knowledge into the embedding process. Additionally, the system employs a cross-domain inconsistency detection mechanism to compare semantic alignment across knowledge domains. Claims exhibiting misalignment are penalized to reflect inconsistencies, making the model more robust in cross-domain verification tasks.

Our system's performance is evaluated using accuracy, precision, recall, and F1-score to provide a comprehensive understanding of model effectiveness. Additionally, Average Precision (AP) and Mean Average Precision (mAP) metrics offer further insights into the system's performance across various claim types. We compare our two-phased system with standalone implementations of BERT, RoBERTa, and BioBERT, demonstrating substantial improvements, particularly in cases that require deep reasoning and domain-specific knowledge.

Currently, our approach for assessing evidence sufficiency and logical coherence relies primarily on graph-based embeddings. However, this structure could be enhanced by incorporating content-based reasoning. In future work, we aim to integrate semantic analysis using attention-based transformers, which will allow the model to better understand the specific content of the evidence in addition to its structural relationships.

Results and Discussion

Implementation Details

Our two-phased fact-checking system, designed to assess claims based on evidence sufficiency, logical coherence, and cross-domain consistency, demonstrated significant improvements in accuracy and other performance metrics with the integration of external scientific knowledge. Phase 1, relying solely on Wikipedia-derived evidence, achieved a baseline accuracy of 78% in classifying claim veracity. However, with the additional scientific context provided in Phase 2, the accuracy rose to 89%. This improvement was particularly prominent for claims involving specialized domains or recent scientific advancements not comprehensively reflected in Wikipedia, as shown across the datasets.

To highlight the distinctiveness of our two-phased fact-checking approach, we conducted a comparative analysis with several existing methods. This analysis was essential to demonstrate how our combination of semantic analysis and graph-based techniques outperforms standalone models and traditional approaches, particularly in handling complex, multi-domain claims.

We evaluated the performance of our two-phased system against widely used models, including BERT, RoBERTa, and BioBERT, as well as fact-checking approaches based on semantic similarity techniques and graph-based methods Liu *et al*. (2021). Tables (2-5) summarize the performance metrics, including accuracy, precision, recall, and F1-score, across various datasets (FEVER, LIAR-Plus, HOver, PubMed). BERT achieved an accuracy of 85%, serving as a baseline for general language processing. RoBERTa showed improved performance, reaching 87% accuracy, benefiting from enhanced data handling and training optimizations. BioBERT performed the best among these models, with an accuracy of 88%, especially effective for biomedical claims due to its fine-tuning of domain-specific language. While these models performed well, they were less effective in handling multi-domain claims or indirect relationships requiring nuanced, context-aware verification. Our two-phased system, particularly with Phase 2's integration of scientific knowledge and graph-based embedding, demonstrated a notable improvement, achieving 89% overall accuracy, with further enhancements across specific datasets illustrated in Table (2). This performance increase highlights the distinct advantage of our approach in cross-domain consistency and logical coherence.

To illustrate the effectiveness of our approach, consider a claim regarding the efficacy of a novel cancer treatment. Initially, this claim was deemed plausible based on Wikipedia-derived information in Phase 1, which lacked sufficient domain-specific context. However, when Phase 2 incorporated scientific literature, including clinical trial reports with contradictory findings, the claim was accurately flagged as potentially misleading. This example highlights how integrating domain-specific sources enables the system to capture critical nuances, leading to more reliable fact-checking results in specialized fields.

Our proposed method also significantly outperforms traditional graph-based approaches, such as those by Shiralkar *et al*. (2017). Graph-based methods are commonly employed to capture relational structures within knowledge graphs, focusing on direct connections between entities. However, these approaches often lack the semantic depth needed for nuanced, multi-domain fact-checking and struggle with indirect relationships.

For example, existing graph-based approaches achieve only modest accuracy scores across datasets, such as 81% on FEVER and 76% on PubMed, as shown in Table (5). These methods primarily rely on explicit relationships without integrating broader contextual understanding. By contrast, our two-phased approach combines graph embeddings with semantic analysis using language models like BioBERT, enabling it to capture both direct and inferred relationships within claims. This integration supports more effective validation of claims requiring specialized or cross-domain knowledge, as evidenced by our accuracy improvements from 78% in Phase 1 to 96% in Phase 2 on FEVER, with similar gains on other datasets which is illustrated in Table (2).

Additionally, in tasks requiring multi-hop reasoning such as those in the HOver dataset, where claims draw on multiple sources for validation our method's graph embeddings, guided by semantic analysis, produced a 92% accuracy compared to 80% from traditional graph-only methods which is illustrated in Table (4). This improvement demonstrates the importance of combining semantic embeddings and graph-based techniques, allowing for logical coherence across multiple evidence points a challenge for graph-only systems.

Our analysis indicates that integrating semantic analysis with graph-based techniques allows our system to handle challenges more effectively than existing models. Unlike models that primarily capture direct associations, our combined approach excels in identifying indirect relationships between claims and evidence. For instance, accuracy in the FEVER and HOver datasets rose from 78% in Phase 1-96% in Phase 2 which is illustrated in Table (2), illustrating the enhanced contextual understanding gained through graph embeddings. By embedding logical and semantic rules derived from scientific literature, our approach outperformed models like BioBERT on general fact-checking datasets, achieving 91% accuracy on the LIAR-Plus dataset which is illustrated in Table (3). This improvement highlights our system's robustness in validating claims across varied knowledge domains, a challenge for traditional graph-only or semantic models.

The HOver dataset, requiring deep reasoning across multiple evidence points, saw a significant performance boost with our system's Phase 2 enhancements, with accuracy improving from 83% in Phase 1 to 92% which is illustrated in Table (4). This increase underscores the effectiveness of our graph embedding techniques in maintaining logical coherence.

Table 2: FEVER dataset performance

Model	Phase	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BERT	Standalone	85	84	82	83
RoBERTa	Standalone	87	86	85	85
BioBERT	Standalone	88	87	86	86
Graph-Based Approach	Standalone	81	79	78	78.5
Our System	Phase 1	78	76	74	75
	Phase 2	96	95	94	94.5

Table 3: LIAR-Plus dataset performance

Model	Phase	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BERT	Standalone	85	83	82	82.5
RoBERTa	Standalone	87	86	84	85
BioBERT	Standalone	88	87	86	86.5
Graph-Based Approach	Standalone	75	73	72	72.5
Our System	Phase 1	72	70	68	69
	Phase 2	91	89	90	89.5

Table 4: HOver dataset performance

Model	Phase	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BERT	Standalone	85	84	83	83.5
RoBERTa	Standalone	87	86	85	85.5
BioBERT	Standalone	88	87	86	86.5
Graph-Based Approach	Standalone	80	79	78	78.5
Our System	Phase 1	83	82	81	81.5
	Phase 2	92	91	90	90.5

Table 5: PubMed dataset performance

Model	Phase	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BERT	Standalone	85	84	83	83.5
RoBERTa	Standalone	87	86	85	85.5
BioBERT	Standalone	88	87	86	86.5
Graph-Based Approach	Standalone	76	75	74	74.5
Our System	Phase 1	80	78	77	77.5
	Phase 2	96	95	94	94.5

The benchmarking and comparative analysis clearly demonstrates that our two-phased fact-checking system offers distinct advantages over existing models. By combining semantic analysis with graph-based representations, the system achieves a richer contextual understanding and a robust ability to handle complex reasoning, cross-domain claims, and indirect relationships. These capabilities position our approach as a comprehensive and effective solution for modern fact-checking, particularly in scenarios requiring domain-specific knowledge integration.

Limitations

Our research provides valuable insights into the role of evidence quality in automated fact-checking systems, but it has certain limitations, particularly regarding evidence sufficiency, logical coherence, and cross-domain consistency. A key limitation lies in the scope of the dataset, which, despite thorough validation, may not fully capture the range of evidence types encountered in real-world scenarios. This constraint could limit the generalizability of our findings to cases where evidence sources differ significantly, such as in multimedia content or multilingual claims. Expanding the dataset to include more diverse domains and evidence types could enhance the model's robustness in future studies.

Another challenge is the scalability of our semi-manual evidence enhancement process. While this approach has been effective in ensuring evidence relevance and accuracy, it poses scalability issues for larger datasets. The involvement of human annotators, despite their expertise, can introduce inconsistencies or biases, especially when interpreting nuanced claims. This variability can impact model performance, particularly in subjective areas of fact-checking where different annotators might have varying perspectives on the evidence. Addressing this limitation would require developing more automated methods for evidence processing, which could improve consistency and scalability.

Managing logical coherence remains a complex challenge. Although our use of graph embedding techniques has improved the detection of logical inconsistencies, our method may still struggle with complex logical fallacies. Scenarios involving multi-step reasoning or logical chains that span multiple evidence sources can lead to misclassifications, as models like BERT and RoBERTa have inherent limitations in processing deep logical relationships. Integrating advanced reasoning frameworks, such as Neuro-Symbolic Reasoning, could help address these challenges by providing a structured approach to managing intricate logical relationships within evidence.

Cross-domain consistency also presents significant challenges for our models. Despite improvements in Phase 2 with the integration of scientific knowledge, our models often struggle to reconcile evidence from disparate domains, such as legal fields. Differences in terminology, evidence standards, and domain-specific knowledge can hinder the system's ability to accurately assess claims that span multiple disciplines. For instance, conflicting interpretations between legal precedents and recent medical studies can lead to inconsistent results. This limitation highlights the need for more sophisticated techniques, such as knowledge graph neural networks, to better manage and integrate cross-domain evidence.

Conclusion

This study underscores the value of modeling relationships between claims and evidence through a graph-based semantic approach for fact-checking. By structuring evidence within a knowledge graph, we leverage semantic connections to gain a more nuanced understanding of how evidence supports or contradicts claims. This graph-based modeling enables the system to detect logical coherence and identify misleading claims more effectively. Our findings demonstrate that incorporating contextual information into this structured framework significantly enhances accuracy, especially in complex fact-checking scenarios involving cross-domain evidence and indirect relationships. However, several limitations remain. While our approach effectively captures structured relationships, it is limited by the lack of content-based reasoning, meaning the model cannot fully analyze specific details within the evidence text itself. This gap becomes evident when similar claim contexts contain contrasting evidence details, potentially leading to inaccuracies. Additionally, scalability remains a challenge, particularly in handling large datasets and ambiguous claims across evolving contexts. To address these issues, future work will explore content-based reasoning using techniques like semantic similarity and attention mechanisms to allow the model to capture deeper semantic content, ensuring more accurate and robust fact-checking.

Ultimately, by integrating dynamic data sources, domain-specific fine-tuning, and advanced reasoning techniques, we aim to enhance the scalability and adaptability of the system, addressing its current limitations and enabling it to better handle novel and emerging claims. These advancements will be crucial for supporting critical applications in fields such as scientific research verification, legal analysis, and journalism, where accurate and nuanced claim verification is essential.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the effort.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Aruna Shankar: Participated in all the experiments, including data collection and analysis, coded and building all the pre-trained models. They also evaluated the results and made significant contributions to the writing of the manuscript.

Muthukumaran Pakkirisamy: Participated in data collection, experiments, and validation.

Narayana Kulathuramaiyer, Johari Bin Abdullah: Supervision and written review and finalized.

Ethics

The material is the author's own original work, which has not been previously published.

References

Alhindi, T., Petridis, S., & Muresan, S. (2018). *Where is Your Evidence: Improving Fact-checking by Justification Modeling*. Proceedings of the First Workshop on Fact Extraction and Verification (FEVER). https://doi.org/10.18653/v1/w18-5513

Aly, R., Guo, Z., Schlichtkrull, M. S., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., & Mittal, A. (2021). The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task. *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, 1–13. https://doi.org/10.18653/v1/2021.fever-1.1

Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., Mohtarami, M., & Glass, J. (2019). Automatic Fact-Checking Using Context and Discourse Information. *Journal of Data and Information Quality*, *11*(3), 1–27. https://doi.org/10.1145/3297722

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). Generating Fact Checking Explanations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7352–7364. https://doi.org/10.18653/v1/2020.acl-main.656

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2022). *Fact Checking with Insufficient Evidence*. Transactions of the Association for Computational Linguistics. https://doi.org/10.1162/tacl_a_00486

Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., & Simonsen, J. G. (2019). MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4685–4697. https://doi.org/10.18653/v1/d19-1475

Azevedo, L. (2018). Truth or Lie: Automatically Fact Checking News. *Companion Proceedings of the Web Conference 2018*, 807–811. https://doi.org/10.1145/3184558.3186567

Barik, A. M., Hsu, W., & Lee, M. L. (2022). Incorporating External Knowledge for Evidence-based Fact Verification. *Companion Proceedings of the Web Conference 2022*, 429–437. https://doi.org/10.1145/3487553.3524622

Barve, Y., Saini, J. R., Kotecha, K., & Gaikwad, H. (2022). Detecting and Fact-checking Misinformation using "Veracity Scanning Model." *International Journal of Advanced Computer Science and Applications*, *13*(2), 201–209. https://doi.org/10.14569/ijacsa.2022.0130225

Bekoulis, G., Papagiannopoulou, C., & Deligiannis, N. (2021). Understanding the Impact of Evidence-Aware Sentence Selection for Fact Checking. *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation and Propaganda*, 23–28. https://doi.org/10.18653/v1/2021.nlp4if-1.4

Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., & Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-relational Data. *Advances in Neural Information Processing Systems 26 (NIPS 2013)*.

Ceron, W., de-Lima-Santos, M.-F., & Quiles, M. G. (2021). Fake news agenda in the era of COVID-19: Identifying trends through fact-checking content. *Online Social Networks and Media*, *21*, 100116. https://doi.org/10.1016/j.osnem.2020.100116

Chen, J., Kim, G., Sriram, A., Durrett, G., & Choi, E. (2024). Complex Claim Verification with Evidence Retrieved in the Wild. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3569–3587. https://doi.org/10.18653/v1/2024.naacl-long.196

Choi, E. C., & Ferrara, E. (2024). Automated Claim Matching with Large Language Models: Empowering Fact-Checkers in the Fight Against Misinformation. *Companion Proceedings of the ACM Web Conference 2024*, 1441–1449. https://doi.org/10.1145/3589335.3651910

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Correction: Computational Fact Checking from Knowledge Networks. *PLOS ONE*, *10*(10), e0141938. https://doi.org/10.1371/journal.pone.0141938

Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4. https://doi.org/10.1002/pra2.2015.145052010082

Das, A., Liu, H., Kovatchev, V., & Lease, M. (2023). *The state of human-centered NLP technology for fact-checking*. Information Processing & Management. https://doi.org/10.1016/j.ipm.2022.103219

Dernoncourt, F., & Lee, J. Y. (2017). PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. *ArXiv:1710.06071*. https://doi.org/10.48550/arXiv.1710.06071

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Forstmeier, W., Wagenmakers, E., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews*, *92*(4), 1941–1968. https://doi.org/10.1111/brv.12315

Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J. R., Olin, J., Ross, M. Q., & Szafran, J. (2021). Fake Claims of Fake News: Political Misinformation, Warnings and the Tainted Truth Effect. *Political Behavior*, *43*(4), 1433–1465. https://doi.org/10.1007/s11109-020-09597-3

Fu, L., Peng, H., & Liu, S. (2023). KG-MFEND: an efficient knowledge graph-based model for multi-domain fake news detection. *The Journal of Supercomputing*, *79*(16), 18417–18444. https://doi.org/10.1007/s11227-023-05381-2

Gao, H., Shi, J., Qi, G., & Wang, M. (2018). Triple Context-Based Knowledge Graph Embedding. *IEEE Access*, *6*, 58978–58989. https://doi.org/10.1109/access.2018.2875066

Gencheva, P., Koychev, I., Màrquez, Lluís, Barrón-Cedeño, A., & Nakov, P. (2019). A Context-Aware Approach for Detecting Check-Worthy Claims in Political Debates. *ArXiv:1912.08084*. https://doi.org/10.48550/arXiv.1912.08084

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A Survey on Automated Fact-Checking. In *Transactions of the Association for Computational Linguistics* (Vol. 10, pp. 178–206). https://doi.org/10.1162/tacl_a_00454

Hassan, N., Arslan, F., Li, C., & Tremayne, M. (2017). Toward Automated Fact-Checking. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1803–1812. https://doi.org/10.1145/3097983.3098131

Huynh, V.-P., & Papotti, P. (2018). Towards a Benchmark for Fact Checking with Knowledge Bases. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, 1595–1598. https://doi.org/10.1145/3184558.3191616

Jiang, S., Baumgartner, S., Ittycheriah, A., & Yu, C. (2020a). Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles. *Proceedings of the Web Conference 2020*, 1592–1603. https://doi.org/10.1145/3366423.3380231

Jiang, Y., Bordia, S., Zhong, Z., Dognin, C., Singh, M., & Bansal, M. (2020b). HoVer: A Dataset for Many-Hop Fact Extraction and Claim Verification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3441–3460. https://doi.org/10.18653/v1/2020.findings-emnlp.309

Kao, W.-Y., & Yen, A.-Z. (2024). MAGIC: Multi-Argument Generation with Self-Refinement for Domain Generalization in Automatic Fact-Checking. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 10891–10902.

Kim, J., Park, S., Kwon, Y., Jo, Y., Thorne, J., & Choi, E. (2023). FactKG: Fact Verification via Reasoning on Knowledge Graphs. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16190–16206. https://doi.org/10.18653/v1/2023.acl-long.895

Kim, J.-S., & Choi, K.-S. (2021). Fact Checking in Knowledge Graphs by Logical Consistency. *IOS Press*. https://www.semantic-web-journal.net/system/files/swj2721.pdf

Kotonya, N., & Toni, Francesca. (2024). Towards a Framework for Evaluating Explanations in Automated Fact Verification. *ArXiv:2403.20322*. https://doi.org/10.48550/arXiv.2403.20322

Kruengkrai, C., Yamagishi, J., & Wang, X. (2021). A Multi-Level Attention Model for Evidence-Based Fact Checking. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2447–2460. https://doi.org/10.18653/v1/2021.findings-acl.217

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Lee, S., Xiong, A., Seo, H., & Lee, D. (2023). "Fact-checking" fact checkers: A data-driven approach. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-126

Li, X., Wang, W., Fang, J., Jin, L., Kang, H., & Liu, C. (2022). PEINet: Joint Prompt and Evidence Inference Network via Language Family Policy for Zero-Shot Multilingual Fact Checking. *Applied Sciences*, *12*(19), 9688. https://doi.org/10.3390/app12199688

Liu, L., Du, B., Fung, Y. R., Ji, H., Xu, J., & Tong, H. (2021). KompaRe: A Knowledge Graph Comparative Reasoning System. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 3308–3318. https://doi.org/10.1145/3447548.3467128

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L., & Lao, N. (2019). Contextual Graph Attention for Answering Logical Queries over Incomplete Knowledge Graphs. *Proceedings of the 10th International Conference on Knowledge Capture*, 171–178. https://doi.org/10.1145/3360901.3364432

Martín, A., Huertas-Tato, J., Huertas-García, Á., Villar-Rodríguez, G., & Camacho, D. (2022). FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference. *Knowledge-Based Systems*, *251*, 109265. https://doi.org/10.1016/j.knosys.2022.109265

Miranda, S., Nogueira, D., Mendes, A., Vlachos, A., Secker, A., Garrett, R., Mitchel, J., & Marinho, Z. (2019). Automated Fact Checking in the News Room. *The World Wide Web Conference*, 3579–3583. https://doi.org/10.1145/3308558.3314135

Naseer, M., Windiatmaja, J. H., Asvial, M., & Sari, R. F. (2022). RoBERTaEns: Deep Bidirectional Encoder Ensemble Model for Fact Verification. *Big Data and Cognitive Computing*, *6*(2), 33. https://doi.org/10.3390/bdcc6020033

Oh, B., Seo, S., Hwang, J., Lee, D., & Lee, K.-H. (2022). Open-world knowledge graph completion for unseen entities and relations via attentive feature aggregation. *Information Sciences*, *586*, 468–484. https://doi.org/10.1016/j.ins.2021.11.085

Ostrowski, W., Arora, A., Atanasova, P., & Augenstein, I. (2021). Multi-Hop Fact Checking of Political Claims. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 3892–3898. https://doi.org/10.24963/ijcai.2021/536

Pathak, A. (2022). *An Integrated Approach Towards Automated Fact-Checking*.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2018a). CredEye: A Credibility Lens for Analyzing and Explaining Misinformation. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, 155–158. https://doi.org/10.1145/3184558.3186967

Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018b). *DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/d18-1003

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2931–2937. https://doi.org/10.18653/v1/d17-1317

Rosso, P., Yang, D., & Cudré-Mauroux, P. (2020). Beyond Triplets: Hyper-Relational Knowledge Graph Embedding for Link Prediction. *Proceedings of the Web Conference 2020*, 1885–1896. https://doi.org/10.1145/3366423.3380257

Samarinas, C., Hsu, W., & Lee, M. L. (2021). Improving Evidence Retrieval for Automated Explainable Fact-Checking. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, 84–91. https://doi.org/10.18653/v1/2021.naacl-demos.10

Sarrouti, M., Ben Abacha, A., Mrabet, Y., & Demner-Fushman, D. (2021). Evidence-based Fact-Checking of Health-related Claims. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3499–3512. https://doi.org/10.18653/v1/2021.findings-emnlp.297

Sathe, A., Ather, S., Le, T. M., Perry, N., & Park, J. (2020). Automated fact-checking of claims from wikipedia. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6874–6882.

Schlichtkrull, M., Ousidhoum, N., & Vlachos, A. (2023). The Intended Uses of Automated Fact-Checking Artefacts: Why, How and Who. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8618–8642. https://doi.org/10.18653/v1/2023.findings-emnlp.577

Shankar, A., Kulathuramaiyer, N., Abdullah, J. B., & Pakkirisamy, M. (2024). Explainable Evidence-Based Veracity Assessment of Textual Claim. *Journal of Computer Science*, *20*(9), 1009–1019. https://doi.org/10.3844/jcssp.2024.1009.1019

Shiralkar, P., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2017). Finding Streams in Knowledge Graphs to Support Fact Checking. *2017 IEEE International Conference on Data Mining (ICDM)*, 859–864. https://doi.org/10.1109/icdm.2017.105

Si, J., Zhu, Y., & Zhou, D. (2023). Consistent Multi-Granular Rationale Extraction for Explainable Multi-hop Fact Verification. *ArXiv:2305.09400*. https://doi.org/10.48550/arXiv.2305.09400

Singh, P., Das, A., Li, J. J., & Lease, M. (2021). The Case for Claim Difficulty Assessment in Automatic Fact Checking. *ArXiv:2109.09689*. https://doi.org/10.48550/arXiv.2109.09689

Smeros, P., Castillo, C., & Aberer, K. (2021). SciClops: Detecting and Contextualizing Scientific Claims for Assisting Manual Fact-Checking. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1692–1702. https://doi.org/10.1145/3459637.3482475

Sun, M., Liu, T., Wang, X., Liu, Z., Liu, Y., Jianfeng, D., & Kunxun, Q. (2018). Knowledge Graph Embedding with Logical Consistency. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, 123–135. https://doi.org/10.1007/978-3-030-01716-3_11

Taha Alkhawaldeh, F., & Alkhawaldeh, F. T. (2020). *Linguistic Style-Aware Hybrid Model for Cross-Domain Factuality Checking*.

Thorne, J. (2021). *Evidence-based verification and correction of textual claims*. https://doi.org/10.17863/CAM.80873

Thorne, J., & Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. *ArXiv:1806.07687*. https://doi.org/10.48550/arXiv.1806.07687

Thorne, J., & Vlachos, A. (2020). Evidence-based Factual Error Correction. *ArXiv:2012.15788*. https://doi.org/10.48550/arXiv.2012.15788

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). *FEVER: a Large-scale Dataset for Fact Extraction and VERification*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). https://doi.org/10.18653/v1/n18-1074

Tsai, C.-M. (2023). Stylometric Fake News Detection Based on Natural Language Processing Using Named Entity Recognition: In-Domain and Cross-Domain Analysis. *Electronics*, *12*(17), 3676. https://doi.org/10.3390/electronics12173676

Vedula, N., & Parthasarathy, S. (2021). FACE-KEG: Fact Checking Explained using KnowledgE Graphs. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 526–534. https://doi.org/10.1145/3437963.3441828

Vladika, J., Schneider, P., & Matthes, F. (2023). HealthFC: Verifying Health Claims with Evidence-Based Medical Fact-Checking. *ArXiv:2309.08503*. https://doi.org/10.48550/arXiv.2309.08503

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Yang, J., Vega-Oliveros, D., Seibt, T., & Rocha, A. (2022). Explainable Fact-Checking Through Question Answering. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8952–8956. https://doi.org/10.1109/icassp43922.2022.9747214

Yang, Q., Christensen, T., Gilda, S., Fernandes, J., Oliveira, D., Wilson, R., & Woodard, D. (2024). Are Fact-Checking Tools Helpful? An Exploration of the Usability of Google Fact Check. *ArXiv:2402.13244*. https://doi.org/10.48550/arXiv.2402.13244

Yao, B. M., Shah, A., Sun, L., Cho, J.-H., & Huang, L. (2023). End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2733–2743. https://doi.org/10.1145/3539618.3591879

Zeng, F., & Gao, W. (2024). JustiLM: Few-shot Justification Generation for Explainable Fact-Checking of Real-world Claims. *Transactions of the Association for Computational Linguistics*, *12*, 334–354. https://doi.org/10.1162/tacl_a_00649

Zhang, J., & El-Gohary, N. M. (2017). Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking. *Automation in Construction*, *73*, 45–57. https://doi.org/10.1016/j.autcon.2016.08.027

Zhang, Z., Rudra, K., & Anand, A. (2021). FaxPlainAC: A Fact-Checking Tool Based on EXPLAINable Models with HumAn Correction in the Loop. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 4823–4827. https://doi.org/10.1145/3459637.3481985

Zhu, B., Zhang, X., Gu, M., & Deng, Y. (2021). Knowledge Enhanced Fact Checking and Verification. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *29*, 3132–3143. https://doi.org/10.1109/taslp.2021.3120636