



Faculty of Computer Science and Information Technology

**A Novel Knowledge Acquisition Framework for Claim Factuality
Disambiguation**

Aruna Shankar

**Doctor of Philosophy
2025**

A Novel Knowledge Acquisition Framework for Claim Factuality
Disambiguation

Aruna Shankar

A thesis submitted

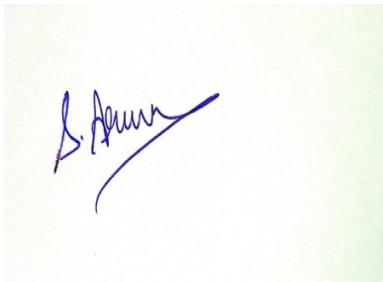
In fulfilment of the requirements for the degree of Doctor of Philosophy

(Software Engineering)

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK

DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Malaysia Sarawak. Except where due acknowledgements have been made, the work is that of the author alone. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.



.....

Signature.

Name: Aruna Shankar

Matric No.: 21010101

Faculty of Computer science and Information Technology

Universiti Malaysia Sarawak

Date: 08/10/2025

ACKNOWLEDGEMENT

I would like to acknowledge all the people who have supported me throughout my work. I would like to express my deepest gratitude to my family, as their support and encouragement were crucial during all these years. Then I would like to acknowledge the most important persons, my mentors and advisors, Prof Dr. Narayanan Kulathuramaiyer and AP Dr. Johari bin Abdullah. Their continuous guidance and support during the proposal, research publications, and thesis write-up were immense. This work would not have been possible without their remarkable input and supervision. They guided me at every single step of this study with their prompt and crucial support. I felt privileged and proud to work under their supervision. My sincere gratitude to the Centre for Graduate Studies and the management of UNIMAS for the advice and support provided during my period of study and stay at Universiti Malaysia Sarawak.

Finally, I would like to thank all my colleagues and friends for their valuable feedback and assistance throughout my studies. Thank you all.

ABSTRACT

The speedy spread of fake news in the digital age has presented significant challenges to accurate and timely truth verification. Traditional fact-checking methods, while valued, are insufficient to address the complexity of online content. Although fact-checking approach offer scalability, they still face critical limitations, including the lack of evidence-supported semantic frameworks for mapping factual coherence between claims and evidence, as well as the absence of mechanisms to capture and represent claim-evidence discrepancies effectively. This research addresses the gap by proposing a novel knowledge acquisition framework that acquires evidence-based knowledge to model the factual coherence relationship between claim and evidence. By systematically mapping claims to retrieved evidence, the framework effectively disambiguates conflicting or ambiguous information, ensuring greater clarity and accuracy in the fact checking process. The significance of the knowledge-level disambiguation approach goes beyond the traditional fact checking methods, as it incorporates a contextual attention mechanism and semantic analysis to detect, represent, and resolve discrepancies. This capability is especially critical for handling complex claims that requires a multi-faceted evidence assessment framework. By ensuring a precise alignment between claim and evidence, the framework effectively resolves semantic uncertainties, thereby enabling a more accurate and context-aware fact-checking. The proposed knowledge level disambiguation mechanism was evaluated using various versions of BERT-based models such as BERT, RoBERTa, and BioBERT on diverse datasets such as FEVER, LIAR-PLUS, HoVer, and PubMed. The outcomes of this research highlight the effectiveness of the proposed hybrid knowledge-level disambiguation approach, which integrates BERT-based models with human-in-the-loop decision-making to ensure accurate claim interpretation, resolve ambiguities, and align claims with relevant evidence. As a result, the approach establishes a robust capability to address the

evidence discrepancies. The findings further underscore the framework's effectiveness in enhancing fact-checking approach by mitigating ambiguities and maintaining precise claim-evidence alignment. This robust performance highlights the framework's potential as a comprehensive solution for combating fake news, even in the presence of conflicting or incomplete information. Building on the current framework, future research will extend the hybrid approach to fact-check multimodal datasets, encompassing text, images, and videos. By integrating multiple data modalities, the approach supports more comprehensive and accurate verification in real-world scenarios, thereby further enhancing its adaptability and effectiveness in mitigating misinformation.

Keywords: Automated fact-checking, misinformation, semantic framework, evidence mapping, discrepancy detection, retrieval, context-based disambiguation

Rangka Kerja Pemerolehan Pengetahuan Novel untuk Nyahkekaburan Faktual Tuntutan

ABSTRAK

Penyebaran pantas berita palsu dalam era digital telah memberikan cabaran besar kepada pengesahan ketepatan pernyataan yang tiba tepat pada masanya. Kaedah semakan fakta tradisional, walaupun bernilai, tidak mencukupi untuk menangani kerumitan kandungan dalam talian. Walaupun pendekatan semakan fakta menawarkan skalabiliti, pendekatan tersebut masih menghadapi kekangan, termasuk kekurangan rangka kerja semantik yang berasaskan bukti untuk memetakan keselarasan fakta antara tuntutan dan bukti, serta ketiadaan mekanisme untuk memperolehi dan mewakili percanggahan di antara bukti dan tuntutan dengan berkesan. Penyelidikan ini menangani jurang tersebut dengan mencadangkan rangka kerja pemerolehan pengetahuan baru yang berasaskan bukti untuk memodelkan hubungan koheren di antara tuntutan dan bukti. Di dalam memetakan tuntutan secara sistematik terhadap bukti yang diperolehi, rangka kerja ini menghasilkan penyahkekaburan maklumat yang berkesan serta, memastikan kejelasan dan ketepatan yang lebih besar dalam proses semakan fakta. Kepentingan pendekatan nyahkekaburan in peringkat pengetahuan melangkaui pencapaian kaedah NLP tradisional, kerana ia menggabungkan mekanisme perhatian kontekstual dan analisis semantik untuk mengesan, mewakili dan menyelesaikan percanggahan. Keupayaan ini adalah amat penting untuk meleraikan tuntutan kompleks yang memerlukan rangka kerja penilaian bukti. Dengan memastikan penjajaran yang tepat antara tuntutan dan bukti, rangka kerja tersebut menyelesaikan ketidakpastian semantik dengan lebih berkesan, sekali gus membolehkan semakan fakta yang lebih tepat dan sesuai dari segi kesedaran konteks. Mekanisme nyahkekaburan tahap pengetahuan seterusnya dinilai menggunakan pelbagai versi

model berasaskan BERT seperti BERT, RoBERTa dan BioBERT serta pelbagai pangkal an data seperti FEVER, LIAR-PLUS, HoVer dan PubMed. Hasil penyelidikan ini menyerlahkan keberkesanan pendekatan nyahkekaburan peringkat pengetahuan hibrid yang dicadangkan, yang menyepadukan model berasaskan BERT dengan pembuatan keputusan manusia dalam gelung untuk memastikan tafsiran tuntutan yang lebih tepat, menyelesaikan kekaburan dan menyelaraskan tuntutan dengan bukti yang berkaitan. Akibatnya, pendekatan ini mewujudkan keupayaan yang teguh untuk menangani percanggahan bukti. Penemuan ini menekankan lagi keberkesanan rangka kerja dalam meningkatkan pendekatan semakan fakta dengan mengurangkan kekaburan dan mengekalkan penjajaran bukti tuntutan yang tepat. Peningkatan prestasi ini menyerlahkan potensi rangka kerja sebagai kaedah penyelesaian komprehensif untuk memerangi ancaman berita palsu, yang semakin mencabar dengan kehadiran maklumat yang bercanggah atau tidak lengkap. Kajian seterusnya akan memanjangkan pendekatan hibrid kepada set data multimodal semakan fakta, merangkumi teks, imej dan video. Dengan menyepadukan pelbagai modaliti data, pendekatan ini bertujuan untuk menyediakan pengesahan yang lebih komprehensif dan tepat dalam senario dunia sebenar, justern meningkatkan lagi kebolehsuaian dan keberkesanannya dalam mengurangkan maklumat salah.

Kata kunci: *Penyemakan fakta automatik, maklumat palsu, rangka kerja semantik, pemetaan bukti tuntutan, pengesanan percanggahan, perolehan semula, nyahkekaburan berdasarkan konteks*

TABLE OF CONTENTS

	Page
DECLARATION	i
ACKNOWLEDGEMENT	ii
ABSTRACT	iii
<i>ABSTRAK</i>	v
TABLE OF CONTENTS	vii
LIST OF FIGURES	x
LIST OF TABLES	xii
CHAPTER 1: INTRODUCTION	1
1.1 Study Background	1
1.2 Problem Statement	8
1.3 The Research Gap	10
1.4 Research Questions	13
1.5 Research Objectives	16
1.6 Research Contributions	17
1.7 Chapter Summary	20
CHAPTER 2: LITERATURE REVIEW	22
2.1 Introduction	22
2.2 Ontology-based Approach	23
2.3 Structure-based Approach	26

2.4	Lexical-based Approach	30
2.5	Content-based Approach	32
2.6	Knowledge-based Approach	34
2.7	Reasoning-based Approach	38
2.8	Logic-based Approach	43
2.9	Graph-based Approach	45
2.10	Context-aware based Approach	50
2.11	Large Language Model in Fact checking	54
2.12	Policy-Level Approaches to Misinformation	56
2.13	Summary of Research Gap	58
2.14	Chapter Summary	60
CHAPTER 3: METHODOLOGY		62
3.1	Overview of the Methodology	62
3.2	Stages of Methodology	65
3.2.1	Problem definition	65
3.2.2	Claim Processing and Evidence Retrieval	66
3.2.3	Factual Disambiguation of Discrepancies Detection Mechanism	74
3.3	Hybrid Framework to Evaluate Factual coherence and Factual Disambiguation	80
3.4	Chapter Summary	82
CHAPTER 4: RESULTS AND DISCUSSION		84
4.1	Introduction	84

4.2	Results of Processing of Claim and Evidence Retrieval	91
4.3	Results of Knowledge Graph Construction	95
4.4	Results of Knowledge Acquisition Framework	100
4.5	Results of Factual Disambiguation for Discrepancy Mechanism	107
4.6	Chapter Summary	121
CHAPTER 5: CONCLUSION AND FUTURE WORK		124
5.1	Introduction	124
5.2	Contributions	124
5.2.1	Evidence-Supported Semantic Mapping Framework for Claim-Evidence Factual Coherence	125
5.2.2	Mechanisms for Identifying and Representing Discrepancies Claim-Evidence Relationship	125
5.2.3	Hybrid Framework to Evaluate Factual Coherence and Discrepancy Detection Mechanism	126
5.3	Limitations of Proposed Work	127
5.4	Future Works	128
5.5	Chapter Summary	129
REFERENCES		130
APPENDICES		151

LIST OF FIGURES

	Page
Figure 1.1 Spread of Misinformation on the Internet	1
Figure 1.2 Examples of Fake News	2
Figure 1.3 The Overview of Research Background	7
Figure 1.4 Research Summary Outline	19
Figure 2.1 Structure-based Approach	26
Figure 2.2 Structure-based Approach in Fact Checking Example	29
Figure 2.3 Linguistics Analysis in Fact Checking	31
Figure 2.4 Knowledge-based Fact-Checking via Entity Linking	35
Figure 2.5 Information Retrieval using Semantics	37
Figure 2.6 Fact-Checking using Rules-based and Logical Inferences	40
Figure 2.7 Pathfinding in Knowledge Graph Fact Verification	47
Figure 2.8 Knowledge Graph Embedding Fact Checking	49
Figure 2.9 Incorporation of Temporal Context Information	51
Figure 2.10 Large Language Model Integration in Fact-Checking	56
Figure 2.11 Literature Review Mapping	59
Figure 3.1 Proposed Knowledge Acquisition Framework for Claim Factuality	
Disambiguation	62
Figure 3.2 Evidence Retrieval and Representation	66
Figure 3.3 Pseudocode of Evidence Retrieval and Representation	68
Figure 3.4 Overview of Knowledge Acquisition for Claim Factuality	71
Figure 3.5 Neighbourhood Context Modeling in Knowledge Acquisition	
Framework	73

Figure 3.6	Pseudocode of Neighbour Context Modeling in Knowledge Acquisition Framework	73
Figure 3.7	Factual Disambiguation of Discrepancies Detection Mechanism	76
Figure 3.8	Evidence sufficiency and logical Coherence Assessment	77
Figure 3.9	Pseudocode of Discrepancies Detection Mechanism	78
Figure 3.10	A Hybrid Human in Loop Evaluation Framework	81
Figure 4.1	Overview of the Framework Validation	84
Figure 4.2	Example of a Constructed Evidence Knowledge Graph	97
Figure 4.3	Evidence Selection Subgraph in Knowledge Acquisition Framework	103
Figure 4.4	Evidence Selection Subgraph before incorporating Human-in-the-Loop	104
Figure 4.5	Evidence Selection Subgraph after incorporating Human-in-the-Loop	104
Figure 4.6	Comparative Results across Benchmarked Approaches	107
Figure 4.7	Case Study of Factual Disambiguation in Evidence Sufficiency	109
Figure 4.8	Evidence Selection Subgraph before incorporating Human-in-the-Loop Correction	112
Figure 4.9	Evidence Selection Subgraph after incorporating Human-in-the-Loop Correction	112
Figure 4.10	Case Study of Factual Disambiguation in Logical Coherence	115
Figure 4.11	Evidence coherence before incorporating Human-in-the-Loop	119
Figure 4.12	Evidence coherence After incorporating Human-in-the-Loop	119

LIST OF TABLES

	Page
Table 4.1 Comparisons Matrix of Evaluation Metrics	87
Table 4.2 Example of Clustering Output for Vitamin C Claim	93
Table 4.3 Evidence Retrieval Performance Metrics	94
Table 4.4 Example of Extracted Triples from PubMed Evidence	98
Table 4.5 Knowledge Graph Construction Performance Metrics	99
Table 4.6 Different ways of Evidence Mapping to Claim	101
Table 4.7 Example of Human-in-the-Loop Adjustment of Evidence Weights	102
Table 4.8 Performance Metrics for Knowledge Acquisition	105
Table 4.9 Contextual Scoring and Sufficiency decision for PubMed Case Study	111
Table 4.10 Comparison Results of Evidence Sufficiency Detection	113
Table 4.11 Comparison Results of Logical Coherence Detection	120

CHAPTER 1

INTRODUCTION

1.1 Study Background

The rise of fake news and misinformation represents one of the most pressing challenges of the digital age. The widespread use of the internet and social media has revolutionized information sharing, enabling content to circulate at unprecedented speed and scale. Unlike traditional media, where editorial oversight provided checks on accuracy, online platforms allow anyone to publish unverified information, creating fertile ground for misinformation (Levak, 2021; Yatid, 2019). Algorithms that prioritize sensational and emotionally charged content further amplify misleading material, ensuring that falsehoods often spread faster and more widely than verified facts. The information disorder framework distinguishes misinformation, disinformation, and malinformation, illustrating the complexity of today's information landscape. Figure 1.1 summarizes the key digital drivers contributing to the spread of misinformation.



Figure 1.1: Spread of Misinformation on the Internet

Source: *Economic Times*. (2020, February 3)

The consequences of misinformation extend across critical domains such as health, politics, and climate change. Political misinformation during the COVID-19 pandemic, for instance, undermined public trust in health authorities, while anti-science propaganda on vaccines and climate change weakened confidence in evidence-based decision-making. Similarly, climate denial narratives framing extreme weather as “natural phenomena” diverted attention from scientific consensus and delayed policy action (Omoregie, 2021). Such cases highlight that misinformation is not a superficial problem but a phenomenon with tangible, real-world consequences for social cohesion, public health, and democratic governance. Figure 1.2 illustrates common examples of misinformation that have been circulated widely online.

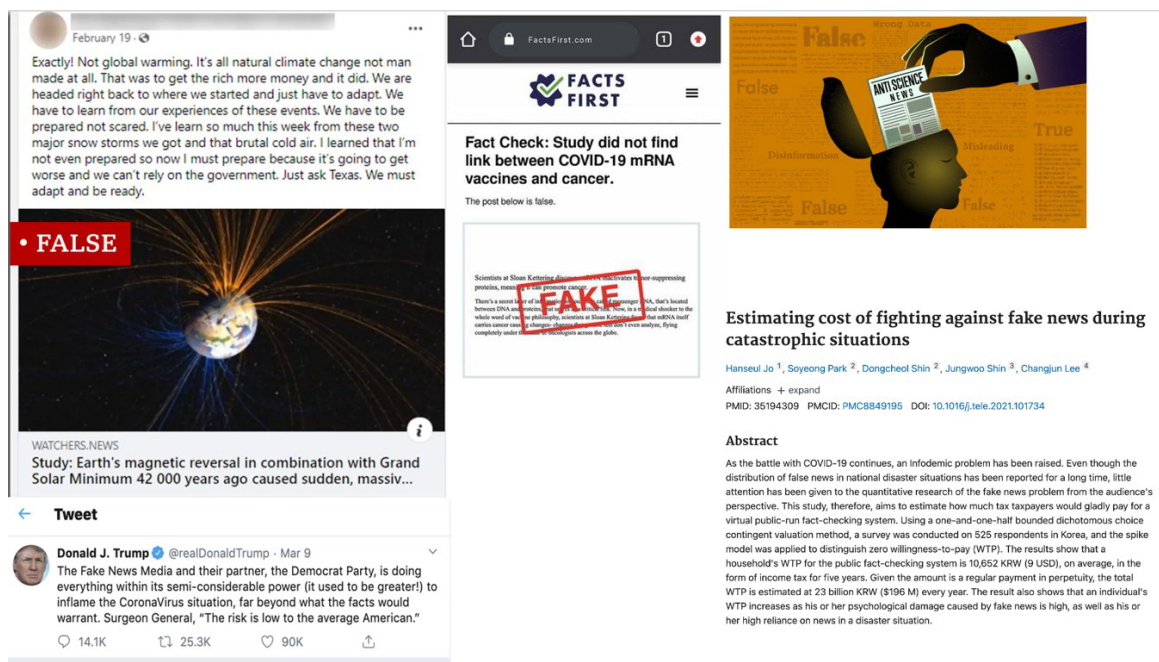


Figure 1.2: Examples of Fake News

Source: <https://www.bbc.com/news/science-environment-59251912>

At the individual level, misinformation creates challenges for unaware users. Many lack the required media literacy, critical reasoning, or technical skills needed to differentiate between credible sources and fabricated claims. As a result, they are more vulnerable to persuasive headlines, doctored images, and narratives framed to exploit emotions. Even digitally savvy users, however, struggle under conditions of information overload, where reliable and unreliable content constantly compete for attention. Cognitive psychology shows that under such pressure, people fall back on heuristics and confirmation biases trusting familiar sources, peer networks, or reinforcing pre-existing beliefs, which deepens polarization and fuels echo chambers. The Covid-19 “infodemic” exemplified this difficulty, where simultaneous exposure to accurate and false claims overwhelmed the public’s ability to discern the truth.

Beyond individual challenges, misinformation must also be understood from a socio-economic perspective. Variations in digital access, literacy levels, and patterns of media consumption shape how different communities encounter and respond to misinformation. For example, reliance on instant messaging applications, social media platforms, or informal community networks influences both the speed and the scale at which false information circulates. At the same time, the rapid expansion of online services and financial technologies exposes users to sophisticated scams and misleading digital content. These dynamics illustrate that misinformation is not merely a technological issue but also a structural and societal challenge, with diverse groups experiencing distinct vulnerabilities depending on how they engage with the digital ecosystem. These socio-economic dynamics illustrate that misinformation is not experienced uniformly but is shaped by the interplay of digital habits, cultural practices, and institutional responses. To better understand these variations, it is important to examine how different countries address the challenge, as each

national context provides unique insights into both the drivers of misinformation and the strategies employed to counter it.

Cross-country comparisons reveal how socio-economic, cultural, and governance contexts shape the manifestation of misinformation. In Malaysia, recent empirical evidence highlights behavioral and socio-economic drivers. A large-scale study of 869 Malaysians revealed that motives for sharing misinformation were primarily rooted in altruism, attitude, and pass-time behavior. Younger users of 18–25 years and individuals with limited access to digital news were particularly prone to these practices (Balakrishnan, 2022). Recognizing these risks, Malaysia has pioneered proactive countermeasures such as the *Rapid Response Troop* and the *sebenarnya.my* fact-checking portal, demonstrating the country's efforts to provide inclusive and accessible fact-checking for its diverse population.

While Malaysia highlights how user behavior and community-driven initiatives shape the misinformation landscape, the United Arab Emirates offers a contrasting perspective where regulatory oversight, digital infrastructure, and demographic diversity play a central role in shaping both the risks and responses to misinformation. In the United Arab Emirates, socio-economic factors take on a different dimension. With one of the world's highest internet penetration rates and a large expatriate population, the UAE faces the dual challenge of misinformation among affluent, digitally connected citizens and among migrant workers who depend on informal, multilingual networks. During the Covid-19 pandemic, false narratives about miracle cures circulated within migrant communities, while phishing and financial scams targeted residents in the context of rapid digitization. The UAE has responded with strict regulatory oversight, cybercrime policing, and national digital literacy campaigns, positioning itself as a governance-led model for combating misinformation (Dubai Police, 2020; UAE Government Portal, 2023).

Together, these cases show that misinformation is not monolithic: in Malaysia, it is shaped by behavioral motives and unequal access to reliable news, while in the UAE, socio-economic diversity and systemic vulnerabilities create distinct risks. Both countries, however, underscore the need for solutions that are technologically robust, socially inclusive, and context aware. Recognizing these complementary perspectives is essential for designing context-aware fact-checking frameworks that address both the social drivers and institutional dimensions of misinformation.

i. Limitations of Manual Fact-Checking

Across both Malaysia and the UAE, fact-checking has emerged as a crucial mechanism for addressing misinformation whether through platforms such as *sebenarnya.my* in Malaysia or through institutional oversight in the UAE. These initiatives reflect a global recognition of the need to verify claims and guide public discourse. However, such approaches remain largely manual in nature. Manual fact-checking, while vital, faces inherent challenges: the overwhelming scale of online content, the rapid pace at which misinformation spreads, and the resource-intensive nature of verification. Corrections often arrive too late to counter viral falsehoods, and fact-checking organizations frequently operate under constraints of time, funding, and reach. These limitations underscore that while manual fact-checking is indispensable, it cannot operate alone in today's fast-moving digital environment (Botnevik et al., (2020); Abdelnabi & Fritz, (2022)).

ii. Need for Automated Fact-Checking

The digital age has significantly transformed how information is created, shared, and consumed. This exceptional access to diverse content has, however, been accompanied by an unprecedented surge in misinformation. Manual fact-checking, though important, cannot

keep pace with the speed and scale of digital information flows. The disparity between the sheer volume of online content and the limited capacity of human fact-checkers underscores the urgent need for scalable, efficient, and real-time automated solutions (Thorne & Vlachos, (2018); Glockner et al., (2022)).

Automated fact-checking leverages advances in Natural Language Processing (NLP) to verify claims at scale by analyzing large datasets in real time. This begins with claim detection, where algorithms identify statements requiring verification and cross-reference them against trusted sources such as academic publications, government records, or verified media outlets. Beyond simple validation, modern frameworks also assess contextual relationships between claims and supporting evidence, significantly enhancing their ability to detect misinformation ((Nikopousios et al., (2023); Du et al., (2022))). Some models go further, incorporating reasoning about argument validity and logical structure, thereby moving fact-checking beyond surface-level matching (Wang et al., (2020)).

iii. Challenges in Automated Fact-Checking

Despite these advances, automated fact-checking still encounters several challenges. Complex claims often involve ambiguous language, evolving domain knowledge, or contradictory evidence. Current systems are efficient at detecting straightforward inconsistencies but struggle with nuanced reasoning—for example, in scientific claims that evolve rapidly or in politically charged statements drawing on conflicting sources. These challenges emphasize the need for computational mechanisms that can model semantic relationships, assess the sufficiency of evidence, and resolve factual discrepancies in a principled way (Liu et al., (2023); Zeng et al., (2021)).

This research addresses these gaps by introducing a knowledge acquisition framework designed to capture semantic relationships between claims and evidence,

evaluate logical coherence, and disambiguate factuality in the presence of uncertainty. By advancing fact-checking from detection to contextual reasoning, the framework strengthens the reliability of automated approaches and enhances their applicability across domains. The overall structure of this research is summarized in Figure 1.3 provides an overview of the research background, situating the research problem and highlighting the challenges of misinformation in the digital age.

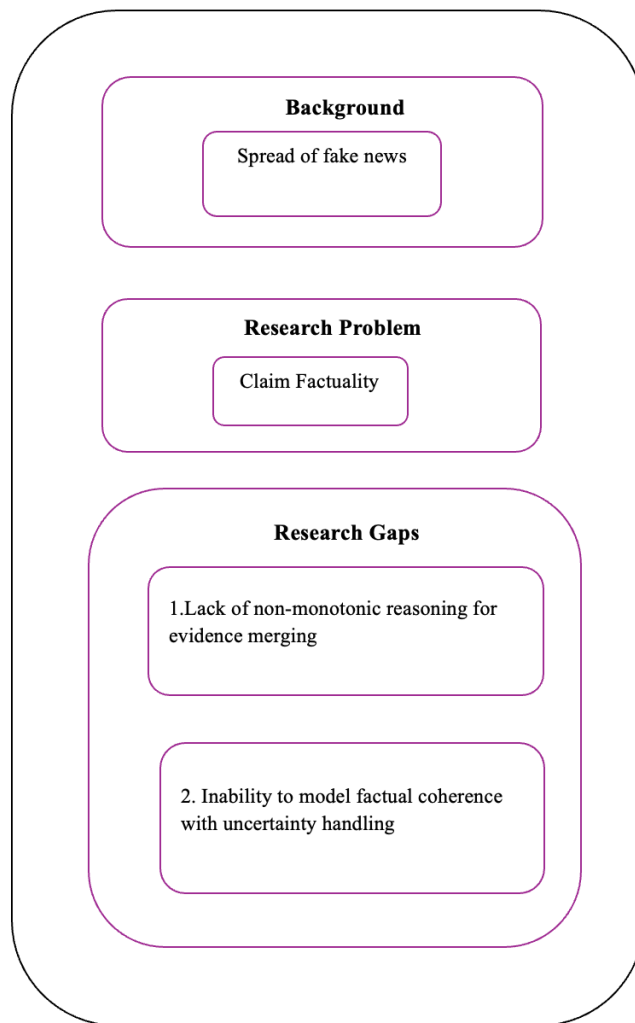


Figure 1.3: The Overview of Research Background

1.2 Problem Statement

While existing fact-checking approaches have made important strides in addressing misinformation, they remain fundamentally constrained by critical weaknesses that undermine their reliability and accuracy. Foremost among these is knowledge-level misalignment, where claims and evidence fail to align in context due to the inability to dynamically map facts to relevant supporting information. This limitation becomes particularly severe in the presence of contradictions, ambiguities, and incomplete evidence, where misinterpretations and overlooked discrepancies tend to be frequent. Compounding the issue, many current approaches rely heavily on surface-level similarity measures, which lack the depth to capture nuanced semantic, logical, and contextual relationships between claims and evidence. Such superficial methods are ill-equipped to operate in today's rapidly evolving and complex information ecosystem, where knowledge is fluid, domain-specific, and often contested. These deficiencies expose a pressing gap: the urgent need for robust, knowledge-driven mechanisms capable of transcending surface matches and incorporating contextual reasoning, coherence assessment, and dynamic adaptability to ensure more accurate, trustworthy, and explainable verification of claims.

i. Challenge 1: Difficulty in Identifying False News in Dynamic Environments

One of the most enduring challenges in automated fact-checking lies in establishing the truth value of claims within highly dynamic information environments. Evidence is rarely static; it is fluid, continuously updated, and often subject to reinterpretation. Claims once false may later be validated by new findings, while previously accepted claims can be overturned as fresh evidence emerges. This unpredictability makes it extremely difficult for

current systems designed largely around static models of fact–evidence alignment to deliver accurate judgments in real time (Sarrouiti et al.,(2021)).

For novice users, who typically lack domain expertise and rely heavily on readily available digital content, this fluidity becomes particularly problematic. They are more likely to accept outdated information as true because they cannot easily discern when the underlying evidence has shifted. Without adaptive fact-checking mechanisms, ill-informed users face increased risks of forming misinformed beliefs based on evolving claims.

ii. Challenge 2: Inability to Resolve Contradictions, and Uncertainties in Evidence

A second, equally significant challenge is the inability of current approaches to effectively manage contradictory, ambiguous, or incomplete evidence. In real-world contexts, it is common for multiple sources to present conflicting conclusions, or for evidence to be too sparse and fragmented to allow confident verification. Current automated systems, which often depend on surface-level comparisons, struggle to navigate such complexities. As a result, they may prematurely validate a claim based on weak or irrelevant evidence, or dismiss it outright due to the absence of sufficient supporting information.

Contradictions and uncertainties are especially difficult to resolve for non-expert users, who are general information consumers lacking training in verification practices. Unlike expert fact-checkers, they often lack the skills to critically evaluate the credibility or contextual relevance of competing sources. When presented with fragmented or conflicting information, non-expert users may either adopt the wrong conclusion or disengage entirely, reinforcing misinformation exposure. This highlights the urgent need for fact-checking approaches that can guide non-expert users through contradictory evidence by making

uncertainties explicit, thereby reducing confusion and strengthening trust in verification systems.

Together, these challenges reveal that misinformation is not only a technical issue but also a user-centric problem, excessively affecting non-expert users. In dynamic environments, novices often lack the contextual awareness needed to distinguish between updated and outdated evidence, leaving them vulnerable to shifts in factual accuracy. Similarly, when confronted with contradictions or incomplete evidence, novices are more likely to adopt oversimplified conclusions or disengage altogether, amplifying misinformation’s impact. This highlights the urgent need for knowledge-driven fact-checking frameworks that not only manage evidence dynamics and contradictions but also present verification outcomes in ways accessible to novice users. By integrating contextual reasoning and knowledge-centric evaluation, such frameworks can bridge the gap between complex verification processes and user comprehension, ensuring accuracy, clarity, and trustworthiness in fact-checking.

1.3 The Research Gap

Despite significant advancements in automated fact-checking, a critical limitation persists in effectively modeling the factual connections between claims and evidence. Current approaches (Chernyavskiy et al., (2021); Nikopensius et al.,(2023)) often treat claims and evidence as isolated entities, relying heavily on surface-level techniques such as keyword overlap or shallow semantic similarity. While useful in simple cases, these methods fail to capture the complex, context-dependent interactions between claims and evidence—whether they support, contradict, or only partially relate to one another. As a result, existing fact-checking systems struggle to ensure reliability in complex scenarios, leading to misclassifications, overlooked discrepancies, and a lack of explanatory depth.

i. Lack of Non-Monotonic Reasoning Framework for Evidence Merging in Factuality Checking

A major gap lies in the absence of robust frameworks for merging evidence across diverse contexts, especially in cases involving contradictory or insufficient evidence. Most existing approaches remain static, linking claims to evidence using keyword-based retrieval (Ostrowski et al., (2020)) or simple semantic similarity measures (Popat et al., (2018)). These methods are inadequate when evidence is distributed, contextually diverse, or evolving, as they fail to reconcile contrasting contexts into a coherent factual judgment. For example, a claim may be partially supported by several pieces of evidence drawn from different domains or time periods. Without mechanisms for non-monotonic reasoning which allows truth values to evolve as new evidence emerges fact-checking systems often either misclassify the claim or ignore critical evidence relationships. Non-monotonic reasoning provides the adaptability required for factuality assessment in dynamic environments, where evolving insights can rapidly alter previously established conclusions. Addressing this gap is crucial for ensuring fact-checking frameworks remain responsive and accurate in real-world contexts.

ii. Lack of Coherent Modelling of Factuality with Uncertainty Handling Capabilities

Closely tied to the absence of non-monotonic reasoning is a second critical shortcoming: the lack of coherent factuality models capable of explicitly handling uncertainty, insufficiency, and contradiction. Most existing systems reduce verification to a binary decision of either true or false, which oversimplifies the complexity of real-world evidence. In practice, claims are often accompanied by insufficient information, where evidence is limited or fragmented, or by contradictions, where multiple sources provide

opposing conclusions. Similarly, evidence may stem from heterogeneous or cross-domain sources, introducing inconsistencies that cannot be resolved through surface-level comparisons alone.

Without mechanisms to weigh the credibility of sources, resolve conflicts across contradictory findings, or represent uncertainty in situations of evidence insufficiency, current systems risk making premature judgments. Claims may be incorrectly validated on the basis of weak or partial evidence, or alternatively, valid claims may be dismissed due to missing contextual details. Such oversimplifications not only compromise accuracy but also limit the explanatory depth of verification outcomes, which is vital for establishing user trust and ensuring transparency in the fact-checking process.

To address these limitations, there is a pressing need for a coherent factuality model that integrates mechanisms for evidence sufficiency, contradiction resolution, and explicit uncertainty representation. By systematically mapping claims to retrieved evidence, disambiguating ambiguous or contextually inconsistent information, and dynamically updating truth values as new evidence emerges, such a model can provide a more reliable and adaptive approach to automated verification. This ensures not only improved robustness in the evaluation of claims but also enhances the overall credibility and transparency of fact-checking systems.

This thesis introduces a Knowledge Acquisition Framework specifically designed to bridge the identified gaps in automated fact-checking. The framework begins by harvesting evidence-based knowledge across heterogeneous sources, ensuring that claims are evaluated within a broad and diverse evidential landscape rather than being confined to isolated datasets. This systematic integration allows the model to capture supporting, contradictory and contextual evidence that is often overlooked by surface-level approaches.

Building on this foundation, the framework applies non-monotonic reasoning to dynamically merge and reconcile evidence. Unlike static models that treat evidence as fixed, non-monotonic reasoning enables the truth value of claims to evolve as new information emerges. This adaptability is critical in domains such as science, health, and current affairs, where knowledge is constantly shifting and previously verified claims may be challenged by new findings. Finally, the framework explicitly models factual coherence under uncertainty, incorporating mechanisms that account for contradictions, insufficiencies, and the dynamic nature of knowledge. By doing so, it moves beyond binary classifications and provides a more nuanced representation of claim validity. These capabilities not only improve accuracy and robustness but also enhance the explainability and transparency of verification outcomes, which are essential for building trust in automated systems. By addressing both the absence of non-monotonic reasoning and the lack of uncertainty-aware factuality models, the proposed Knowledge Acquisition Framework advances the field toward fact-checking systems that are more reliable, adaptable, and capable of operating effectively in complex, real-world evidence scenarios.

1.4 Research Questions

Fact-checking is essential for preserving the integrity of knowledge, particularly in an era where misinformation significantly shapes public opinion and decision-making. Yet, the online information landscape is highly complex evidence is often contradictory, incomplete, cross-domain, and contextually ambiguous, posing critical challenges for automated fact-checking systems. Current approaches frequently lack the ability to dynamically adapt to evolving evidence, limiting their robustness in real-world, continuously changing environments.

The challenge therefore extends beyond simply labeling of claims as true *or* false. Fact-checking in real-world settings requires mechanisms that can systematically identify evidence insufficiencies, effectively resolve contradictions and inconsistencies, and manage uncertainty within dynamic and evolving contexts. Without such capabilities, verification systems risk oversimplifying complex claim–evidence relationships, resulting in reduced reliability and accuracy.

In order to address these limitations, this research introduces a Knowledge Acquisition Framework that seeks to model claim–evidence relationships more coherently. By incorporating techniques to reconcile discrepancies, account for evolving evidence, and represent uncertainty, the framework enhances the robustness and adaptability of automated fact-checking. In doing so, it establishes a stronger foundation for reliable claim factuality assessment. To guide this investigation, the following research questions are formulated.

i. Central Research Question

The rise of misinformation in the digital spaces underscores the need for a robust mechanisms that can evaluate the credibility of claims in real time. Traditional fact-checking often struggles with evolving evidence, fragmented sources, and contradictory information, making it difficult to ensure the consistent and reliable verification. To address these challenges, this thesis is guided by the following central research question: How can the truthfulness of a claim be dynamically established in evolving evidence and discrepancies to effectively mitigate the spread of fake news on the web?

ii. Supporting Research Questions

The supporting research questions refine the central inquiry by addressing how coherence can be modeled, how discrepancies can be identified, how external knowledge

can be integrated and which evaluation metrics are most suitable for the benchmarking reliability. These questions serve as the intellectual backbone of the study, linking theoretical gaps to the practical contributions of the framework.

a. How can coherence between claims and evidence be effectively captured and modeled to establish claim factuality?

Coherence between claims and evidence can be captured through semantic modeling of neighborhood context, which systematically maps the relational pathways between entities in structured knowledge representations. This allows the system to move beyond surface similarity and assess deeper contextual alignment.

b. What are the key indicators of discrepancies in claim–evidence relationships, how can they be detected?

Discrepancies such as semantic contradictions, incomplete evidence, contextual misalignments, and cross-domain inconsistencies can be detected using contextual attention mechanisms. These mechanisms assign weights to the most salient parts of evidence, enabling precise identification of conflicts, insufficiencies, and ambiguities.

c. How can external knowledge sources be integrated to strengthen contextual interpretation of claims and evidence?

The incorporation of external knowledge sources, such as PubMed and other domain-specific repositories, enhances contextual understanding and reduces misalignment. By validating claims against authoritative and cross-domain sources, the system becomes more competent to address the knowledge gaps and better equipped to manage domain-specific factuality.

d. What evaluation metrics and benchmarks are most suitable for assessing factual coherence modeling and discrepancy detection?

The evaluation requires benchmarking against established datasets such as FEVER, LIAR-PLUS, HoVer, PubMed and state-of-the-art baselines such as BERT, RoBERTa, BioBERT). Metrics such as precision, recall, F1-score, and accuracy are used to assess both accuracy and robustness in handling insufficiency, contradiction, and evolving evidence.

1.5 Research Objectives

Building upon the identified research questions, this thesis aims to achieve the following objectives:

i. Develop an Evidence-Supported Semantic Mapping Framework For Claim-Evidence Factual Coherence

The first objective is to design a semantic framework that systematically models the coherence between claims and evidence. Current fact-checking approaches often rely on shallow techniques such as keyword matching or basic similarity measures, which fail to capture deeper contextual relationships. This leads to inaccuracies when claims involve indirect references, conditional statements, or implicit reasoning. The proposed framework will integrate semantic embeddings and contextual cues to enhance factual alignment, and thereby enabling more reliable evaluations of claim truthfulness across diverse scenarios.

ii. Design Mechanisms to Identify and Represent Discrepancies in Claim-Evidence Factual Coherence

The second objective focuses on developing mechanisms to detect and represent discrepancies such as logical inconsistencies, contradictory evidence,

and insufficient information. Current models often reduce fact-checking to binary outcomes (true/false), oversimplifying complex relationships. This research proposes the use of contextual attention mechanisms to identify critical evidence components, assess semantic coherence, and differentiate between partial support, contradictions, and inconsistencies. By representing discrepancies explicitly, the framework will improve the accuracy, reliability, and interpretability of automated fact-checking in real-world contexts.

iii. Develop a Hybrid Framework for Evaluating Factual Coherence and Discrepancy Detection Mechanisms

The third objective is to design a hybrid evaluation framework that integrates both benchmark-driven testing and human-in-the-loop validation. The evaluation will rely on established datasets such as FEVER, LIAR-PLUS, HoVer, and PubMed to measure precision, recall, F1-score, and accuracy across diverse verification scenarios. Beyond automated benchmarking, expert reviewers will validate edge cases, verify coherence mappings, and refine discrepancy detection mechanisms. This dual evaluation approach ensures that the framework not only meets quantitative performance standards but also aligns with human judgment in handling ambiguous or complex cases.

1.6 Research Contributions

This thesis outlines the key contributions made in this research, focusing on the development of techniques for factual coherence modelling, discrepancy detection and evaluation mechanisms in automated fact-checking process. The research addresses critical challenges, including ensuring contextual accuracy, identifying claim-evidence

inconsistencies, and improving process performance through a hybrid evaluation framework. The contributions are organized around three primary research objectives, each playing a crucial role in enhancing the reliability of fact-checking approach.

i. Evidence-Supported Semantic Mapping Framework For Claim-Evidence Factual Coherence

The research introduces a novel semantic framework that goes beyond surface-level keyword matching to capture deeper contextual relationships between claims and evidence. By leveraging contextual embeddings and semantic similarity analysis, the framework ensures accurate alignment even when claims and evidence are expressed differently or drawn from diverse contexts. This contribution significantly reduces semantic gaps by improving the accuracy, contextual clarity, and robustness of fact-checking systems.

ii. Mechanisms for Identifying and Representing Discrepancies in Claim-Evidence Relationships

The second contribution is the development of mechanisms to identify and represent three critical forms of discrepancies: non-substantiated claims, contradictory evidence, and incomplete evidence. Through the use of contextual attention, the system can highlight inconsistencies and focus on the most relevant parts of the evidence. Ambiguous or complex cases are flagged for expert validation, ensuring human oversight where automated techniques may fall short. This contribution strengthens fact-checking outcomes by making them more accurate, explainable, and adaptable to evolving information.

iii. Hybrid Framework to Evaluate Factual Coherence and Discrepancy Detection Mechanism

The final contribution is the introduction of a hybrid evaluation framework that combines automated benchmarking with human-in-the-loop validation. Benchmarking across diverse datasets such as FEVER, LIAR-PLUS, HoVer, and PubMed ensures quantitative rigor, while expert reviews add contextual and interpretive depth. This dual approach enhances the framework’s credibility, ensuring reliable performance not only in controlled experimental settings but also in complex, real-world fact-checking scenarios.

Together, these contributions advance the state of automated fact-checking by integrating semantic depth, discrepancy awareness, and robust evaluation. The framework thus offers a more accurate, transparent, and resilient approach to establishing claim factuality, addressing the core challenges of misinformation in today’s digital ecosystem. A summary of the research outline, including key components of the framework, is illustrated in Figure 1.4.

Problem Statement	Research Gap	Research objectives	Research Contributions
Spread of fake news on the web	1. Lack of Non-Monotonic Reasoning Framework for Evidence Merging in Factuality Checking	1. Develop an Evidence-Supported Semantic Framework for Mapping Claim-Evidence Factual Coherence	1. Evidence-Supported Semantic Mapping of Claim-Evidence Factual Coherence
	2. Lack of Coherent Modelling of Factuality with Uncertainty Handling Capabilities	2. Design Mechanisms to Identify and Represent Discrepancies in Claim-Evidence Factual Coherence	2. Mechanisms for Identifying and Representing Discrepancies in Claim-Evidence Relationships
		3. Design a Hybrid Framework for Evaluating Factual Coherence Mapping and Discrepancy Detection Mechanisms Using Benchmarking and Human-in-the-Loop Validation	3. Hybrid Framework to Evaluate Factual Coherence and Discrepancy Detection Mechanisms

Figure 1.4: Research Summary Outline

1.7 Chapter Summary

Chapter 1 has laid the foundation for this research by situating it within the broader challenge of combating misinformation and fake news in the digital age. It has emphasized the importance of automated fact-checking as a scalable solution but highlighted that existing approaches face significant limitations when confronted with complex, evolving, and contradictory evidence. It has been underscored that misinformation not only misguides public discourse but also undermines trust in digital information systems, making the need for advanced fact-checking mechanisms both timely and critical.

Two central research gaps were identified as key barriers. The first is the absence of non-monotonic reasoning frameworks that can merge and reconcile evidence dynamically, adapting to new or contradictory information as it emerges. Without such reasoning capabilities, automated systems struggle to incorporate evolving knowledge or revise earlier conclusions. The second is the lack of coherent factuality models that explicitly address uncertainty, insufficiency, and contradictions in evidence. Current approaches often simplify fact-checking into binary true/false outputs, which leads to premature validation of claims based on weak evidence or the dismissal of potentially valid claims due to incomplete context. These gaps highlight the need for frameworks that move beyond surface-level similarity and shallow matching toward context-aware, knowledge-driven approaches capable of handling real-world complexity.

The first objective explores the design and development of an evidence-supported semantic framework for mapping claim–evidence coherence, ensuring deeper contextual alignment between claims and supporting evidence. The second objective explores the design mechanisms for identifying and representing discrepancies, such as contradictions, partial support, or incomplete information, within claim–evidence relationships. The third

objective then formulates a hybrid evaluation framework that combines quantitative benchmarking on standard datasets such as FEVER, LIAR-PLUS, HoVer, PubMed with expert-informed validation, ensuring both statistical rigor and contextual reliability. Together, these objectives aim to improve the adaptability, precision, and transparency of fact-checking systems.

Chapter 2 provides a comprehensive literature review, examining existing fact-checking methods, highlighting their strengths and weaknesses, and situating the identified research gaps within the broader field. Chapter 3 presents the methodology, detailing the design and implementation of the proposed knowledge acquisition framework, including evidence retrieval, claim–evidence mapping, discrepancy detection, and evaluation protocols. Chapter 4 reports on the experimental results, demonstrating how the framework performs across multiple datasets and evaluation scenarios, and validating its effectiveness in addressing the stated research gaps. Chapter 5 concludes the thesis by summarizing key findings, discussing limitations, and suggesting directions for future research, particularly in enhancing the adaptability of fact-checking systems in increasingly complex digital ecosystems.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents a structured review of the literature on automated fact-checking, establishing the foundation for this research. It explores key concepts, methodologies, and challenges that have shaped the fact-checking approach, providing insights to guide the development of the proposed claim factuality framework. The review begins with an evaluation of existing fact-checking methodologies, emphasizing their strengths, limitations, and effectiveness in modeling factual coherence, particularly in knowledge-centric claim-evidence relationships where semantic accuracy and contextual alignment are critical.

A critical assessment of current approaches reveals significant research gaps, including the inability to effectively model factual coherence, resolve contradictions, and handle incomplete or ambiguous evidence. The chapter further reviews diverse fact-checking strategies, including ontology-based models, graph-based techniques, and embedding-based methods, comparing their effectiveness in verifying claims. By critically engaging with prior work, this chapter identifies key gaps in existing fact-checking approaches, underscoring the necessity for a novel framework that enhances factual coherence modeling and discrepancy detection. These insights provide the foundation for the design and evaluation of the proposed knowledge acquisition framework for claim factuality disambiguation in the following chapters.

2.2 Ontology-based Approach

Ontology-based approaches have long been viewed as the backbone of structured reasoning in fact-checking because they allow claims to be verified by modeling explicit relationships between entities. Instead of treating text as a bag of words, these approaches create meaning by asking: “*How do entities relate to each other in a structured knowledge framework?*” For example, if the claim is “*Barack Obama was born in Hawaii,*” an ontology can check the *isBornIn* relation between the entity *Obama* and the entity *Hawaii* in a structured knowledge base. This makes them particularly appealing in domains where semantic accuracy is critical.

Early works such as (Lin et al., (2017)) demonstrated how rule-based reasoning aligned natural language claim representations with ontological structures. For instance, given the claim “*The book Harry Potter was written by J.K. Rowling,*” the system matched entities such as *Harry Potter* $\rightarrow$ Book and *J.K. Rowling* $\rightarrow$ Author and validated the *writtenBy* relationship using predefined rules. The result was a verification decision that supported the claim when the relation was present or refuted it otherwise. Similarly, (Ren et al., (2017)) extended rule-based reasoning to Chinese text, extracting entities such as *Confucius* and *Analects* and mapping them into ontology classes before applying rules to assess factuality. However, such rigid reasoning struggled with claims in datasets like FEVER, where evidence often comes from loosely structured Wikipedia text. For example, the claim “*The film Inception was directed by Christopher Nolan*” required reasoning over a *directedBy* relation for newly emerging films and directors, something static ontologies often failed to capture.

To address scalability, another stream of work introduced edge-distance semantic similarity techniques to quantify relationships mathematically within ontology graphs.

Sébastien et al. (2020) measured semantic relatedness based on path distances between nodes. In a biomedical context, a claim like “*Heparin is an allergen*” was evaluated by computing the shortest path between *Heparin* and the class *Allergen*. A shorter distance yielded stronger semantic similarity, producing a numeric factuality score. While this made integration with large-scale datasets such as LIAR-PLUS possible, such methods could only indicate relatedness and could not capture contradictions and contextual insufficiencies in evidence.

Recent advances combine ontological reasoning with neural embeddings and generative AI to adapt to dynamic knowledge. (Li et al., (2023)) introduced DRAGON-AI, which integrated domain ontologies with large language models (LLMs) to continuously update biomedical knowledge graphs. When testing a PubMed claim like “*Aspirin reduces the risk of heart attack,*” the system not only validated the *reducesRiskOf* relation in the ontology but also assimilated new biomedical findings to strengthen the factuality assessment. In 2024, Ontology-Guided Retrieval-Augmented Generation (OG-RAG) emerged, constraining generative LLMs to ontology-aligned evidence. For example, when applied to a HoVer claim such as “*The Apollo 11 mission landed on the Moon in 1969,*” OG-RAG ensured that the model only retrieved evidence consistent with ontology-defined space mission data, thereby reducing hallucinations and improving factual reliability (Sharma, Kumar, & Li, 2024). Most recently, (Kang et al., (2025)) introduced contrastive ontology embeddings, embedding relations directly into vector spaces for contradiction detection. This enabled nuanced reasoning; for instance, the claim “*COVID-19 is caused by bacteria*” was flagged as contradictory by aligning against embedded biomedical ontologies that define COVID-19 as a viral disease.

Taken together, this trajectory illustrates a clear evolution in claim representation, reasoning techniques, and verification of outputs. Early ontology systems relied on rigid rule-based reasoning to produce binary verification. Graph-based distance measures improved scalability but lacked contradiction handling. Hybrid neural–ontology models (Li, OG-RAG, Kang) now merge structured reasoning with adaptive learning, yielding dynamic and explainable factuality assessments that adapt to evolving knowledge.

However, despite their semantic precision, ontology-driven frameworks face well-documented limitations. (Lin et al., (2017)) reported that rule-based reasoning failed when evidence did not strictly conform to predefined relations, leading to misclassifications in loosely structured corpora like FEVER. Similarly, Sébastien et al. (2020) noted that while path-distance measures improved scalability, they could not distinguish between supportive and contradictory relations, producing false positives in biomedical contexts. More recently, (Li et al., (2023)) acknowledged that even hybrid ontology–LLM systems struggled with indirect claims and rapidly evolving biomedical findings, requiring continuous updates to maintain relevance. (Kang et al., (2025)) further highlighted that contradiction detection with ontology embeddings remained domain-sensitive, with errors occurring when relations were under-specified. Taken together, these findings demonstrate that static ontologies—while effective for explicit relational checks—lack the flexibility to handle contextual variability, contradictions, and dynamically evolving knowledge.

These persistent gaps motivate the exploration of structure-based approaches, which relax the rigid schema dependency of ontologies. Instead of enforcing predefined class–relation mappings, structure-based methods emphasize connectivity patterns, entity–relation graphs, and topological reasoning. This shift allows systems to model claim–evidence relationships more flexibly, enabling verification even in domains with incomplete, contradictory, or continuously evolving information.

2.3 Structure-based Approach

While Ontology-based approaches excel at enforcing semantic precision, their rigid reliance on predefined schemas limits adaptability to the fluid and context-dependent nature of real-world discourse. Structure-based approaches address this by modeling claims, premises, and conclusions as interconnected structures, emphasizing relational links and discourse flows. As illustrated in Figure 2.1, this process begins with extracting claims, identifying premises, and mapping their relationships into a structured representation such as argument trees or graphs. This forms the basis for reasoning about factual coherence beyond surface-level textual similarity.

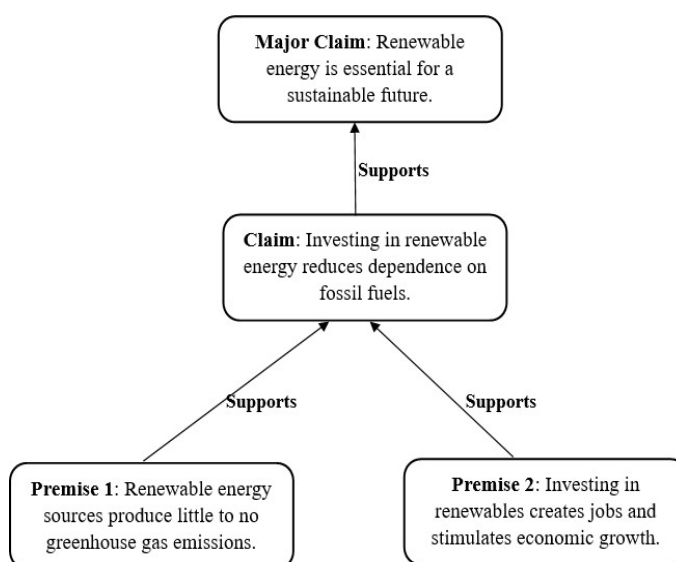


Figure 2.1: Structure-based Approach

Source: (Stab & Gurevych, 2014; Peldszus & Stede, 2013)

The earliest attempts at structure-based fact-checking relied on linguistic markers such as “because” and “therefore” to identify argumentative components. (Peldszus & Stede (2013)) demonstrated that discourse cues could reveal premises and conclusions in short texts. However, they also reported that such methods were brittle across languages and domains, since argumentative markers are expressed differently depending on linguistic and

cultural norms. This limitation set the stage for machine learning approaches, which aimed to capture argument structures more flexibly.

Building on this foundation, (Stab & Gurevych (2014)) applied machine learning classifiers such as Support Vector Machines (SVMs) and Random Forests to argumentative essay corpora. Their models learned features that distinguished claims from premises, improving classification accuracy over rule-based systems. Yet, they also noted that models trained on English academic writing performed poorly when applied to political debates or biomedical text. This confirms that structure-based models face language and domain transferability issues (Stab & Gurevych, (2014); Peldszus & Stede, (2013)). The need for deeper semantic modeling then motivated the adoption of neural embedding methods.

The introduction of transformer-based models addressed this semantic gap. BERT (Devlin et al., (2019)) captured nuanced relationships between claims and premises, enabling more robust discourse-level analysis. For example, in FEVER, BERT aligned the claim “Inception was directed by Christopher Nolan” with supporting evidence, recognizing a premise–conclusion dependency. Domain-specific fine-tuning (Chakrabarty et al., (2019)) further improved stance detection and evidence alignment in social media contexts. However, despite these advances, neural embeddings alone often failed in multi-hop reasoning or when evidence contained implicit premises, leading to the rise of graph-based reasoning approaches.

To overcome these challenges, researchers introduced graph-based reasoning and dependency parsing. (Lippi & Torroni (2016)) modeled support and attack relations as argument graphs, while Nguyen & Litman (2016) built hierarchical argument trees with neural networks to represent layered debate structures. More recently, Graph Attention Networks (GATs) and discourse parsing extended this capability, allowing models to track

argumentative flows across multiple documents. For example, (Li et al., (2023)) demonstrated that verifying the HoVer claim “*The Apollo 11 mission commander’s birthplace is Ohio*” required traversing multiple relations (Apollo 11 → Neil Armstrong → birthplace: Ohio), something structural graph models captured effectively. However, while graphs improved logical reasoning, they often lacked interpretability for human evaluators, which brought renewed focus to argumentation schemes.

Argumentation schemes such as “*argument from authority*” or “*cause-to-effect*” provide interpretable logical templates (Walton et al., (2008); Feng & Hirst, (2011)). They enable explicit reasoning patterns but face difficulties in real-world datasets where arguments are implicit or fragmented. For example, assessing the biomedical claim “*Aspirin reduces the risk of heart attack*” requires synthesizing diverse premises from clinical trials and risk factor studies. (Xie et al., (2024)) addressed this challenge by embedding argumentation schemes into transformer models, creating hybrid discourse–semantic frameworks that captured implicit premises more effectively. These hybrid models also led to renewed emphasis on evaluation metrics.

Evaluation techniques evolved in parallel with modeling methods. (Rinott et al., (2015)) pioneered probabilistic scoring of arguments, which was later extended with probabilistic graphical models (Thorne et al., (2018)) and credibility grounding. (Chen et al., (2025)) further incorporated source reliability into evaluation. For example, in LIAR-PLUS, evaluating the claim “*A senator voted against the healthcare bill*” requires integrating supportive and contradictory premises from multiple news sources, while weighting them by credibility. These studies highlight that evaluation is not only about sufficiency of evidence, but also about consistency, trustworthiness, and structural coherence.

Figure 2.2 illustrates how structure-based discourse analysis operates in fact-checking by decomposing the claim “*Taking vitamin C supplements prevents the common*

cold” into argumentative components and aligning them with multiple evidence statements. Evidence 1 highlights immune support and the fact that Vitamin C is commonly recommended, which structurally represents a supporting premise. Evidence 2 notes that users of Vitamin C report fewer cold symptoms, also functioning as a supportive premise, though expressed in experiential rather than causal terms. Evidence 3, by contrast, discusses Vitamin C’s role in collagen synthesis and wound healing, which is only indirectly related and therefore a structurally weaker form of support. By mapping these relations, the analysis reveals that while some evidence supports the claim, other evidence is only tangentially connected. The resulting outcome is a “Mixed Results” contextual implication, demonstrating how structure-based approaches uncover the strength, weakness, and relevance of premises. Unlike surface-level similarity methods, discourse analysis exposes the logical connections or the lack thereof between a claim and its supporting evidence.

Claim: Taking vitamin C supplements prevents the common cold

Evidence:

1. Vitamin C support the immune system and is commonly recommended during cold and flu season.
2. Many people who take vitamin C supplements report fewer cold symptoms.
3. Vitamin C is essential for collagen synthesis, which plays a crucial role in skin health and wound healing.

Contextual Implications: Mixed Results

Figure 2.2: Structure-based Approach in Fact Checking Example

Despite these advances, structure-based methods face significant challenges. Many arguments are indirect, requiring inference to uncover unstated premises or conclusions. As (Peldszus & Stede (2013)) and (Stab & Gurevych (2014)) both noted, cross-lingual and domain-diverse settings remain problematic, with accuracy dropping when models are applied beyond their training domains. Multi-modal arguments (text + image + video) are also underexplored, though (Zhang et al., (2025)) introduced multimodal discourse parsing to tackle this for misinformation detection. Moreover, argument persuasiveness, rhetorical

strategies, and subjective framing complicate the development of standardized evaluation metrics.

Taken together, structure-based approaches have expanded the frontier of fact-checking by shifting the focus from surface-level evidence matching to argument-level reasoning. They provide a flexible foundation for modeling factual coherence, enabling systems to evaluate not only whether evidence exists, but whether it logically supports or contradicts a claim. These strengths and gaps directly motivate the framework proposed in this thesis, which integrates structural reasoning with contextual and semantic alignment.

Although structure-based approaches excel at modeling the logical relations between claims and premises, they often depend on higher-level discourse representations and can overlook the finer lexical cues that signal factuality, subjectivity, or verifiability. This gap highlights the complementary role of lexical-based approaches, which focus on the surface-level properties of words and phrases to provide interpretable and efficient entry points for fact-checking.

2.4 Lexical-based Approach

Lexical analysis represents one of the earliest and most fundamental pillars of automated fact-checking. These methods focus on the surface-level properties of language—words, phrases, and structural patterns to distinguish factual statements from unverifiable or opinion-based content. By examining verifiability markers, literalness, quantifiers, and subjectivity in language, lexical methods provide interpretable signals for detecting potentially misleading claims.

Early work by Hermann & Rubinfeld (1985) demonstrated that factual statements often include literal, verifiable terms such as dates, named entities, and quantifiable facts, whereas opinionated or subjective statements rely more heavily on emotive or evaluative vocabulary. Building on this foundation, Kyle & Crossley (2015) showed how lexical

sophistication, and concreteness could serve as reliable indicators of verifiability, enabling systems to classify statements more effectively as factual or interpretive.

Figure 2.3 illustrates this role of lexical analysis in practice. In the claim *“Taking vitamin C supplements prevents the common cold”*, the word *“prevents”* acts as a strong verifiability marker, signaling a factual assertion. Similarly, in the evidence statements, lexical cues such as *“commonly recommended”*, *“many people”*, and *“essential for collagen synthesis”* indicate factuality, frequency, or causality. These highlighted features exemplify how lexical analysis operates at the word and phrase level, segmenting input text into claim–evidence components based on linguistic signals. This aligns directly with early findings (Hermann & Rubenfeld, 1985; Kyle & Crossley, 2015) and establishes lexical analysis as a critical entry point in fact-checking pipelines.

Claim: Taking vitamin C supplements prevents the common cold
Evidence:
1. Vitamin C support the immune system and is commonly recommended during cold and flu season.
2. Many people who take vitamin C supplements report fewer cold symptoms.
3. Vitamin C is essential for collagen synthesis, which plays a crucial role in skin health and wound healing.
Vocabulary: Terms relate directly to health and wellness

Figure 2.3: Linguistics Analysis in Fact Checking

Source: (Xu et al. (2024))

With the growth of large-scale misinformation datasets, lexical features were increasingly embedded into machine learning architectures. For instance, (Lee et al., (2018)) integrated lexical tagging into a decomposable attention model, improving evidence retrieval by aligning lexical cues with semantic similarity. More recently, hybrid lexical–semantic approaches have demonstrated the significant gains such as (Thakur et al., (2023)) used lexical–semantic embeddings to detect subtle manipulations in political discourse, while (Xu et al., (2024)) combined lexical concreteness with transformer encoders to enhance stance

detection in health misinformation. (Chen et al., (2025)) further advanced this line by introducing a dynamic lexical–contextual scoring model, which integrates lexical literalness with contextual embeddings to reduce errors in multi-domain claims such as climate and biomedical misinformation.

Despite these advancements, lexical approaches have well-documented limitations. They often misclassify claims that rely on inference or deep contextual reasoning. For example, the sentence “*The senator stood against healthcare reform*” cannot be fact-checked accurately without historical and political context. Polysemy also complicates analysis: the word “*lead*” may refer to a chemical element or an action, depending on context, making surface-level interpretation unreliable (Thurmair et al., (2012)). Figure 2.3 reflects this limitation as well, since while “*prevents*” indicates a strong factual assertion, it may exaggerate causality without corroborating biomedical evidence.

Lexical-based approaches provide interpretable and efficient cues, but their focus on surface-level signals limits their effectiveness. This motivated the shift toward content-based approaches, which emphasize semantic meaning and domain-specific knowledge to achieve more accurate factual verification.

2.5 Content-based Approach

Content-based approaches represent a core dimension of automated fact-checking, as they emphasize the direct alignment between the textual content of claims and corresponding evidence. These methods operate by measuring the degree of semantic overlap or divergence between statements, thereby identifying whether evidence supports, contradicts, or remains unrelated to a claim. A defining feature of content-based approaches is their ability to handle paraphrases and semantic equivalence, since claims often appear in multiple forms across diverse information sources.

The earliest implementations of content-based verification relied heavily on surface-level similarity measures. Techniques such as cosine similarity and TF-IDF quantified lexical overlap between claim and evidence texts (Ferreira et al., (2018)). While computationally efficient, these methods were often limited in scope: they failed to capture subtle semantic differences or indirect contradictions where wording differed but meaning diverged. For example, the claims *“The senator voted for the bill”* versus *“The senator opposed the bill”* may appear lexically similar but are semantically opposite, a distinction that early methods frequently overlooked.

The advent of word embedding models marked a turning point. Representations such as Word2Vec, GloVe, and FastText embedded words into dense vector spaces, allowing models to capture semantic relationships that transcended surface-level word matching (Reddy et al., (2019)). This meant that paraphrased statements, for instance, *“Obama signed the healthcare act”* and *“The healthcare act was enacted by Obama”* could be recognized as equivalent even in the absence of direct lexical overlap.

The subsequent rise of transformer-based architectures revolutionized content-based verification. Pre-trained models such as BERT (Devlin et al., (2019)) and RoBERTa (Liu et al., (2019)), fine-tuned for natural language inference (NLI), significantly enhanced the detection of semantic equivalence and contradiction. These models leveraged contextual embeddings, enabling systems to interpret meaning at the phrase and sentence level. For example, BERT could correctly flag contradiction in the FEVER-style claim *“The Eiffel Tower is located in Berlin”* when paired with evidence stating, *“The Eiffel Tower is in Paris”*, despite both sentences sharing high lexical similarity.

Recent advances continue to extend the capabilities of content-based methods. Hybrid models integrate semantic embeddings with external knowledge sources to improve robustness across domains such as politics, science, and health. (Wein et al., (2022))

employed contrastive learning for semantic equivalence, Zhang et al., (2024) introduced context-aware contradiction detection to handle evolving misinformation in news, and (Chen et al., (2025)) proposed a cross-domain NLI model that leverages multi-task learning to align claims with heterogeneous evidence drawn from scientific articles, news reports, and social media posts.

As a result, these developments illustrate a clear trajectory: from early reliance on lexical overlap to embedding-based semantic modeling that captures deeper textual equivalence and contradiction. Content-based approaches remain indispensable to fact-checking frameworks, ensuring that claims are assessed not only at the surface level but also in terms of their underlying meaning. However, even state-of-the-art transformer-based systems are constrained by the limits of textual evidence; they struggle when claims require background knowledge, contextual grounding, or domain-specific information that is not explicitly present in the text. These challenges have motivated the emergence of knowledge-based approaches, which integrate structured knowledge graphs and external resources to extend verification beyond what is directly observable in content.

2.6 Knowledge-based Approach

Knowledge-based methods map claims into structured representations by linking textual elements to entities and relations in knowledge bases such as DBpedia, Freebase, or Wikidata. Enhanced entity linking techniques including multilevel encoding (Chen et al., (2021)) and context-aware typing (Dai et al., (2019)) address ambiguity, ensuring that terms like “*Apple*” are correctly identified as *Apple Inc.* rather than the fruit. Once entities are resolved, knowledge graphs (KGs) provide the relational context necessary for reasoning.

To operationalize this reasoning, Knowledge Graph Embeddings (KGEs) such as TransE and DistMult (Wang et al., (2017); Li et al., (2021)) represent entities and relations

in vector spaces, enabling efficient relational inference. More advanced interaction models like CrossE (Zhang et al., (2019)) model entity–relation interactions directly. Hybrid methods, such as logic-informed embedding models (Du et al., (2017)), further enhance interpretability by incorporating symbolic reasoning rules into embedding space.

A persistent challenge, however, is KG incompleteness and inconsistency. To address this, probabilistic and rule-based methods have been introduced. Techniques such as Bayesian reasoning (Santos & Santos, (1999)), backward chaining (Nonfjall & Larsen, (1992)), and argumentation frameworks (Bouzeghoub et al., (2017)) provide mechanisms for detecting contradictions and maintaining integrity across noisy or conflicting knowledge. Figure 2.4 illustrates the pipeline, a claim is decomposed into entities and mapped into a KG for relational verification. For instance, the claim “*Barack Obama was born in Hawaii*” is represented as $(Obama \rightarrow Person, Hawaii \rightarrow Location)$ and validated against the relation *isBornIn*. This process exemplifies how entity linking and knowledge graphs form the backbone of structured fact-verification pipelines.

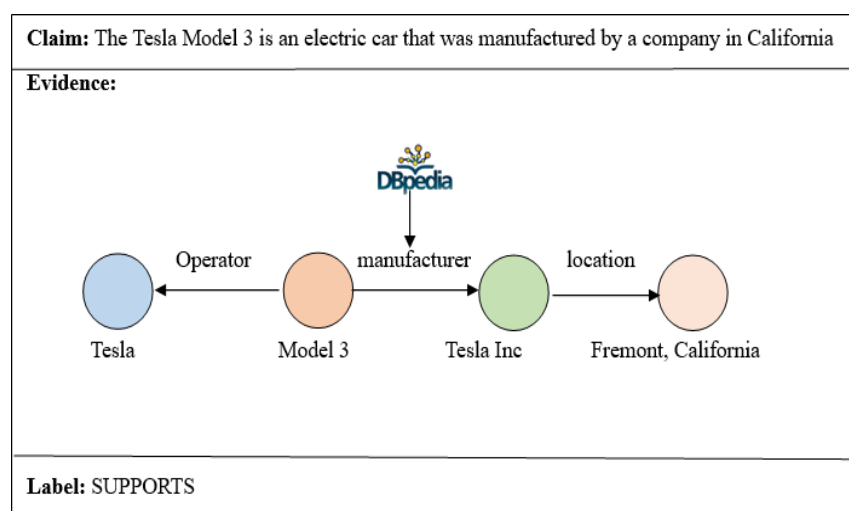


Figure 2.4: Knowledge-based Fact-Checking via Entity Linking

Source: (Kim, J et al., (2023))

Knowledge-based approaches are typically evaluated on benchmark KG datasets such as FB15K, WN18, and YAGO (Bordes et al., (2013); (Suchanek et al., (2007))), which test link prediction and fact verification accuracy. Results consistently show that embedding-based methods improve scalability (Wang et al., (2017)), while rule-augmented systems enhance interpretability (Guo et al., (2016)). Complementing KG-based reasoning, retrieval-based approaches extract relevant evidence from unstructured corpora such as Wikipedia, news archives, or scientific articles. Early lexical retrieval methods like BM25 were limited to keyword overlap, but dense semantic retrieval has since become dominant. Dense Passage Retrieval (DPR) (Karpukhin et al., (2020)), evaluated on natural Questions and TriviaQA, demonstrated that embedding-based retrieval significantly outperforms lexical search. Hybrid models (Zhang et al., (2024)) extend this by integrating dense and sparse signals. More recent advances include ANCE-PRF (Pan et al., (2024)), which improves retrieval quality via query feedback loops. These methods retrieve candidate evidence documents, which are subsequently ranked and verified for factual alignment with the claim.

At the boundary, Retrieval-Augmented Generation (RAG) (Khaliq et al., (2024)) integrates retrieval with generative models, grounding responses in retrieved evidence. Variants such as OG-RAG (Sharma et al., (2024)) constrain retrieval to ontology-aligned evidence, thereby reducing hallucinations, while KGRAG (Zhu et al., (2025)) integrates knowledge graphs into the retrieval process to enrich chunking and produce coherent summaries across domains. Figure 2.5 illustrates this workflow, a claim is encoded into dense embeddings, which guide retrieval from a large corpus. Retrieved documents are then ranked by semantic similarity and passed into fact-checking models. This pipeline

demonstrates how retrieval augments knowledge-based reasoning by expanding the evidence space to include both structured (KGs) and unstructured (textual) sources.

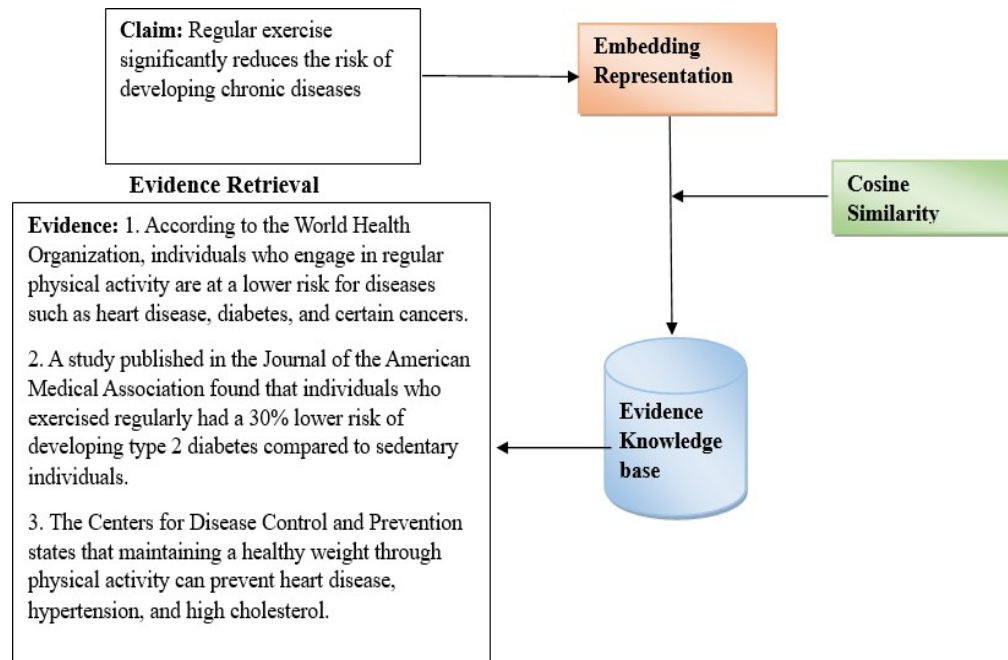


Figure 2.5: Information Retrieval using Semantics

Source: (Sharma et al. (2024) and Zhu et al. (2025))

Together, knowledge- and retrieval-based methods extend fact-checking beyond surface-level content by grounding claims in external evidence. They excel at enhancing factual precision and contextual coverage but are not without limitations. Errors in entity linking can propagate through the pipeline, incomplete knowledge graphs limit reasoning power, and dense retrieval may introduce irrelevant or noisy passages. Moreover, while these methods provide access to broader evidence, they still require higher-level integration with semantic and logical reasoning frameworks to fully disambiguate factuality in dynamic, cross-domain contexts.

These challenges have motivated the development of hybrid approaches, which combine knowledge-based reasoning, semantic embeddings, and contextual alignment to achieve more robust and explainable fact verification. The next section discusses how such hybrid systems integrate multiple axes of reasoning to overcome the limitations of individual methods.

2.7 Reasoning-based Approach

In the early years of automated fact-checking, researchers turned to logic as a way of moving beyond shallow keyword or embedding similarity. The reasoning-based approach promised something more structured: the ability to weigh claims against formal rules and to infer truthfulness through deduction. Unlike lexical or content-based techniques, reasoning introduced explicit explainability, where every inference could be traced to a logical principle. Broadly, reasoning-based approaches evolved along two paths such as, rule-based reasoning and inference-based reasoning.

2.7.1 Rule-based Reasoning

Rule-based reasoning rests on carefully crafted logical constraints that act as gatekeepers of factuality. Early contributions, such as analysis of paradoxes (Yablo, (1993)) and (Wu et al.'s (1993)) development of resolution principles, illustrated how logical rules could expose inconsistencies hidden within knowledge systems. Later, (Hussain et al., (2011)) introduced MIRC's for contradiction detection in ontologies, while (Carnielli and Rodrigues (2015)) argued that contradictions should not be treated as fatal errors but as signals of conflicting evidence. This perspective allowed reasoning-based systems to act as interpreters of information tension rather than rigid referees.

The appeal of rule-based approaches lies in their transparency: every conclusion can be justified by the underlying rule. However, hand-crafted rules proved difficult to scale in dynamic and fast-changing domains. To address this, semi-automated and hybrid solutions emerged. (Momii et al., (2023)) showed how explicit reasoning rules layered on knowledge graphs can preserve interpretability, while (Qudus et al., (2023)) explored methods to mine rules automatically from large datasets. More recent directions involve leveraging LLMs such as (Luo et al., (2023)) demonstrated how large language models can generate interpretable rules automatically, and (Nikopensius et al., (2023)) applied reinforcement learning to guide reasoning paths across graphs.

2.7.2 Inference-based Reasoning

In contrast, inference-based reasoning moves away from rigidly predefined rules and instead relies on probabilistic and argumentation frameworks to capture uncertainty and context. Bayesian networks, for instance, allowed systems to assign degrees of belief rather than binary truth values, while natural logic offered reasoning mechanisms more closely aligned with everyday language. Recent studies further advanced this direction: Zeng (2023) described hybrid probabilistic embedding systems, (Chen et al., (2025)) proposed GraphCheck, which integrates knowledge graphs as soft prompts for multi-step reasoning in LLMs, and (Huang et al., (2025)) introduced GraphFC, a graph-based framework validating claim–evidence triplets step by step. (Hammond (2025)) further demonstrated the use of Bayesian reasoning to uncover subtle interactions in biological networks principles directly applicable to fact verification.

Despite their promise, reasoning-based methods face scalability and complexity challenges. As knowledge graphs grow to billions of nodes and edges, identifying shortest reasoning paths (Ciampaglia et al., (2015)), preventing rule explosion (Gu et al., (2020)),

and mining meaningful patterns (Shi & Weninger, (2015)) become computationally demanding. Moreover, ambiguity in natural language often undermines rigid logical systems rules struggle when claims are vague, context-dependent, or domain-specific.

As illustrated in Figure 2.6, reasoning-based fact-checking integrates a pool of logical axioms with structured knowledge graphs to validate claims. The axiom pool defines relations such as transitivity, symmetry, and inverse operators, which are then applied to a knowledge graph of entities and relationships. This process enables the system to infer new relations and validate existing ones. For example, if “*hasParent*” is defined as the inverse of “*hasChild*”, the system can automatically derive missing links or verify their consistency. Verified triplets are not only lexically matched but logically validated, ensuring factual soundness at a structural level.

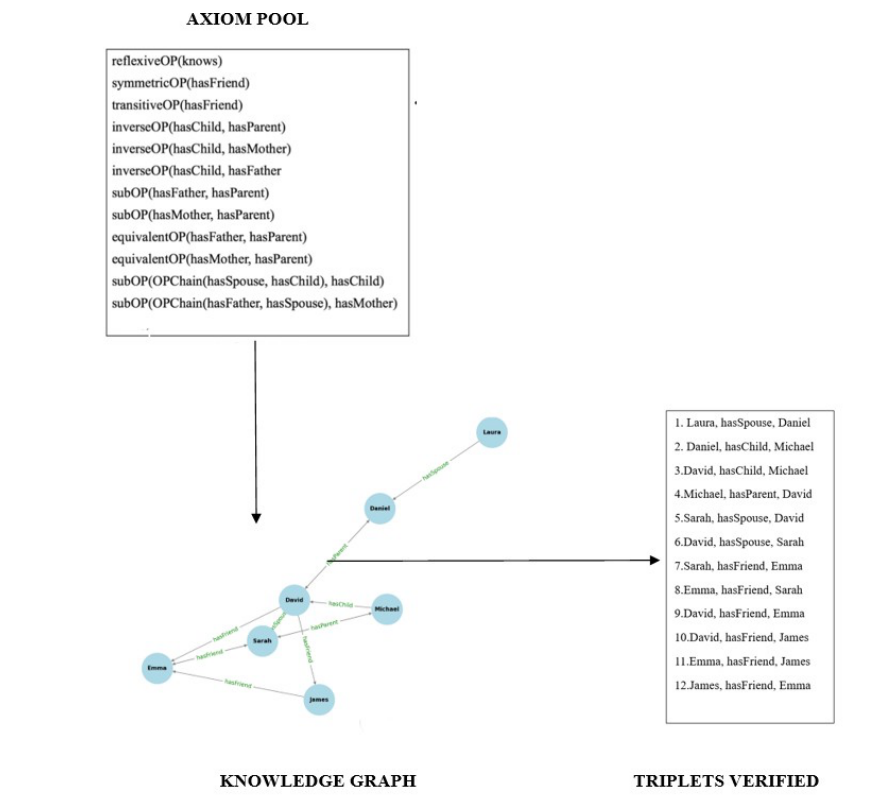


Figure 2.6: Fact-Checking using Rules-based and Logical Inferences

Source: (Momii et al., (2023))

By embedding axioms directly into graph-based reasoning, these systems preserve explainability and consistency. Each inference can be traced to a specific rule, making reasoning-based verification transparent and interpretable ((Baader et al., (2008); (Hussain et al., (2011)); (Carnielli & Rodrigues, (2015)); (Ciampaglia et al., (2015)); (Momii et al., (2023))). The trajectory of reasoning-based fact-checking reflects a shift from manual rule design toward semi-automated, hybrid neural-symbolic systems. Emerging models like ChatRule (Luo et al., (2023)) and GraphCheck (Chen et al., (2025)) show how symbolic reasoning can be fused with neural embeddings to combine transparency with adaptability. The future of reasoning-based approaches thus lies not in choosing between rules and inference, but in weaving them together into systems that reason with both rigor and resilience.

2.7.2.1 Bayesian Network for Modeling Uncertainty

Among inference-based methods, Bayesian Networks (BNs) have emerged as a cornerstone for managing uncertainty in automated fact-checking. Unlike rule-based systems, which operate in binary terms, BNs model probabilistic dependencies between variables, enabling systems to infer the likelihood of a claim given uncertain or incomplete evidence (Marcot, (2012); Yu et al., (1999)). This makes them particularly valuable in dynamic contexts where not all information is available or where evidence may conflict.

Their key strength lies in quantifying how multiple pieces of information jointly contribute to claim verification. For example, in complex misinformation scenarios, BNs can weigh conflicting signals such as partially verified statements, missing sources, or ambiguous evidence to generate a probability distribution rather than a definitive true/false outcome. However, constructing effective BN models requires extensive data and domain

expertise, which can limit scalability in fast-evolving domains (Carriger et al., (2016)).

Recent research has significantly extended their applicability. (Vink, (2024)) demonstrated how template-driven BNs streamline reasoning in forensic evidence analysis, while (Getoor & Taskar (2023)) integrated BNs with statistical relational learning (SRL) to reason over structured relational data rather than isolated variables. In parallel, efforts to handle uncertain or conflicting evidence such as soft, virtual, and contradictory inputs have enhanced BN inference reliability in real-world verification tasks (Zhou et al., (2023)). Together, these advances position Bayesian Networks as dynamic and adaptable tools for probabilistic reasoning, strengthening their role in fact-checking systems.

Inference-based reasoning also extends into argumentation frameworks and hybrid probabilistic models. Natural logic offers a way to align reasoning closer to human language, while probabilistic embeddings allow inference over graph-structured claims and evidence (Zeng, (2023)). More recently, (Chen et al., (2025)) introduced GraphCheck, which integrates knowledge graphs as soft prompts for large language models, enabling multi-step probabilistic reasoning. Similarly, (Huang et al., (2025)) proposed GraphFC, a framework that validates claim–evidence triplets step by step, while (Hammond (2025)) demonstrated how Bayesian inference principles can uncover subtle interactions in biological networks with direct implications for fact verification.

Despite their strengths, both rule-based and inference-based systems face challenges at scale. Expanding knowledge graphs to billions of nodes introduces computational bottlenecks in path search (Ciampaglia et al., (2015)), risks of rule explosion (Gu et al., (2020)), and difficulties in mining meaningful patterns (Shi & Weninger, (2015)). Ambiguity in natural language compounds the problem, where vague or context-dependent claims defy both rigid rules and probabilistic models.

To overcome these barriers, the field has shifted toward hybrid neural-symbolic reasoning. These approaches integrate symbolic logic’s interpretability with the adaptability of deep learning, forming systems that are both rigorous and scalable. Luo et al.’s ChatRule (2023) uses LLMs to generate symbolic rules dynamically, while Chen et al.’s GraphCheck (2025) shows how probabilistic reasoning embedded in graph structures can be coupled with transformer-based models for stepwise verification. This emerging synthesis represents the future of reasoning-based fact-checking: transparent, resilient, and adaptive.

Reasoning-based approaches whether through explicit rules, Bayesian inference, or hybrid neural-symbolic models demonstrate the importance of structured reasoning in fact-checking. While earlier methods struggled with scale and ambiguity, modern advances point toward systems that can balance logical rigor with probabilistic flexibility. This evolution lays the foundation for the next dimensions such as context-based approaches, which extend reasoning beyond isolated claims and evidence to the broader situational and discourse context in which truth claims operate.

2.8 Logic-based Approach

Logic-based approaches aim to capture the semantics of statements to enable common-sense reasoning in automated fact-checking. Unlike surface-level text matching, these methods apply formal rules of inference, allowing systems to derive structured and interpretable conclusions about the truthfulness of claims. A notable example is Natural Logic, demonstrated by Angeli and Manning (2014), which showed the potential of logical inference by achieving high precision in generating valid conclusions. However, such approaches face important challenges, particularly in relation to non-monotonic reasoning, where the introduction of new information can invalidate earlier conclusions (Dieussaert et al., (2004)). This difficulty underscores the complexity of maintaining consistency in dynamic environments where knowledge is constantly evolving. Furthermore, reliance on

static knowledge bases constrains adaptability, limiting the ability of logic-based systems to cope with the rapidly changing nature of real-world discourse.

Recent research has sought to overcome these limitations by extending logic-based reasoning into hybrid frameworks that combine symbolic inference with neural models. For instance, (Dalvi et al., (2021)) introduced Neuro-Symbolic Fact Verification, integrating logical rule induction with neural text encoders to improve inference over ambiguous claims. Similarly, (Madaan et al., (2022)) proposed Logic-Guided Fact Verification, where symbolic rules are used to regularize transformer models, thereby enhancing consistency under uncertain conditions. To address non-monotonicity, (Kalyanpur et al., (2023)) explored defeasible reasoning frameworks that dynamically revise conclusions when new evidence emerges, offering adaptability to evolving information contexts. More recently, (Liu et al., (2024)) demonstrated that combining Natural Logic with large language models can significantly reduce hallucination by grounding outputs in interpretable inference rules, while (Zhang et al., (2025)) introduced Temporal Defeasible Logic, explicitly modeling time-sensitive updates to knowledge for fact verification tasks.

In parallel, probabilistic reasoning frameworks have been integrated to complement symbolic methods. Bayesian Networks provide structured mechanisms for probabilistic inference, enabling systems to manage uncertainty and incomplete evidence more effectively than rigid rule-based approaches. Their ability to quantify relationships between evidence and claims makes them particularly valuable in real-world fact-checking scenarios. Together, Natural Logic, probabilistic reasoning, and modern neuro-symbolic hybrids offer a multifaceted solution to the challenges of common-sense reasoning in fact-checking. While logic-based systems continue to face constraints of scalability and adaptability, their emphasis on interpretability and rule-grounded inference ensures that they remain a critical component of automated fact-checking frameworks.

2.9 Graph-based Approach

While reasoning-based methods emphasize logical inference and probabilistic structures, graph-based approaches shift the focus toward explicit modeling of entities, relations, and paths within knowledge graphs (KGs). By structuring information in this way, fact-checking systems can move beyond surface-level text matching to leverage graph topology for pathfinding, constraint consistency, and interpretable reasoning. This structural advantage makes graph-based methods particularly effective for handling multi-hop claims, dynamic knowledge, and hybrid reasoning with large language models (LLMs).

A natural starting point for graph-based verification is pathfinding that traces the connections between claim entities. Unlike text-only approaches, graphs explicitly encode relationships, making it possible to discover multi-hop reasoning chains that support or refute a claim. Recent work illustrates this: LLM-SKAN augments a language model with structured triplets for fine-grained reasoning (Zhang et al., (2024)), while GraphFC constructs claim–evidence graphs and verifies them step by step, improving interpretability (Huang et al., (2025)). These advances demonstrate how path-based models expose the logic behind decisions rather than producing opaque predictions. However, real-world claims often evolve over time, meaning that reasoning chains cannot be static. This motivates the need for temporal and hyper-relational KGs.

To address the dynamic nature of knowledge, researchers have extended KGs with temporal and qualifier-rich edges. For example, TemporalFC incorporates time-aware knowledge to assess evolving claims (Kim et al., (2023)), while Findings of EMNLP 2024 proposed methods for reasoning over hyper-relational graphs, capturing contextual details like time, source, or location. Reinforcement learning methods also support temporal multi-hop reasoning, enabling systems to explore sparse time-stamped graphs. While temporal

reasoning enhances adaptability, scaling these approaches to massive graphs introduces computational challenges, hence the efficient retrieval mechanisms have become central.

As KGs expand into billions of nodes, brute-force reasoning is impractical. Systems such as VKGFR (Li et al., (2024)) address this by constructing *virtual knowledge graphs* lightweight, query-specific subgraphs that enable fast retrieval without sacrificing accuracy. Such designs significantly improve throughput, especially in large-scale settings like Wikipedia verification. Yet even efficient retrieval is not sufficient if entity references are ambiguous. This has led to a stronger emphasis on entity disambiguation and heterogeneous graph construction. Ambiguous or overlapping entities frequently mislead verification systems. To mitigate this, pipelines like VeGraph decompose claims into triples, apply graph-guided entity disambiguation, and then verify the refined structures (Wang et al., (2024)). Further, heterogeneous graph models integrate sentence-level evidence with KG concepts such as YAGO and Wikidata, allowing systems to reason jointly over structured and unstructured sources. While these methods enrich graph reasoning, recent trends show that the greatest promise lies in combining symbolic KG structure with the adaptive reasoning of LLMs.

A growing research direction unifies the rigor of KGs with the flexibility of LLMs. In such hybrids, KGs serve as *soft prompts* or *constraints* that reduce hallucination and ground reasoning in structured evidence. For example, (Chen et al. (2025)) demonstrated that integrating KG constraints into LLM reasoning improves both accuracy and interpretability in multi-hop verification. Surveys in this space highlight the synergy of KG structure and neural reasoning, particularly for complex, open-domain claims. Taken together, these directions illustrate a common workflow for graph-based verification. In practice, graph-based verification as illustrated in Figure 2.7 relies heavily on pathfinding algorithms that identify both positive and negative reasoning chains between entities.

Positive paths represent supportive evidence such as correctly linking Malcolm Gladwell as the author of *David and Goliath* or *Outliers* while negative paths expose contradictions, for example, misclassifying *David and Goliath* as fiction. By explicitly distinguishing between supporting and contradictory links, pathfinding provides interpretable reasoning trails that not only strengthen verification accuracy but also enhance transparency in automated fact-checking systems.

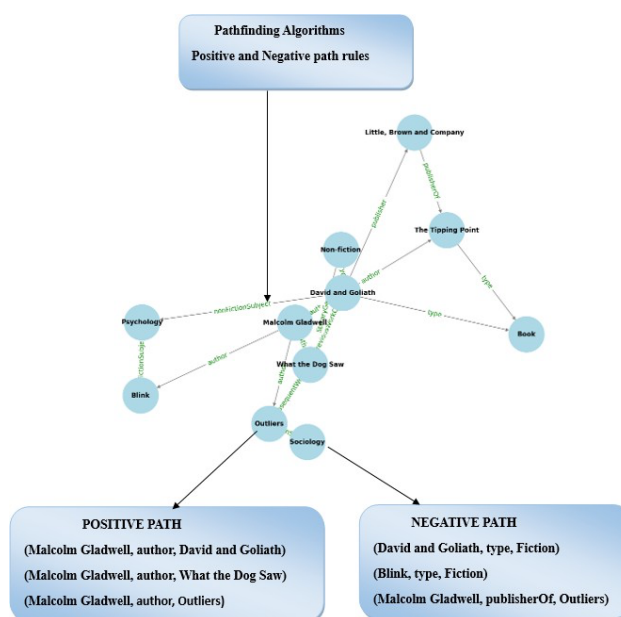


Figure 2.7: Pathfinding in Knowledge Graph Fact Verification

Source: (Kim et al., (2023) and Zhang et al., (2025))

2.9.1. Graph Embeddings and Reasoning Techniques

Graph embeddings translate knowledge graphs into low-dimensional vector spaces, enabling computational reasoning for fact-checking, link prediction, and knowledge graph completion. Early approaches such as TransE (Bordes et al., (2013)) modeled relationships as translations in vector space, offering scalability and efficiency for large graphs. However, they struggled with complex relations like one-to-many and circular dependencies. To address this, RotatE (Sun et al., (2019)) extended the paradigm by representing relations as rotations in complex space, capturing patterns such as symmetry,

inversion, and hierarchy. This marked a shift from simple translation-based methods to models capable of handling more nuanced relational structures.

Building on these foundations, researchers began tackling real-world challenges of uncertainty. For instance, Uncertain Knowledge Graph Embeddings (UKGE) (Chen et al., (2018); (Yang et al., (2021))) introduced confidence-aware embeddings to manage noisy or incomplete facts. In parallel, Heterogeneous Graph Neural Networks (HGNNs) (Li et al., (2021)) employed attention mechanisms to integrate diverse semantic relations, while tensor-based embeddings (Padia et al., (2019)) leveraged factorization methods to capture deeper relational dependencies. These approaches collectively expanded embeddings beyond deterministic links, enabling reasoning in more complex, multi-domain verification scenarios.

The next wave of research emphasized robustness and temporal reasoning. Methods such as CosG (Zhang et al., (2022)) applied contrastive learning to better distinguish claim–evidence pairs, while temporal models like TemporalFC (Qudus et al., (2023)) and DNCL (Feng et al., (2025)) incorporated time-sensitive embeddings and noise-aware training to verify claims that evolve over time. By explicitly modeling temporal dynamics, these frameworks made embeddings more adaptive to the fluid nature of real-world information.

More recently, advances have focused on hybrid and LLM-integrated systems. Ensemble methods such as HybridFC (Qudus et al., (2024)) and retrieval-driven pipelines like Fact or Fiction (Opsahl, (2024)) demonstrated how subgraph retrieval and embedding reasoning can be combined to scale fact verification. Pushing further, GraphCheck (Chen et al., (2025)) embedded knowledge graphs directly into large language model prompts, reducing hallucination while enhancing multi-hop reasoning across long-form text. These innovations illustrate how embeddings now serve not only as compact graph representations but also as mediators between structured graph reasoning and neural fact-checking.

In summary, this evolution from translation-based embeddings to uncertainty-aware, temporal, and hybrid neuro-symbolic systems which highlights the central role of embeddings in bridging structured knowledge graphs with automated reasoning for claim verification. Figure 2.8 illustrates this progression, showing how embeddings derived from methods such as TransE and RotatE feed into higher-level reasoning modules, ultimately enabling scalable and interpretable fact-checking

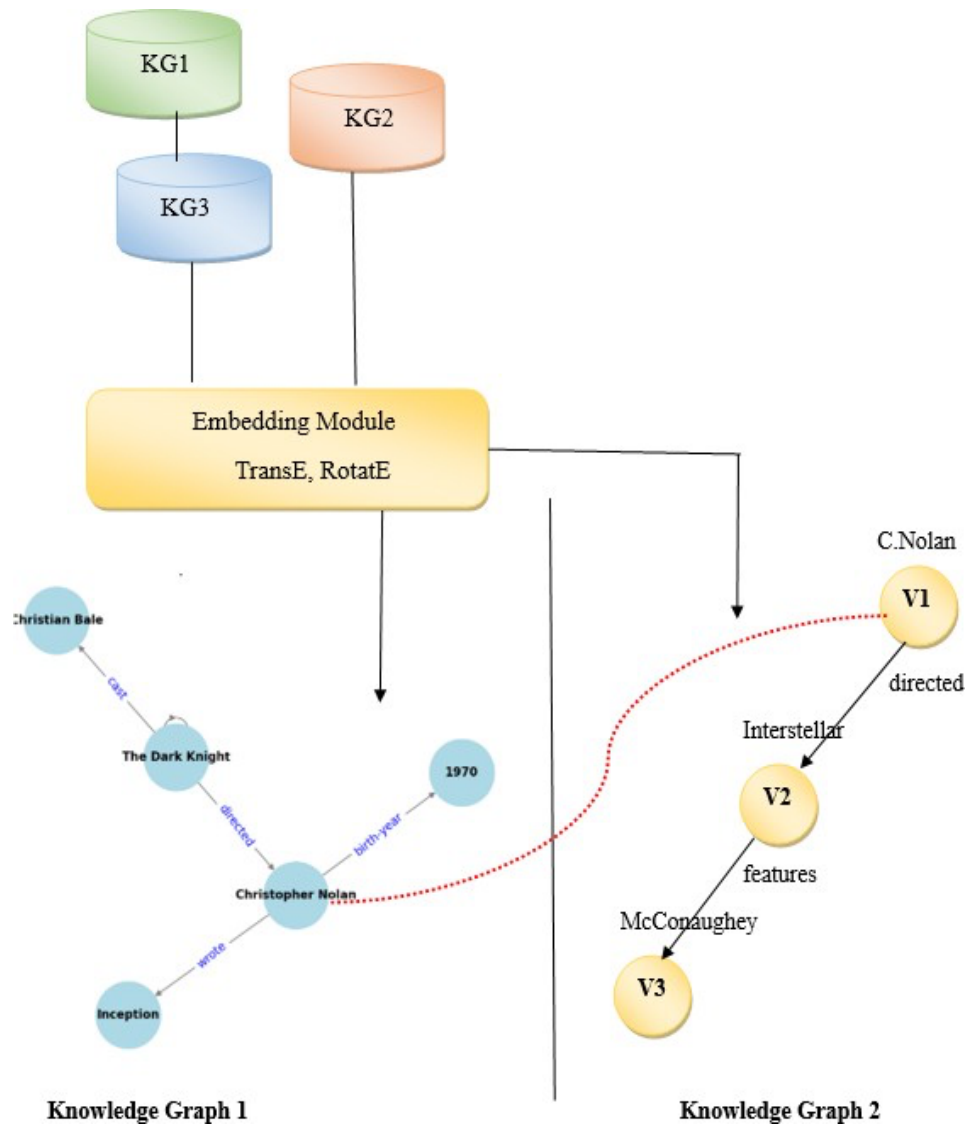


Figure 2.8: Knowledge Graph Embedding Fact Checking

Source: (Li et al., (2021), Zhang et al., (2022))

2.10 Context-aware based Approach

The integration of contextual information is increasingly recognized as a cornerstone of automated fact-checking. Unlike surface-level text matching, context-aware approaches incorporate discourse structures, temporal cues, and metadata to produce more accurate and trustworthy verification results. Early studies highlighted how context fundamentally shapes truth perception. (Cohen et al., (2015)) demonstrated through computational experiments that incorporating contextual weighting significantly improves classification outcomes, showing that information is interpreted differently depending on situational cues rather than surface features alone. In a complementary direction, Van (Bavel et al., (2016)) provided empirical evidence from social psychology, illustrating that contextual framing such as political identity or group alignment that strongly influences whether individuals perceive information as true. Together, these works established the foundation for recognizing context as pivotal in fact-checking, inspiring subsequent computational models to explicitly capture such factors.

Recent neural-based approaches have extended this principle by embedding contextual signals into fact-checking pipelines. For example, (Atanasova et al., (2019)) developed neural models for political discourse, while (Samarinas et al., (2021)) proposed Quin+, a system that improved retrieval performance even under noisy conditions. Temporal reasoning has also advanced significantly (Allein et al., (2023)) showed that grounding claims within shared timelines leads to higher verification precision, especially in dynamic misinformation environments where narratives evolve quickly.

Metadata-based methods further strengthen claim verification by integrating credibility signals and contextual cues. (Petroni et al., (2015)) introduced the CORE model, leveraging source credibility in verification, while (Prasath et al., (2013)) explored the

contextual tagging techniques to capture auxiliary features. More recent work, such as (Groza & Katona (2022)), emphasizes explainability by linking verdicts with interpretable contextual evidence, a point reinforced by (Nakov et al., (2021)) and (Gupta et al., (2023)) who argue that transparency and interpretability are critical in real-world misinformation detection.

Emerging research between 2022 and 2024 highlights the growing importance of multi-modal contextual signals. For example, (Shaar et al., (2022)) incorporated conversation threads for social media claims, while (Thorne et al., (2023)) showed that modeling temporal evolution signals boosted verification performance by up to 15% of Macro-F1. More recently, (Alam et al., (2024)) demonstrated that multimodal metadata such as text, images, and user history enhances misinformation detection robustness, bridging the gap between textual fact-checking and real-world misinformation ecosystems. As illustrated in Figure 2.9, discourse analysis linguistic framing, temporal information such as timelines, dates and metadata like source credibility and auxiliary context are jointly analyzed to refine claim verification. This layered integration ensures that models capture not only the literal meaning of a claim but also the surrounding circumstances in which it was made.

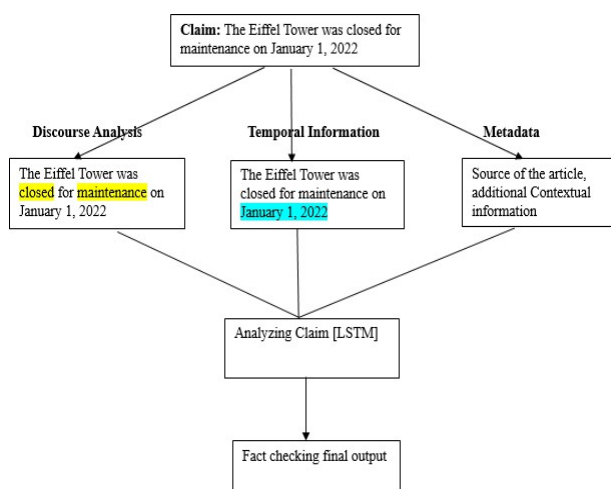


Figure 2.9: Incorporation of Temporal Context Information for Veracity Assessment

Source: ((Atanasova et al., (2019)); (Allein et al., (2023)); (Alam et al., (2024))

Despite these advancements, several challenges persist. A key limitation lies in the integration of multiple parts of evidence such as existing models often evaluate evidence in isolation, making it difficult to reconcile supportive, contradictory, or partially aligned findings into a coherent judgment. Interpretability also remains a concern, particularly when verification must adapt across dynamic and cross-domain contexts. In addition, many approaches narrowly emphasise specific contextual dimensions, such as temporal or discourse cues, while overlooking broader socio-cultural and psychological factors that shape truth perception (Atanasova et al., (2019)). Subtle language phenomena including sarcasm, irony, and metaphor continue to present difficulties for automated systems (Prasath et al., (2013)). Finally, reliance on static knowledge bases constrains adaptability in fast-evolving misinformation environments, especially in multimodal settings where text interacts with video, images, or other sources of context (Groza & Katona, (2022)). These contextual approaches laid the groundwork for the next wave of fact-checking techniques, where pretrained transformer-based embeddings such as BERT and its variants extended context-awareness by modelling semantic alignment at a much deeper level.

2.10.1 Transformer-based Embeddings for Fact-Checking

A major advancement in context-aware verification came with the introduction of transformer-based embeddings, particularly the BERT family of models. Earlier static word embeddings such as Word2Vec (Mikolov et al., (2013)) and GloVe (Pennington et al., (2014)) produced a single fixed vector for each word regardless of its context. While effective in capturing semantic similarity at a surface level, they failed to model contextual features, a limitation especially problematic in fact-checking, where the same term may shift in meaning across political, social, or biomedical domains.

To address this, (Devlin et al., (2019)) introduced BERT (Bidirectional Encoder Representations from Transformers), which employed a bidirectional pretraining strategy

based on Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). Trained on large corpora including BooksCorpus and Wikipedia, BERT generated contextualised embeddings that vary depending on sentence context. This allowed richer semantic representations of claim–evidence pairs, making BERT particularly effective in tasks such as textual entailment and sentence similarity, both central to claim verification.

Building on BERT, (Liu et al., (2019)) proposed RoBERTa, which removed the NSP objective, introduced dynamic masking, and scaled training with larger datasets. These modifications led to significant performance gains on benchmarks, demonstrating more robust semantic alignment. In fact-checking pipelines, RoBERTa proved especially useful for improving retrieval accuracy when evidence is expressed through paraphrases or semantically complex structures.

For domain-specific applications, specialised variants were developed. (Lee et al., (2020)) introduced BioBERT, pretrained on PubMed abstracts and PMC full-text articles, showing marked improvements in biomedical fact verification by capturing domain-specific terminology and semantic relations. Similarly, (Beltagy et al., (2019)) presented SciBERT, pretrained on a large corpus of scientific publications, which enhanced evidence retrieval and reasoning performance in technical domains such as scholarly claim validation. These adaptations highlight the importance of domain-sensitive embeddings for ensuring accuracy in specialised fact-checking contexts.

Beyond retrieval, BERT-based models are widely used for veracity classification, labelling claims as *supported*, *refuted*, or *not enough information (NEI)*. For instance, (Hanselowski et al., (2019)) demonstrated that BERT substantially outperformed earlier baselines such as TF-IDF, Jaccard Similarity, and TransE embeddings in the FEVER dataset, particularly in cases requiring multi-hop reasoning. Similarly, (Soleimani et al., (2020)) reported improved F1-scores when applying BERT-based architectures for evidence

ranking and claim classification, further confirming their superiority over shallow lexical and symbolic methods.

However, despite these advances, limitations remain. Standard BERT-based methods typically treat evidence sentences in isolation, without integrating multiple strands of supportive, contradictory, or conditional evidence. As a result, they risk overestimating sufficiency or collapsing contradictory findings into a single outcome, thereby reducing interpretability in contested domains. This limitation underscores the need for more advanced frameworks. Building on these limitations of transformer-based models, recent research has turned toward Large Language Models (LLMs), which extend contextual representation with generative reasoning and multimodal adaptability, offering new opportunities for advancing automated fact-checking.

2.11 Large Language Model in Fact checking

Building on the foundations of context-aware methods, Large Language Models (LLMs) represent a paradigm shift in automated fact-checking. Unlike earlier approaches that relied on manually engineered features, graph embeddings, or contextual tagging, LLMs leverage massive pretraining on diverse corpora to capture linguistic nuance, contextual reasoning, and world knowledge within a unified representation. This makes them particularly effective in addressing the complexity, ambiguity, and dynamism of real-world misinformation.

The transformative role of LLMs lies in their ability to enhance a range of Natural Language Processing (NLP) tasks central to verification, including semantic similarity, question answering, retrieval, and summarization. For example, conversational search has been significantly advanced through query rewriting, with (Galimzhanova et al., (2023))

demonstrating notable improvements in retrieval accuracy. In domain-specific contexts such as medical fact-checking, the *expand-guess-refine* alignment strategy introduced by (Manathunga and Hettigoda (2023)) boosted performance on clinical datasets like USMLE. Similarly, prompt optimization has improved the contextual faithfulness of LLM outputs (Zhou et al., (2023)), while frameworks that combine LLM capabilities with logical reasoning and temporal verification have enhanced accuracy in verifying dynamic claims (Li et al., (2023)).

To address persistent issues such as hallucinations and flawed reasoning, self-correction mechanisms incorporating automated feedback loops have been explored (Pan et al., (2023)). Furthermore, efficiency gains have been achieved through smaller fine-tuned models, which deliver results comparable to large-scale systems such as GPT-3.5. (Zha et al., (2023)) reported up to a 31.7% increase in precision for fact-checking tasks using these compact models, highlighting their suitability for real-time, scalable verification workflows.

Recent research has extended these advancements in several directions. (Guo et al., (2024)) introduced *Fact-LLM*, a hallucination-aware benchmark that evaluates LLM robustness in verification tasks. Shen et al., (2024) focused on prompt engineering to detect implicit claims, a long-standing challenge in misinformation analysis. (Chakraborty et al. (2024)) proposed *LLMFact*, enabling cross-domain and multimodal verification, while (Huang et al., (2025)) integrated external knowledge graphs with LLM reasoning to enhance reliability in misinformation detection. More recently, (Lee et al., (2025)) developed adaptive LLM-based agents capable of real-time verification in news and social media streams, pushing LLM integration closer to practical deployment.

As illustrated in Figure 2.10, LLMs enhance fact-checking by pairing claims with supporting evidence and identifying error categories such as entity misrepresentation, fabricated events, flawed reasoning, and irrelevant evidence. This structured error analysis

highlights the interpretability and transparency LLMs bring to verification tasks that addresses one of the major shortcomings of earlier methods.

Claim: Bill Gates stated in an interview that vaccines are a form of population control
Evidence: Bill Gates has publicly advocated for vaccines as a critical tool in improving global health, particularly in developing countries. He has spoken extensively about reducing child mortality through vaccination, leading to healthier families. In a 2011 TED Talk, Gates mentioned that as health improves and child mortality decreases, families tend to have fewer children. However, there is no credible evidence that Gates ever claimed vaccines are a tool for population control in a harmful or conspiratorial sense. Fact-checking platforms like Snopes and Reuters have debunked this false narrative.
Error and Explanation Entity Error: Steve Jobs is not involved in this issue, and the correct entity is Bill Gates. This is a clear entity misrepresentation. Event Error: This event is completely fabricated and not supported by any credible evidence. Reasoning Coherence Error: The explanation lacks logical flow, leading to flawed reasoning. Irrelevant Evidence Error: While the statement about Gates' wealth is true, it is irrelevant to the specific claim about his views on vaccines and population control.

Figure 2.10: Large Language Model Integration in Fact-Checking
Source: (Chen et al., (2023))

In summary, LLMs have advanced fact-checking from surface-level verification toward explainable, adaptive, and trustworthy reasoning frameworks. Their ability to integrate contextual understanding, temporal reasoning, and multimodal evidence underscores their central role in the next generation of misinformation detection systems.

2.12 Policy-Level Approaches to Misinformation

Beyond computational techniques, governments and international institutions have also implemented strategies to combat misinformation at the national and global levels. These initiatives highlight the societal importance of fact-checking but also expose persistent challenges in managing evolving and contradictory evidence.

At the global scale, organizations such as the World Health Organization (WHO) have spearheaded efforts to counter health-related misinformation, particularly during the

COVID-19 pandemic. The WHO's *myth-busting* campaigns and collaboration with digital platforms such as Google, Facebook, and Twitter aimed to provide verified information to the public in real time (WHO, 2021). Similarly, the European Union's Code of Practice on Disinformation (EU Commission, 2022) established a multi-stakeholder framework requiring technology companies to improve transparency in political advertising, disrupt financial incentives for disinformation, and provide researchers with access to platform data. These initiatives highlight a regulatory shift towards accountability in mitigating misinformation.

At the national level, governments have established specialized task forces and fact-checking units. For instance, the United States Cybersecurity and Infrastructure Security Agency (CISA) launched campaigns to safeguard elections from disinformation, while the UK's Rapid Response Unit focused on countering viral falsehoods in public communication (Cabinet Office, 2021). Similarly, ASEAN countries have formed regional collaborations such as the *ASEAN Fact-Checking Network* to share resources in combating misinformation across languages and borders (ASEAN, 2023).

Despite their significance, these institutional approaches face limitations. Political and cultural contexts often shape what is labeled as misinformation, raising concerns of bias and censorship. Moreover, manual verification by government task forces cannot keep pace with the speed and scale of misinformation spread on social media. These challenges underscore the need for automated, explainable fact-checking frameworks that complement institutional efforts by offering scalable and transparent verification pipelines.

In summary, while institutional and policy-level interventions provide critical governance and accountability mechanisms, they remain constrained in scalability and adaptability. This reinforces the necessity of computational frameworks, such as the one

proposed in this thesis, which can dynamically assess evidence sufficiency, logical coherence, and cross-domain consistency.

2.13 Summary of Research Gap

The review of prior approaches to automated fact-checking highlights significant advancements across ontology-based, structure-driven, lexical, semantic, graph-based, and LLM-powered methods. While each approach has contributed meaningfully to the field, two persistent challenges remain unresolved. First, existing systems struggle to dynamically merge evolving or contradictory evidence due to the absence of a non-monotonic reasoning framework. Current models often treat evidence as static, limiting their ability to adapt to real-world claims where information is continuously updated and contested. Second, there is a lack of a coherent factuality model capable of handling uncertainty. Methods that rely heavily on static features, lexical overlap, or shallow heuristics are prone to inconsistencies when confronted with ambiguous, incomplete, or conflicting evidence.

Figure 2.11 illustrates this synthesis by mapping research findings to their associated limitations, revealing the two central gaps, the lack of dynamic reasoning for evolving and contradictory evidence, and the inability to model factuality under uncertainty.

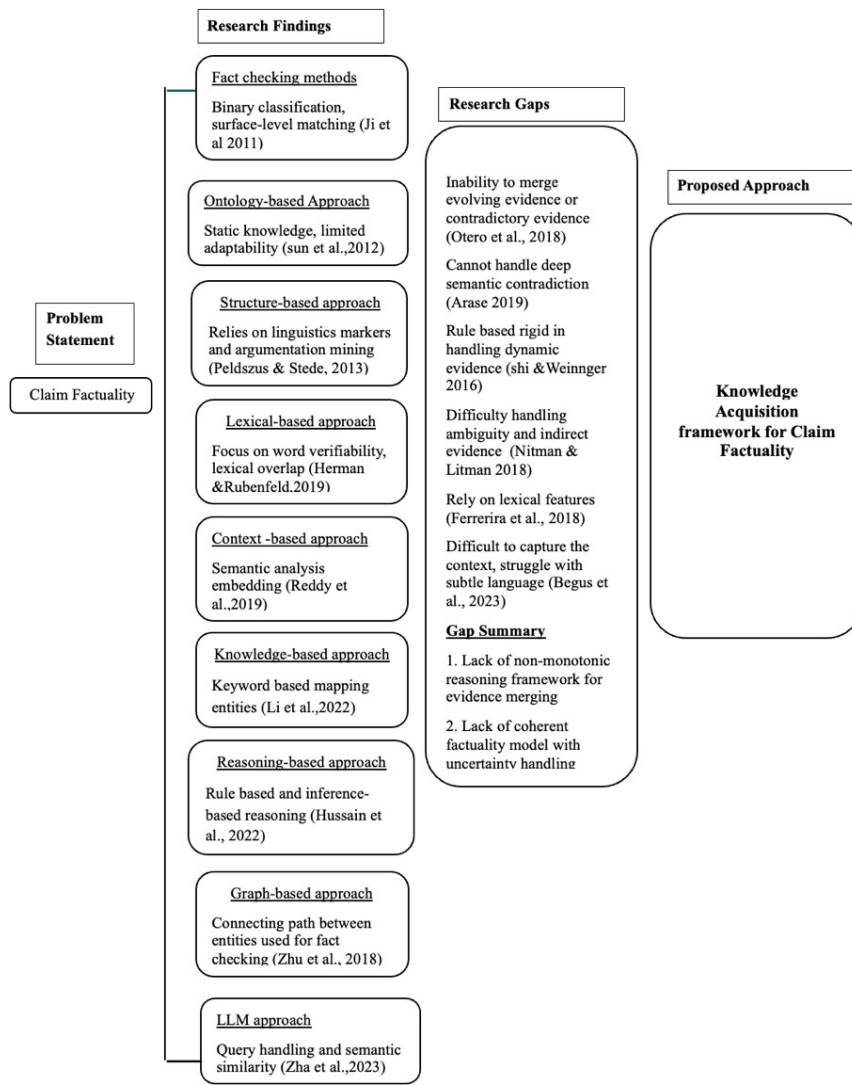


Figure 2.11: Literature Review Mapping

To address these gaps, this thesis proposes a Knowledge Acquisition Framework for Claim Factuality, built on three integrated mechanisms that collectively strengthen automated fact-checking. The Knowledge Acquisition Mechanism enables continuous updating and enrichment of evidence sources, ensuring that the system adapts to evolving claims and dynamic contexts. The Factuality Disambiguation component resolves contradictions and effectively manages uncertainty, providing a more coherent assessment of claim validity. Finally, the Hybrid Evaluation Framework aligns contextual, logical, and semantic reasoning to deliver balanced and reliable veracity judgments. Together, these mechanisms ensure adaptability to evolving information, robustness against contradictory evidence, and reliability in uncertainty-laden environments, thereby advancing the state of automated fact-checking.

2.14 Chapter Summary

This chapter provided a comprehensive review of the existing literature on automated fact-checking, tracing the progression of methods from traditional lexical matching to more advanced ontology-based, structure-driven, semantic, graph-based, and large language model (LLM)-driven frameworks. Each approach has contributed important advancements, particularly in areas such as semantic similarity, evidence retrieval, discourse-level reasoning, and knowledge graph construction. However, the review revealed that despite these contributions, fact-checking systems remain constrained in their ability to handle the complexity, dynamism, and contradictions of real-world misinformation.

One recurring challenge lies in the reliance on shallow methods such as keyword matching, basic semantic similarity, or static knowledge bases that fail to capture deeper contextual and discourse-level relationships between claims and evidence. While ontology-based and structure-based approaches improved logical consistency, their rigid schemas often lack the flexibility needed to address the fluid nature of evolving claims. Similarly, semantic and neural approaches enhanced contextual modeling but struggled with contradictions, sarcasm, and ambiguity in evidence. Even knowledge-graph-based methods, though powerful for multi-hop reasoning, face limitations when dealing with uncertainty or dynamically evolving evidence sources.

The review also highlighted the transformative role of recent LLM-based methods, which unify linguistic nuance, world knowledge, and contextual reasoning. These models significantly improve retrieval, reasoning, and claim verification. Yet, even LLMs face challenges in handling hallucinations, implicit claims, multimodal misinformation, and domain adaptation. Emerging benchmarks and hybrid approaches indicate promising directions, but also reinforce that no single method fully resolves the intertwined challenges of evolving evidence, logical coherence, and uncertainty.

Synthesizing these insights, two critical gaps remain unresolved such as the inability of current systems to dynamically merge evolving or contradictory evidence due to reliance on monotonic reasoning, and the lack of coherent models to assess factuality under uncertainty, leading to inconsistencies in complex, real-world scenarios. Addressing these gaps requires frameworks that can adapt to new evidence, manage contradictions, and ensure coherent factuality judgments even in uncertainty-laden environments.

To overcome these limitations, this thesis proposes a Knowledge Acquisition Framework for Claim Factuality. The framework integrates three complementary mechanisms such as Knowledge Acquisition, which enables continuous updating and enrichment of evidence sources, Factuality Disambiguation, which resolves contradictions and insufficiency in claim-evidence relationships; and a Hybrid Evaluation Framework, which aligns contextual, logical, and semantic reasoning for robust factuality assessment. Together, these mechanisms ensure adaptability to evolving claims, resilience against contradictory or incomplete evidence, and reliability under uncertainty.

In conclusion, the insights drawn from the literature review not only demonstrate the significant progress achieved in automated fact-checking but also establish the necessity of a more dynamic, adaptive, and integrated approach. These findings lay the foundation for the methodology chapter, which details how the proposed Knowledge Acquisition Framework is operationalized to address the identified research gaps and advance the state of claim factuality disambiguation.

CHAPTER 3

METHODOLOGY

3.1 Overview of the Methodology

The proposed methodology was derived directly from the research gaps identified in the literature review, particularly the limitations of existing fact-checking approaches in handling insufficient evidence, contradictions, and uncertainty. Traditional methods often relied on surface-level semantic overlap and rigid rule-based reasoning, which proved inadequate in dynamic, real-world contexts where claims evolve and evidence spans multiple domains. Motivated by these research gaps, this thesis introduces a Knowledge Acquisition Framework for Claim Factuality Disambiguation, designed as a multilayered approach that integrates deep semantic modeling, knowledge representation, evidence sufficiency with a logical reasoning mechanism.

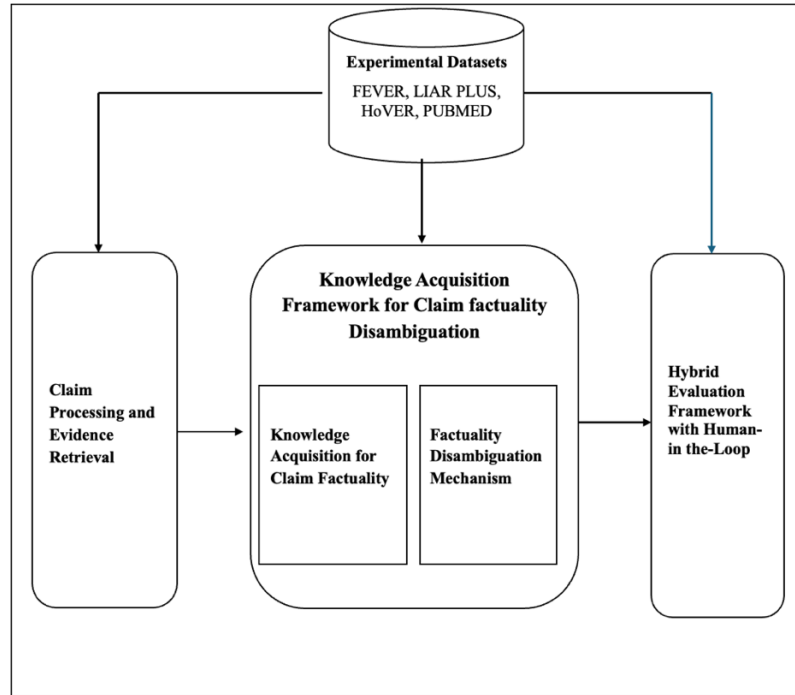


Figure 3.1: Proposed Knowledge Acquisition Framework
For Claim factuality disambiguation

As shown in Figure 3.1, the framework is organized into three interconnected stages such as (1) claim processing and evidence retrieval, (2) knowledge acquisition framework, and (3) hybrid evaluation with human-in-the-loop validation. Together, these stages create a structured process that can adapt to the complexities of fact verification while maintaining both scalability and interpretability.

The first stage establishes the foundation of the framework by identifying and structuring potential evidence for a given claim. Here, unstructured claims are transformed into machine-readable form, and semantically relevant evidence is retrieved with high coverage. Transformer-based language models such as BERT, RoBERTa and BioBERT are employed to parse claims and compute contextual embeddings that capture deep semantic alignment rather than relying on surface-level keyword overlap. Entity linking is applied to reduce ambiguity and ensure cross-domain consistency by mapping references to canonical entities.

The processed claims and retrieved evidence are then structured into a knowledge graph, constructed using Neo4j, where claims, entities, and evidence are represented as nodes and relationships. This structured representation is critical because it provides the basis for reasoning and enrichment in the next stage. Once the evidence is retrieved and organized, the framework transitions toward evaluating not just the existence of evidence but also its sufficiency and coherence.

The next stage is knowledge acquisition and factuality disambiguation which represents the core reasoning engine of the framework, where the system moves beyond evidence retrieval to refine and interpret knowledge for factual assessment. It begins with knowledge acquisition, in which the evidence space is dynamically enriched by extracting only the most essential and contextually relevant information. Through the use of contextual

attention mechanisms, the system explores the evidence neighborhood of each claim, identifies meaningful connections, and filters out noisy or irrelevant data. This ensures that only evidence with strong semantic alignment is retained, forming a robust foundation for deeper analysis. Once this essential knowledge is consolidated, the process advances to factuality disambiguation, where the sufficiency, coherence, and reliability of the evidence are systematically evaluated. Here, semantic similarity measures are combined with rule-based reasoning to address contradictions, manage uncertainty, and reconcile information originating from different domains. This integrated process allows the framework to reach a clear determination of whether a claim is Supported, Refuted, or Not Enough Information. The outcome of this stage is therefore a refined set of evidence paths that are both semantically precise and logically consistent, providing a solid basis for the final hybrid evaluation stage, where robustness and interpretability are further ensured.

The third stage ensures the reliability and transparency of the system’s factuality judgments. Automated evaluation metrics such as precision, recall, accuracy, and F1-score are used to assess the performance of each component. However, full reliance on automation risks misclassification in cases of insufficient, contradictory, or ambiguous evidence. This challenge is addressed by selectively incorporating a human-in-the-loop mechanism.

Importantly, human validation is not applied to all claims but only to those that fall into a “gray zone” of uncertainty. When millions of claims are processed, the majority are handled automatically by the pipeline, while only borderline cases are escalated to expert review. Human involvement thus serves two roles such as validating edge cases for reliability and providing iterative feedback that refines thresholds, rule-based reasoning, and contextual attention mechanisms over time. This design ensures scalability while retaining expert oversight where it matters most.

The outcome of this stage is a factuality decision that classifies each claim as Supported, Refuted, or Not Enough Information (NEI), accompanied by the relevant strands of evidence that justify the decision. To achieve this, the methodology integrates a range of state-of-the-art platforms and frameworks. Hugging Face Transformers and PyTorch are employed for model training and deployment, while Neo4j provides graph-based knowledge representation and reasoning. Amazon SageMaker supports large-scale experimentation, model optimisation, and evaluation, and supporting libraries such as spaCy enable entity recognition and linking. Python-based pipelines orchestrate preprocessing, semantic similarity computation, and evidence ranking, ensuring that the entire process is both modular and scalable.

Importantly, the outputs of this technically sophisticated pipeline are designed with end-user interpretability in mind. Rather than opaque predictions, the framework produces structured factuality decisions such as Supported, Refuted, or NEI alongside the key evidences that underpin the outcome. This design ensures transparency by exposing not only the decision but also the reasoning behind it. Furthermore, when human validation has contributed to refining edge cases, this intervention is explicitly flagged, providing additional assurance of reliability. By combining advanced computational infrastructure with explainable outputs, the framework balances technical rigour with usability, making it trustworthy to both technical and non-technical stakeholders.

3.2 Stages of Methodology

3.2.1 Problem definition

This study addresses the critical challenge of claim factuality within the fact-checking process. The primary objective is to develop a novel framework that acquires the essential knowledge required for effectively map claims to corresponding evidence while

addressing the disambiguation of discrepancies that often arise between claims and evidence. The framework ensures that conflicting or incomplete evidence is properly identified and resolved, facilitating the accurate alignment of claims with the most relevant evidence. Additionally, it enhances the system’s ability to adapt to evolving information and complex, real-world contexts, providing a robust and reliable approach to the challenges posed by misinformation.

3.2.2 Claim Processing and Evidence Retrieval

3.2.2.1 Claim Parsing and Entity Identification

The claim processing and evidence retrieval stage forms the foundation of the framework, ensuring that unstructured claims are systematically transformed into structured representations that can be analyzed computationally. As illustrated in Figure 3.2, this stage begins with claim parsing and entity identification. Constituency parsing is applied to decompose the claim into its core components such as subjects, predicates, and objects. Unlike keyword-based parsing, which often oversimplifies complex or comparative statements, constituency parsing captures the deeper syntactic dependencies that strengthen claim meaning.

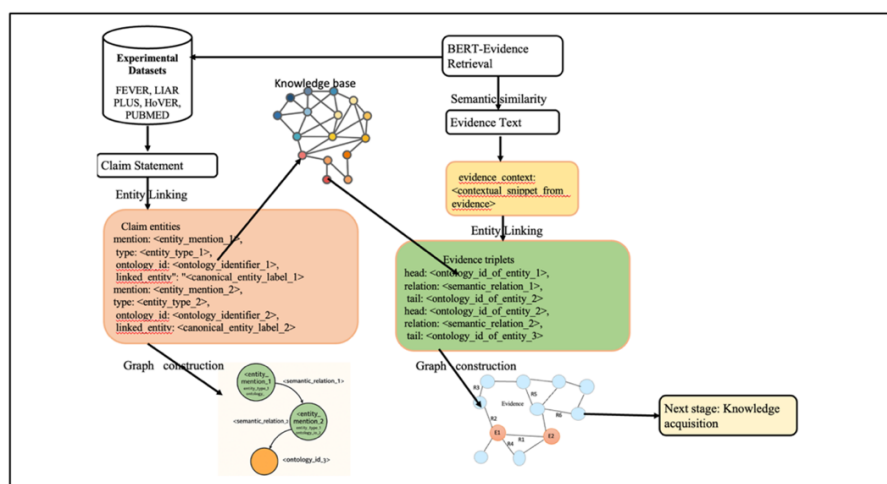


Figure 3.2: Evidence Retrieval and Representation

The extracted entities are then normalized and linked to canonical identifiers within an ontology, ensuring semantic consistency. For instance, the mention “*heart attack*” is mapped to its canonical identifier “*myocardial infarction*.” This structured representation includes the entity mention, type, ontology identifier, and linked canonical entity. The pseudocode in Figure 3.3 demonstrates this process, that begins with the parsing of claims into structured units. Validation against annotated subsets confirmed high reliability in both entity detection and normalization.

Once the claim has been parsed and normalized, the next operation involves retrieving relevant evidence from benchmark datasets such as FEVER, LIAR-PLUS, HoVer, and PubMed. This step, also shown in Figure 3.3, leverages BERT-based contextual embedding models to retrieve evidence text that is semantically aligned with the claim. Within the pseudocode, this is formalized through the functions `RETRIEVE_CANDIDATES()` and `RERANK_CANDIDATES()`. The first function retrieves a broad pool of candidate passages by aligning the semantic meaning of the claim with the corpus. This ensures that even lexically distinct but semantically related expressions are retrieved, for example, aligning “hypertension” with “high blood pressure”.

The second function, `RERANK_CANDIDATES()`, then refines this pool through pairwise scoring, comparing each candidate more closely with the claim. This stage eliminates loosely related passages and retains only those that strongly align contextually. To further prevent semantic drift, an additional filtering layer is introduced, combining cosine similarity thresholds with domain-specific keyword checks. Together, this two-step retrieval and re-ranking mechanism overcomes the weaknesses of traditional methods such as TF-IDF and BM25 (Manning, Raghavan, & Schütze, (2008)), which rely heavily on surface-level word overlap. By leveraging contextual embeddings, the framework ensures that evidence is not only retrieved efficiently but also filtered for contextual precision and

semantic reliability. Crucially, the outcome of this stage is a high-quality, semantically coherent pool of evidence strands that can be systematically structured in the subsequent knowledge graph construction phase. By moving from broad retrieval to precise contextual alignment, the framework lays the foundation for modelling higher-order relationships between claims and evidence, enabling downstream mechanisms such as sufficiency detection and logical coherence evaluation.

```

ALGORITHM Stage1_Evidence Parsing, Retrieval, Representation

INPUT: CLAIMS, CORPUS, KNOWLEDGE_BASES, CONFIG
OUTPUT: KG

PROCEDURE STAGE1():
  parsed ← PARSE_CLAIMS(CLAIMS)
  cand ← RETRIEVE_CANDIDATES(parsed, CORPUS, CONFIG)
  ev ← RERANK_CANDIDATES(parsed, cand, CONFIG)
  links ← LINK_ENTITIES(parsed, ev, KNOWLEDGE_BASES, CONFIG)
  KG ← BUILD_KNOWLEDGE_GRAPH(parsed, ev, links, CONFIG)
  RETURN KG

PROCEDURE PARSE_CLAIMS(CLAIMS):
  out ← ∅
  FOR claim IN CLAIMS:
    structure ← EXTRACT_SVO_STRUCTURE(claim)
    mentions ← IDENTIFY_ENTITY_MENTIONS(claim)
    normalized ← NORMALIZE_MENTIONS(mentions)
    out ← out ∪ { id, text=claim, structure, mentions=normalized }
  RETURN out

PROCEDURE RETRIEVE_CANDIDATES(parsed, CORPUS, CONFIG):
  map ← ∅
  FOR p IN parsed:
    initial ← FIND_SEMANTIC_NEIGHBORS(p.text, CORPUS, CONFIG.topK_recall)
    filtered ← APPLY_RECALL_FILTERS(initial, CONFIG)
    map ← map ∪ { p.id : filtered }
  RETURN map

PROCEDURE RERANK_CANDIDATES(parsed, cand, CONFIG):
  evmap ← ∅
  FOR p IN parsed:
    pairs ← FORM_PAIRS(p.text, cand[p.id])
    scored ← SCORE_PAIRWISE_RELEVANCE(pairs)
    ranked ← SORT_DESC(scored)
    topk ← SELECT_TOPK(ranked, CONFIG.topK_precision)
    pruned ← APPLY_PRECISION_FILTERS(topk, CONFIG)
    evmap ← evmap ∪ { p.id : pruned }
  RETURN evmap

PROCEDURE LINK_ENTITIES(parsed, ev, KBs, CONFIG):
  LINKS ← ∅
  FOR p IN parsed:
    clinks ← MAP_MENTIONS_TO_KB(p.mentions, KBs, CONFIG)
    elinks ← ∅
    FOR e IN ev[p.id]:
      m ← IDENTIFY_ENTITY_MENTIONS(e.text)
      elinks ← elinks ∪ MAP_MENTIONS_TO_KB(m, KBs, CONFIG)
    LINKS ← LINKS ∪ { claim_id=p.id, claim_links=clinks, ev_links=elinks }
  RETURN LINKS

PROCEDURE BUILD_KNOWLEDGE_GRAPH(parsed, ev, links, CONFIG):
  KG ← INIT_GRAPH()
  FOR p IN parsed:
    c ← ADD_CLAIM_NODE(KG, p.id, p.text)
    ADD_ENTITY_NODES_AND_MENTION_EDGES(KG, c, links.claim_links(p.id))
    FOR e IN ev[p.id]:
      en ← ADD_EVIDENCE_NODE(KG, e.id, e.text, e.meta)
      CONNECT(KG, c, en, REL="SUPPORTED_BY", WEIGHT=e.score)
      trip ← EXTRACT_TRIPLETS(e.text, links.ev_links(e.id))
      ADD_TRIPLETS_TO_GRAPH(KG, trip, SOURCE=e.id)
    IF CONFIG.extend_neighbors:
      AUGMENT_WITH_KB_NEIGHBORS(KG, around=links for p.id)
  RETURN KG

```

Figure 3.3: Pseudocode of Evidence Retrieval and Representation

3.2.2.2 Entity Linking and Knowledge Graph Construction

Once relevant evidence passages are retrieved, they are subjected to entity linking, represented in the pseudocode as the `ENTITY_LINKING()` function. At this stage, contextual snippets from evidence are transformed into structured triplets in the form of head–relation–tail structures. For example, the evidence sentence “*Hypertension increases the risk of stroke*” is represented as $[Hypertension] \rightarrow increases \rightarrow [Risk\ of\ Stroke]$. Each element of the triplet is linked to a canonical ontology identifier such as Dbpedia and Wikidata ensuring that diverse surface forms and synonyms are unified. This step mitigates ambiguity caused by synonymy such as “heart attack” vs. “myocardial infarction” and polysemy, thereby enhancing both consistency and interpretability across the knowledge base. The final operation in this stage is knowledge graph construction, implemented through the `BUILD_KNOWLEDGE_GRAPH()` function. Here, the claim entities and evidence triplets are integrated into a coherent graph structure, where nodes correspond to entities and edges encode semantic relations. Importantly, provenance metadata is preserved so that each edge can be traced back to its original evidence source, ensuring transparency and auditability. The resulting knowledge graph is not merely a storage artifact but a reasoning resource which supports downstream tasks such as evidence sufficiency assessment, logical coherence validation, and contradiction detection.

To guarantee the reliability of the constructed graph, multiple validation steps were applied. These included verifying entity identifier consistency, ensuring relation alignment with ontology constraints, and testing graph-level coherence (i.e., whether entities and edges formed logically valid substructures). Results from these checks confirmed that the constructed triples conformed to ontology-based expectations and maintained structural integrity across domains. The outcome of this stage is a semantically enriched knowledge graph that consolidates both claims and their supporting or refuting evidence into a unified,

explainable representation. This graph provides the structural foundation for the subsequent stages of the framework, including knowledge acquisition and factuality disambiguation, where higher-level reasoning is required to address evidence sufficiency, logical coherence, and contradiction handling.

In summary, this stage demonstrates how raw claims are transformed into structured representations through a coordinated sequence of parsing, retrieval, linking, and graph construction. By anchoring entities to canonical identifiers and encoding their relationships within a transparent graph, the framework overcomes the limitations of surface-level evidence retrieval. The result is a knowledge graph that serves as a robust and explainable foundation for automated fact-checking, ensuring that subsequent reasoning operates over evidence that is both semantically aligned and structurally coherent.

3.2.2.3 Knowledge Acquisition for Claim Factuality

The objective of this stage is to advance from raw, retrieved evidence towards structured, context-enriched knowledge that ensures accurate factual coherence mapping between claims and evidence. Unlike the earlier retrieval phase, which simply identifies potentially relevant evidence, this stage addresses a more fundamental question: *How can we ensure that retrieved evidence is not only related, but also factually coherent with the claim?* The answer lies in moving beyond retrieval to a process of semantic refinement and structural validation, where evidence is assessed in terms of its coherence rather than surface-level overlap. As illustrated in Figure 3.4 the Overview of the Knowledge Acquisition for claim factuality, the process begins with a claim represented in triplet form ($E1 \rightarrow relation \rightarrow E2$) alongside a pool of retrieved evidence. At this point, the framework applies the Knowledge Acquisition phase, which explores and evaluates candidate evidence paths for their semantic and structural alignment with the claim. This represents a shift from

retrieval-centric approaches to contextual and reasoning-driven modeling, ensuring that only the most meaningful evidence is retained.

Traditional approaches such as keyword overlap or Jaccard similarity (Zhang et al., (2019)) and embedding-based models such as TransE (Bordes et al., (2013)) often overlook the neighbourhood context of entities. While these methods can retrieve direct triples, they struggle with multi-hop connections and indirect evidence, leading to incomplete or incorrect fact-checking outcomes, particularly in domains like biomedical science or political discourse where context is crucial.

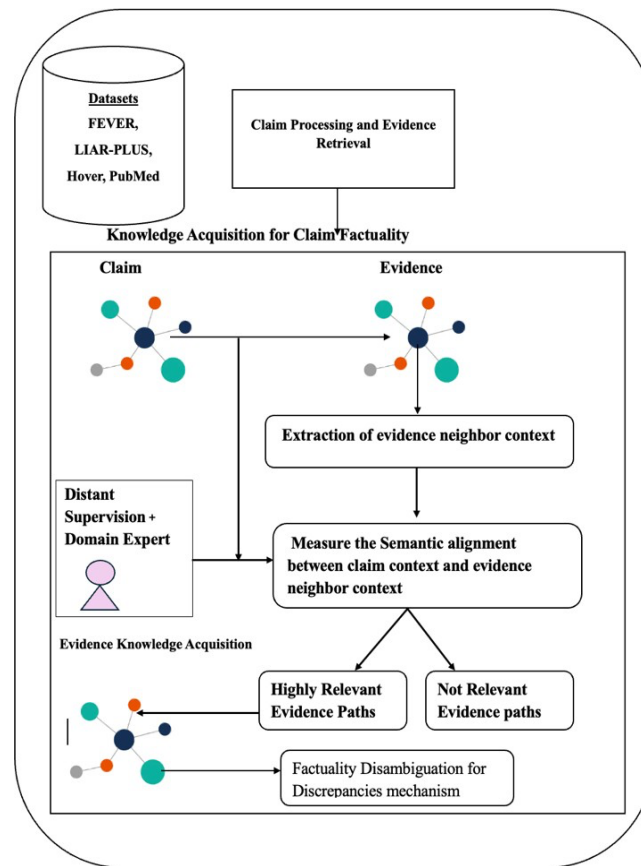


Figure 3.4: Overview of Knowledge Acquisition for Claim Factuality Disambiguation

The proposed framework introduces a neighbour-context-based semantic mapping approach to overcome these limitations. Instead of analyzing triples in isolation, the method integrates both local context (immediate entity neighbours) and path context (multi-hop relation sequences). This design enables the framework to reconstruct relational pathways

when direct evidence is missing. For example, the claim “*Barack Obama was born in Hawaii*” may not appear as a single triple in the knowledge graph. This claim may not appear as a single triple in the knowledge graph because factual knowledge is often represented at a finer level of granularity. Knowledge graphs typically encode atomic relations such as $(Barack\ Obama \rightarrow born_in \rightarrow Honolulu)$ and $(Honolulu \rightarrow located_in \rightarrow Hawaii)$, but avoid directly storing of all higher-level generalizations to reduce redundancy. As a result, validating the claim requires reasoning across multiple hops, reconstructing the pathway $(Barack\ Obama \rightarrow born_in \rightarrow Honolulu \rightarrow located_in \rightarrow Hawaii)$. This reflects a recurring limitation of existing knowledge graphs: direct evidence is often sparse or entirely absent. Neighbourhood-context reasoning addresses this gap by reconstructing indirect yet semantically valid chains of evidence. In doing so, it acts as a bridge between incomplete direct links and the richer contextual pathways that ultimately underpin factual verification.

The internal workings of the Neighbourhood Context Modeling technique in knowledge acquisition framework is illustrated in Figure 3.5, while the step-by-step procedural flow is further detailed in the accompanying pseudocode presented in Figure 3.6. Beginning with the claim triplet $(E1 \rightarrow relation \rightarrow E2)$, the system first identifies direct neighbours, then performs controlled multi-hop expansion up to 2–3 hops to generate candidate relational paths. Each path is subsequently encoded with structural and semantic features, such as relation sequence, entity participation, neighbourhood density, and claim–context similarity. These features are then evaluated using a semantic alignment model with contextual attention. Paths that exhibit strong semantic alignment are retained, while irrelevant or weak paths are discarded. If no coherent path is identified, the system explicitly flags the case as *Not Enough Information (NEI)*, thereby avoiding incorrect refutations when context is missing.

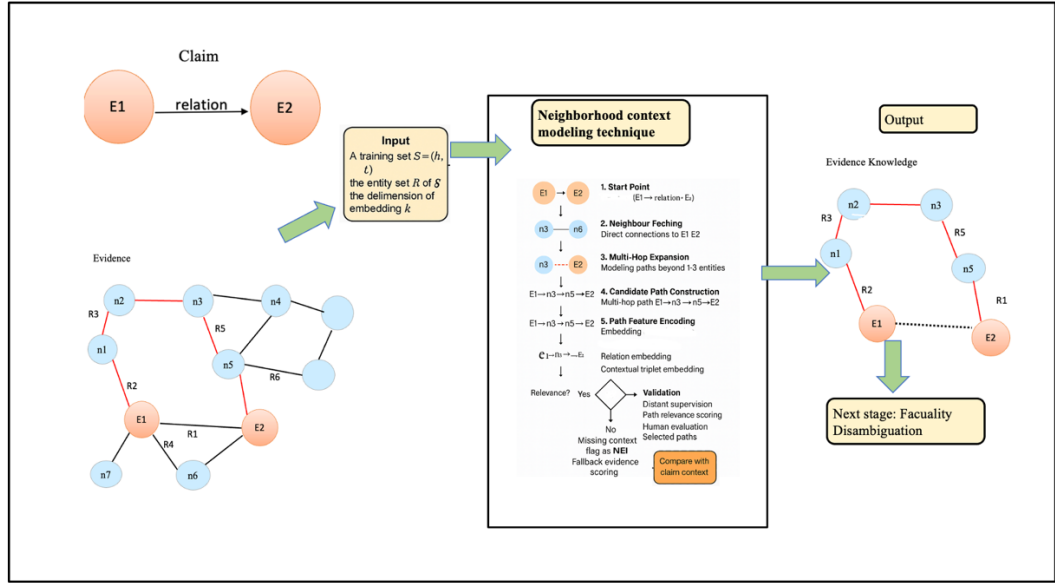


Figure 3.5: Neighbourhood Context Modeling in Knowledge Acquisition Framework

```

PROCEDURE MAP_CLAIM_TO_EVIDENCE(CLAIM, KNOWLEDGE_GRAPH):
  (E1, REL, E2) ← PARSE_TRIPLET(CLAIM)
  NEIGH_E1 ← GET_NEIGHBOURS(E1, max_hops=2)
  NEIGH_E2 ← GET_NEIGHBOURS(E2, max_hops=2)
  PATHS ← FIND_PATHS(E1, E2, KNOWLEDGE_GRAPH, max_length=3)
  FOR path IN PATHS:
    CONTEXT ← ENCODE_CONTEXT(path.entities, path.relations)
    SCORE ← SEMANTIC_ALIGNMENT(CLAIM, CONTEXT)
    STORE(path, SCORE)
  HIGHLY_RELEVANT ← FILTER(PATHS, threshold=τ)
  RETURN HIGHLY_RELEVANT
END PROCEDURE

```

Figure 3.6: Pseudocode of Neighbour Context Modeling in Knowledge Acquisition Framework

The output of this stage is a refined subset of evidence paths that are both semantically aligned and structurally coherent. This output is visualized in Figure 3.6, where only strong, validated paths remain to connect the claim entities. For example, even if the direct edge $(E1 \rightarrow E2)$ is weak or absent, a multi-hop path such as $(E1 \rightarrow n1 \rightarrow n2 \rightarrow n3 \rightarrow n5 \rightarrow E2)$ may provide the required factual linkage. By filtering evidence in this way, the

framework ensures that subsequent reasoning stages operate on high-quality, contextually meaningful knowledge rather than noisy or misleading connections.

The importance of this stage lies in its ability to distinguish relevance of knowledge as opposed to coherence. Retrieval alone does not guarantee factual support; it merely identifies potential connections. By embedding triplets within their neighbourhood context (a.shankar et al., (2025)), this stage separates incidental associations from genuinely coherent evidence. This prevents error propagation to later stages and strengthens the robustness of the entire fact-checking pipeline.

By filtering and structuring the evidence, this stage prepares the groundwork for the next stage, where the framework focuses on Factuality Disambiguation, since even coherent evidence may still be incomplete or contradictory and thus requires further resolution before a final factuality decision can be made. Once coherent evidence paths have been established, the system must address cases where the evidence is contradictory, incomplete, or inconsistent across domains. Without the structured filtering provided by knowledge acquisition, disambiguation would be forced to reason over unreliable data. By serving as both a semantic filter and a structural bridge, knowledge acquisition ensures that factuality disambiguation begins with the strongest possible foundation, thereby enhancing the accuracy and trustworthiness of the overall framework.

3.2.3 Factual Disambiguation of Discrepancies Detection Mechanism

While the knowledge acquisition stage ensures that claims are aligned with semantically coherent and structurally relevant evidence, this alone does not guarantee that the evidence is sufficient and logically consistent. In real-world scenarios, retrieved evidence may still be incomplete, contradictory, or contextually weak. If left unresolved, such discrepancies undermine the credibility of the fact-checking process and lead to unreliable

conclusions. The objective of this stage, therefore, is to disambiguate discrepancies in evidence by assessing both its sufficiency and its logical coherence, ensuring that the final factuality decision is made on the basis of robust and trustworthy knowledge. Traditional approaches often treat sufficiency and logical coherence in isolation. For example, methods based on cosine similarity or keyword overlap ((Mihalcea et al., (2006); Reimers & Gurevych, (2019); Thorne et al., (2018)) can identify surface-level lexical matches but fail when evidence is indirect or incomplete. Conversely, symbolic logic-based approaches ((Sikka et al., (2013); Chen et al., (2020)) can validate consistency once explicit rules are defined, but they lack the semantic flexibility to handle multi-hop or ambiguous evidence. These limitations result in frequent misclassifications when evidence is distributed across multiple domains or when contradictions exist within the same evidence pool.

To address this, the proposed mechanism integrates both semantic and logical reasoning within a single stage. As illustrated in Figure 3.7, the framework first applies contextual attention mechanisms to evaluate evidence sufficiency, determining whether the available information is strong enough to support or refute the claim. Unlike traditional similarity measures, contextual attention assigns adaptive weights to entities and relations in the knowledge graph, thereby highlighting the most informative segments of evidence while filtering out noise.

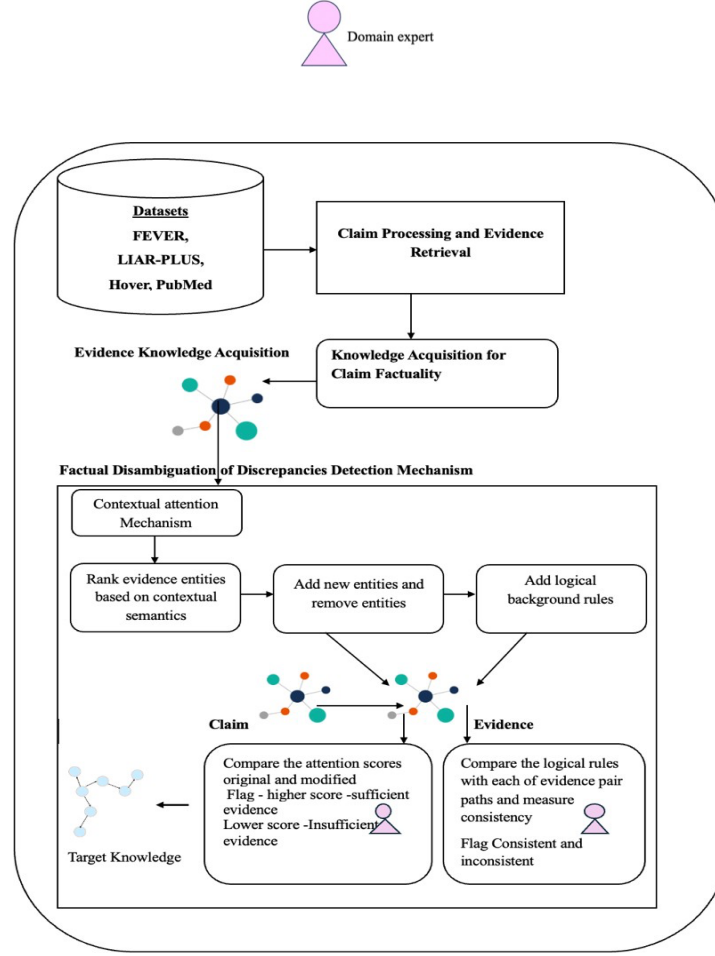


Figure 3.7: Factual Disambiguation of Discrepancies Detection Mechanism

Building on sufficiency assessment, the second layer introduces logical coherence validation. Here, evidence paths are checked against structured rules derived from domain expertise. Rules such as *causality* (if X reduces Y , then Y should not remain unchanged) or *temporal ordering* (an effect must not precede its cause) are applied to ensure consistency across multiple strands of evidence. As illustrated in Figure 3.8, this process enables the framework to reconcile indirect yet coherent evidence while filtering out internally contradictory strands. By doing so, the system avoids the weaknesses of purely semantic models that might accept contradictory inputs, and purely logical approaches that might reject incomplete yet useful evidence.

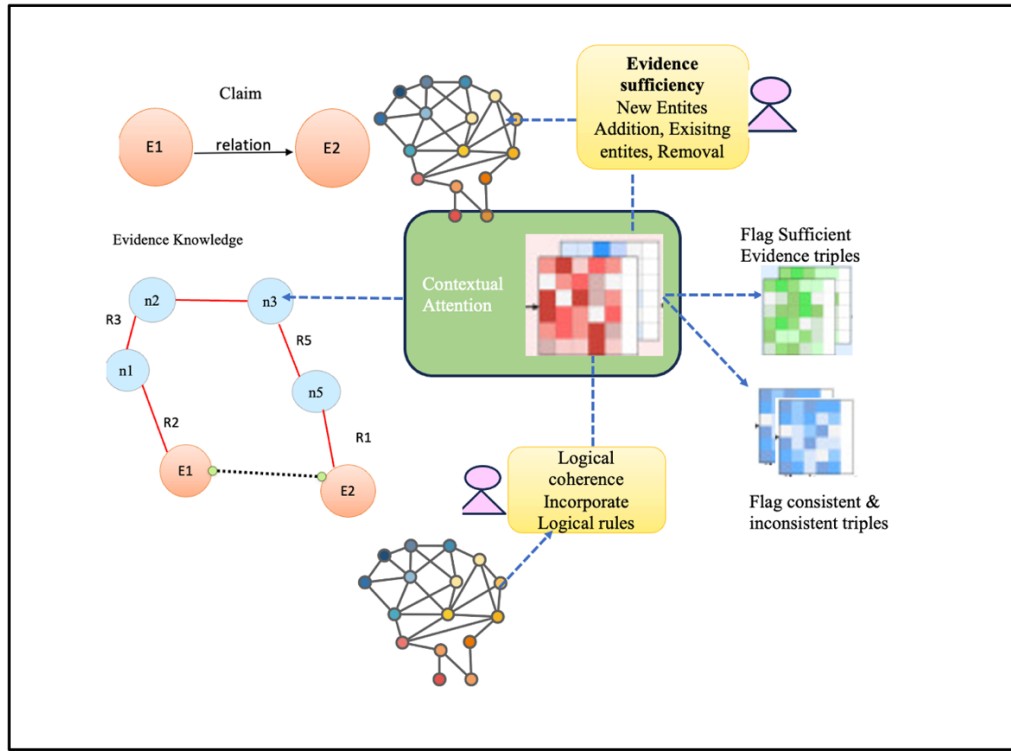


Figure 3.8: Evidence sufficiency and logical Coherence Assessment

The workflow is formalized in Pseudocode as shown in Figure 3.9, which outlines the step-by-step process. Beginning with claim–evidence triplets, the system first evaluates contextual sufficiency using attention scores. Next, paths are refined by adding or removing entities based on semantic alignment, before being validated against logical rules. This combined evaluation ensures that only evidence paths which are both semantically relevant and logically consistent are retained.

```

PROCEDURE FACTUALITY_DISAMBIGUATION(CLAIM, EVIDENCE_GRAPH):
  PATHS ← RETRIEVED_CONTEXTUAL_PATHS(CLAIM)
  FOR path IN PATHS:
    SCORE ← CONTEXTUAL_ATTENTION(CLAIM, path)
    IF SCORE ≥  $\tau_1$ :
      FLAG(path, "Sufficient")
    ELSE:
      FLAG(path, "Insufficient")
      path ← ENTITY_REFINEMENT(path) # add/remove nodes
      SCORE ← RECOMPUTE_ATTENTION(CLAIM, path)

  LOGICAL_VALIDITY ← APPLY_RULES(path)
  IF LOGICAL_VALIDITY = False:
    FLAG(path, "Inconsistent")
  ELSE:
    FLAG(path, "Consistent")

  FINAL_SCORE ← COMBINE(ATTENTION_SCORES, LOGICAL_FLAGS)
  RETURN LABEL(FINAL_SCORE) # Supported, Refuted, NEI
END PROCEDURE

```

Figure 3.9: Pseudocode of Discrepancies Detection Mechanism

The outcome of this stage is a refined set of evidence paths annotated with both semantic sufficiency and logical consistency. As shown in Figure 3.6, the framework classifies evidence as sufficient, contradictory, or insufficient for claim verification. In cases where no coherent evidence remains after evaluation, the system outputs a “*Not Enough Information (NEI)*” label, thereby avoiding misleading conclusions and ensuring that claims are not prematurely judged on incomplete or inconsistent grounds. What makes this stage particularly significant is its ability to move beyond surface-level similarity measures. Instead of treating evidence fragments in isolation, the framework integrates contextual sufficiency with logical validation. This dual-layered assessment ensures that verification outcomes are not only grounded in relevant evidence but are also consistent with domain-specific reasoning patterns. Such integration is especially critical in cases where evidence may initially appear supportive but on closer inspection reveals contradictions.

Equally important, however, is that the outcome remains interpretable to users. Beyond producing a label of Supported, Refuted, or Not Enough Information (NEI), the framework makes the reasoning paths behind each decision explicitly visible. For example, the claim “*Vitamin D supplementation reduces the risk of influenza*” may be validated through the reasoning path *Vitamin D* $\rightarrow$ *reduces risk* $\rightarrow$ *influenza incidence*. In contrast, contradictory strands such as *Vitamin D* $\rightarrow$ *no significant effect* $\rightarrow$ *influenza risk* are highlighted and excluded from the final judgement. When no consistent evidence remains, the system assigns an explicit NEI label, ensuring that the absence of reliable support is transparently communicated. This structured outcome serves two complementary functions. First, it distinguishes between sufficient, insufficient, and contradictory evidence, preventing diverse findings from being collapsed into a single verdict. Second, it enhances transparency by showing not only which evidence paths were retained or discarded but also the rationale behind each decision. Together, these functions ensure that factuality assessments are not only accurate but also explainable and trustworthy, thereby strengthening user confidence in the verification process.

In summary, the Factual Disambiguation of Discrepancies Detection stage grounds claim verification in both contextual sufficiency and logical coherence. By explicitly modelling these two dimensions, the framework overcomes the limitations of earlier approaches that relied solely on similarity metrics. The result is a system that produces outputs that are robust, interpretable, and adaptable across domains. Having now established the mechanisms for knowledge acquisition and factual disambiguation, the next chapter moves to their empirical evaluation on benchmark datasets, testing how well these components perform in practice under real-world fact-checking scenarios.

3.3 Hybrid Framework to Evaluate Factual coherence and Factual Disambiguation

The evaluation of the proposed framework constitutes a critical stage of this research, providing a structured and systematic assessment of how effectively it addresses the challenges of claim factuality disambiguation. Importantly, the evaluation is designed not only to measure the performance of individual components in isolation but also to examine their integration as a coherent pipeline. This ensures that the framework can be validated under conditions that reflect the complexity of real-world fact-checking rather than controlled or artificially simplified settings.

The primary objective of this evaluation is to test whether the framework can consistently retrieve relevant evidence, construct structured knowledge, and evaluate both sufficiency and coherence. To capture a wide spectrum of claim–evidence relationships, four diverse datasets are employed: FEVER, LIAR-PLUS, HoVer, and PubMed. These datasets vary significantly in both domain and complexity ranging from short, politically oriented claims to multi-hop biomedical reasoning. Their inclusion ensures that the framework is tested against varied misinformation scenarios, thereby validating its adaptability beyond narrow domains.

The evaluation proceeds in four main stages, each targeting a specific dimension of the framework. First, evidence retrieval is measured using precision, recall, and F1 metrics, ensuring that relevant passages are retrieved with broad coverage while minimizing irrelevant noise. Second, knowledge acquisition is evaluated by examining both the structural accuracy of the constructed knowledge graphs and their semantic alignment with claims, confirming that retrieved evidence is meaningfully organised for downstream reasoning. Third, evidence sufficiency is tested to determine whether the framework can correctly detect inadequate evidence, with ambiguous cases explicitly labelled as Not Enough Information (NEI). This is critical for preventing overconfident predictions when no

reliable supporting data exists. Finally, logical coherence is validated to ensure that evidence not only aligns semantically with claims but also maintains internal consistency, particularly when information is derived from multiple sources.

Although the evaluation relies primarily on quantitative metrics, expert validation is selectively incorporated to cross-check outputs in ambiguous cases. This is presented as a complementary alignment, ensuring that computational outputs remain interpretable and aligned with domain reasoning. Figure 3.10 summarises this evaluation framework, showing how retrieval, knowledge acquisition, sufficiency detection, and logical coherence are benchmarked across datasets, with human-in-the-loop checks providing selective oversight.

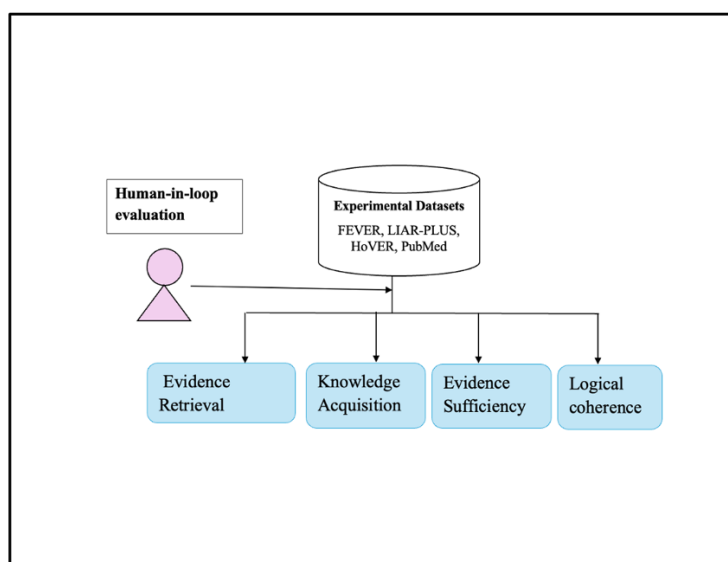


Figure 3.10: A Hybrid Human in Loop Evaluation Framework

The results of this evaluation confirm that the framework performs reliably across all four dimensions: retrieval accuracy, graph construction fidelity, sufficiency detection, and logical coherence. By combining heterogeneous datasets with a hybrid evaluation strategy that integrates automated benchmarking and selective expert review, the framework demonstrates both reliability and adaptability in handling insufficient or conflicting evidence

or in alignment of cross-domain evidences. Crucially, this evaluation design moves beyond traditional baselines, which often assess retrieval or classification performance in isolation, without considering how sufficiency and coherence interact in complex verification tasks. By explicitly benchmarking these dimensions together, the framework provides a more holistic and explainable measure of claim verification.

Having established this rigorous evaluation methodology, the next chapter presents the empirical outcomes of these assessments, systematically comparing the proposed framework against baseline models and state-of-the-art techniques to highlight its strengths, limitations, and contributions to advancing claim factuality disambiguation.

3.4 Chapter Summary

This chapter has presented the methodology supporting the proposed framework for claim factuality disambiguation. Rather than revisiting each component in isolation, the emphasis was placed on showing how the components collectively operate as an integrated pipeline that transforms raw claim–evidence inputs into structured and interpretable factuality assessments.

The methodological design was guided by three central themes (1) to establish reliable mechanisms for retrieving and representing evidence in structured form, (2) to move beyond surface-level similarity by incorporating semantic knowledge acquisition and context-aware mapping, and (3) to embed evaluative mechanisms capable of addressing the challenges of evidence insufficiency, inconsistency, and ambiguity. Together, the combined outcomes provide the structural foundation and analytical depth required for a transparent and trustworthy fact-checking process.

A further strength of the methodology lies in its adaptability. The framework was deliberately designed to operate across heterogeneous datasets and domains while remaining resilient to the evolving and uncertain nature of misinformation. This adaptability makes the framework both methodologically novel and practically relevant, addressing persistent limitations of prior approaches that struggled to ensure evidence sufficiency and logical coherence.

Finally, this chapter has established the rationale, structure, and methodological foundations of the knowledge acquisition framework for claim factuality disambiguation. The next chapter builds on this foundation by presenting the empirical evaluation, systematically assessing the performance of the framework against baseline and state-of-the-art methods to validate its effectiveness and contribution to more robust, adaptive, and explainable fact-checking systems.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the results and discussion of the experimental evaluation of the proposed *Knowledge Acquisition Framework for Claim Factuality Disambiguation*. The aim of this evaluation is to demonstrate the effectiveness of the framework in addressing the key challenges of automated fact-checking, namely evidence insufficiency and logical incoherence. By analyzing outcomes at both the component and system levels, the chapter illustrates how the framework advances beyond baseline methods to deliver more reliable veracity assessments.

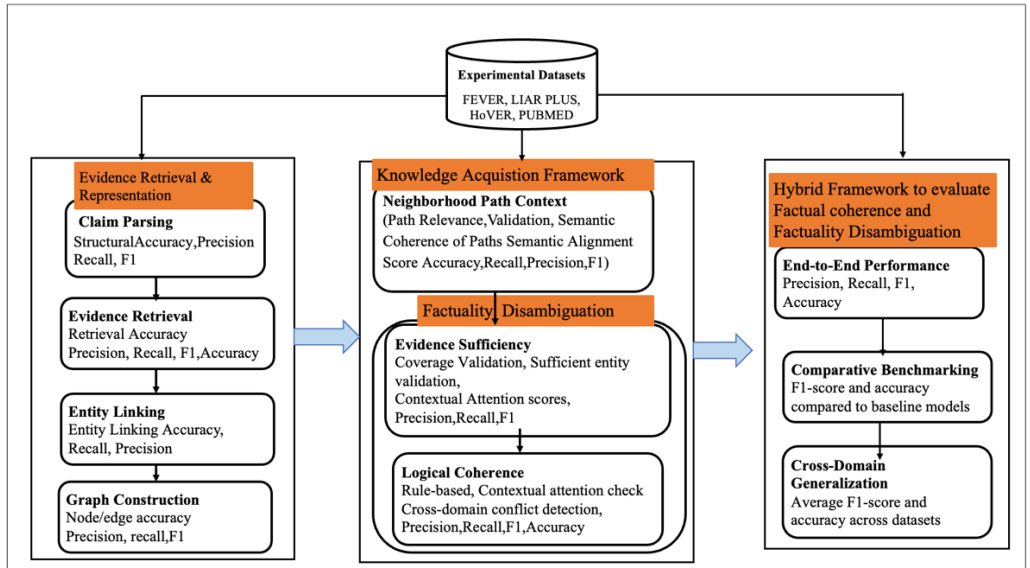


Figure 4.1: Overview of the Framework Validation

The evaluation process, outlined in Figure 4.1, is organized into three interconnected stages, each associated with specific validation checkpoints and performance metrics. This layered structure ensures that evaluation progresses systematically from fine-grained validation at the component level to growing robustness at the system level. Instead

of treating the stages as independent modules, the framework organizes them as a progressive sequence such as structural reliability is first established in the Evidence Retrieval and Representation stage where semantic adequacy and logical coherence are then evaluated in the Knowledge Acquisition and Factuality Disambiguation stage and finally, overall robustness is benchmarked in the Framework Evaluation stage. The first stage, Evidence Retrieval and Representation, evaluates the framework’s ability to parse claims, retrieve relevant evidence, and structure this evidence into graph-based representations. This stage establishes structural reliability, ensuring that claims and evidence are correctly identified and represented in a triplet format. Validation is carried out through retrieval accuracy, structural integrity checks, and node/edge-level consistency, laying the foundation for higher-level reasoning.

The second stage, Knowledge Acquisition and Factuality Disambiguation, evaluates whether the retrieved and structured evidence is both sufficient and logically coherent in relation to the claim. To achieve this, the framework applies neighbourhood path exploration to uncover relevant multi-hop connections within the knowledge graph, complemented by contextual attention mechanisms that assess the semantic alignment of these paths with the claim. Validation at this stage is multi-layered. First, node and edge accuracy ensures that entities and their relations are correctly represented within the graph, preventing structural inconsistencies. Second, evidence sufficiency is validated at the entity level, confirming that the retrieved paths contain all necessary components to meaningfully support or refute the claim rather than relying on incomplete fragments. Third, semantic coherence is checked to ensure that evidence paths remain aligned with the contextual meaning of the claim. Fourth, logical coherence is validated through rule-based checks, where domain-specific reasoning rules are applied to verify that the relationships among entities follow causal connection. Finally, coherence consistency is assessed across multiple

parts of evidence, ensuring that supportive, contradictory can coexist transparently without collapsing into misleading or oversimplified outcomes. By integrating these validation mechanisms, the framework moves beyond surface-level retrieval and establishes a deeper, context-aware representation of claim–evidence relationships. This layered validation ensures that only structurally valid, semantically coherent, and contextually sufficient evidence is advanced to the next stage of reasoning, providing a robust foundation for factuality assessment

The third and final stage is Framework Evaluation which provides a complete assessment of the proposed system as an end-to-end fact-checking framework. At this stage, the framework is benchmarked against state-of-the-art language models such as BERT, RoBERTa, and BioBERT as well as established baseline methods such as TransE and Jaccard Similarity. Evaluation spans multiple benchmark datasets like FEVER, LIAR-PLUS, HoVer, and PubMed which collectively capture the diversity of claim verification tasks across political, cross-domain, and biomedical contexts. This design ensures that performance is not confined to narrow conditions but is tested against heterogeneous forms of misinformation. The evaluation itself is structured around three dimensions such as end-to-end performance which validates the pipeline as a whole, comparative benchmarking which highlights gains over baseline models, methods and cross-domain generalisation, which confirms the adaptability of the framework to varied evidence types and claim structures. Overall, the evaluation process ensures that the proposed framework is examined at multiple levels, beginning with structural validation, extending to semantic and logical adequacy, and concluding in general benchmarking. To make these validation dimensions more explicit, Table 4.1 summarises how precision, recall, accuracy, and F1-score are operationalised across different stages of the framework. This layered design allows the framework to be assessed both as a collection of interdependent components and as an

integrated system. The following sections present the results of each stage in detail, together with critical discussion of their implications for advancing automated fact-checking research.

Table 4.1: Comparisons Matrix of Evaluation Metrics

Technique	Precision	Recall	Accuracy	F1-Score
Claim–Evidence Factual Coherence (Triplet Embedding)	Measures how many pieces of evidence are semantically aligned with the claim.	Evaluates how much of the relevant evidence was successfully retrieved.	Improved through human-in-the-loop validation that aligns claim and evidence, reduces misclassifications, and enhances overall correctness.	Balances precision and recall by verifying semantic consistency and ensuring comprehensive evidence retrieval.
Evidence Sufficiency (Contextual Attention)	Measures how many pieces of evidence are sufficient to support or refute the claim.	Considers how much of the required evidence is captured to verify the claim.	Strengthened through human-in-the-loop validation, where experts prioritize sufficient evidence, avoid irrelevant evidence, and identify additional relevant evidence.	Balances sufficiency and relevance, providing well-rounded fact-checking decisions based on the adequacy of evidence.
Logical Coherence (Contextual Attention)	Measures how many pieces of evidence are logically consistent with the claim.	Assesses how effectively the system identifies all logically relevant and consistent evidence.	Reinforced through human intervention, which addresses complex logical connections and prevents false classifications.	Ensures balance by including only logically consistent evidence, minimizing contradictions, and preventing contradictory evidence from being overlooked.

4.1.1 Case Study: Evaluating the Claim “Vitamin C Supports Common Cold”

This section illustrates how the proposed framework operates in practice through a case study on the claim “*Vitamin C supports the common cold*”. This claim has long been debated within the health domain, with some clinical studies reporting modest benefits in

reducing the duration of symptoms and others finding little or no evidence for preventive effects. Its combination of conflicting and incomplete findings makes it a particularly suitable example for testing the framework’s ability to manage insufficiency, and contradiction in evidence. The process begins with claim parsing, where the statement is decomposed into its core components. In this case, *Vitamin C* is identified as the subject, *supports* as the predicate, and *common cold* as the object. By isolating these elements, the framework identifies the key entities and relationships that form the basis for targeted evidence retrieval.

From the structured representation, the framework proceeds to retrieve candidate evidence from multiple sources, including PubMed-indexed clinical trials. The retrieved body of evidence illustrates the diversity of findings. Large-scale studies concluded that daily Vitamin C intake does not prevent colds in the general population. However, they did report a modest effect on symptom duration, with colds shortened by approximately 10 percent, alongside slightly milder symptoms. By contrast, studies where Vitamin C was taken only after the onset of illness found no measurable benefit for either duration or symptom severity. A third line of evidence examined individuals in unique conditions, such as marathon runners or soldiers engaged in strenuous activity in extreme cold. In these cases, Vitamin C was associated with a marked preventive effect, reducing the risk of developing colds by about half. This mixture of modest support, null results, and conditional preventive effects highlights the challenges of assessing the claim in absolute terms.

Once the relevant evidence is gathered, the framework proceeds with entity linking and graph construction to transform the raw text into a structured representation. Entities such as *Vitamin C*, *common cold*, *symptom duration*, *prevention*, *athletes in extreme cold*, and *adverse effects* are mapped to canonical identifiers within the knowledge graph. Their relationships are then encoded as edges, for example: *Vitamin C* $\rightarrow$ *reduces duration* $\rightarrow$

common cold; *Vitamin C* → *no preventive effect* → *general population*; *Vitamin C* → *preventive effect* → *athletes in strenuous activity*; and *Vitamin C (high dose)* → *mild side effect* → *diarrhea*.

To enrich the graph beyond direct pairings, the framework applies neighbourhood context modelling, which systematically explores both local neighbours and multi-hop connections around each entity. Through this process, the framework moves beyond isolated evidence triplets and instead captures relational chains that link entities indirectly, allowing the graph to represent not just individual facts but also the broader pathways through which claims and evidence are connected. For example, evidence paths such as *Vitamin C* → *reduces duration* → *shorter illness* → *milder symptoms* or *Vitamin C* → *preventive effect* → *subgroup (athletes)* → *reduced risk of colds* are constructed through controlled multi-hop traversal. Each path is evaluated for path relevance, semantic coherence, and node/edge accuracy, ensuring that only structurally consistent and semantically meaningful connections are preserved (a.shankar et al., (2024)). This neighbourhood-based exploration enables the framework to reconstruct dependencies across fragmented pieces of evidence, making explicit the conditional and context-dependent nature of the findings. In practice, this means that even when direct links are weak or absent, the graph still retains meaningful multi-step pathways that reflect how different strands of evidence interact. By embedding these extended connections, the knowledge acquisition stage provides a richer and more explanatory graph that forms the foundation for the subsequent sufficiency and coherence evaluations.

Once entity linking and neighbourhood context expansion have been applied, the result is a knowledge graph enriched with both direct and multi-hop evidence paths. This graph captures not only the immediate relationships between *Vitamin C* and the *common cold* but also the broader, conditional dependencies that emerge across different strands of

evidence. By embedding both direct and indirect connections, the framework ensures that the evidence base is comprehensive enough to reflect the true diversity of clinical findings. However, constructing the graph is not the endpoint. For the framework to provide a trustworthy factuality assessment, it must now evaluate whether the evidence represented in this graph is sufficient and whether it forms a logically coherent.

Having established these structured connections, the framework then evaluates the adequacy and coherence of the evidence. Sufficiency analysis determines that while the evidence is adequate to support limited interpretations of the claim notably, reductions in duration and mild symptom relief, it is insufficient to support prevention in the general population. Logical Coherence analysis further examines whether the evidence forms a consistent narrative or reveals contradictions. Here, the framework identifies contradictions across the evidences, for example, prevention is supported in specific subgroups but not in general populations, while symptom relief is confirmed in some contexts but absent in others. By distinguishing between partial sufficiency and broader inconsistency, the framework highlights the fragmented and conditional nature of the evidence base.

Finally, the outcomes of sufficiency and coherence evaluation are synthesised into a factuality decision. Given that Vitamin C demonstrates modest, context-dependent benefits but lacks comprehensive and consistent support for the broader claim, the framework classifies “Vitamin C supports the common cold” as Not Enough Information (NEI). This outcome reflects the reality that while Vitamin C may influence cold duration or symptom severity under certain conditions, the overall evidence remains inconclusive with respect to its preventive effect.

This case study demonstrates the capacity of the framework to process claims that rely on heterogeneous and context-relevant evidence. By progressing systematically from

claim parsing and retrieval, through structured graph representation and neighbourhood context modelling, and finally to sufficiency and coherence evaluation, the framework shows how subtle and context-sensitive claims can be assessed without falling into simplistic true/false categories. In doing so, it validates the framework’s ability to produce transparent, interpretable, and reliable factuality decisions in domains where evidence is evolving and uncertain.

4.2 Results of Processing of Claim and Evidence Retrieval

This section discuss the outcomes of the proposed Knowledge Acquisition Framework during the stages of claim parsing, evidence retrieval, and clustering of retrieved evidence. The purpose of this evaluation is to determine how effectively the framework identifies the structural components of claims, retrieves semantically aligned evidence, and organises this evidence into coherent groups that form the foundation for factuality assessment. To ensure robustness, four benchmark datasets were employed such as FEVER (Thorne et al., (2018)), LIAR-PLUS (Alhindi et al., (2018)), HoVer (Jiang et al., (2020)), and PubMed (Qin et al., (2021)) along with performance assessed using accuracy, precision, recall, and F1-score (Manning, Raghavan & Schütze, (2008)).

The first stage of the process involved claim parsing, which ensures that the essential elements of a claim are accurately captured for downstream analysis. For example, in the case study claim “*Vitamin C supports the common cold*”, the framework extracted *Vitamin C* as the subject, *supports* as the predicate, and *common cold* as the object. These structural extractions were then validated against annotated ground truth. Precision was defined as the proportion of correctly extracted entities, recall as the proportion of all relevant entities successfully identified, and the F1-score as their harmonic mean (Powers, (2011)). Establishing high accuracy at this stage was crucial, as errors in parsing would directly propagate into subsequent evidence retrieval.

Building on this foundation, the next stage focused on evidence retrieval, where the framework assessed claim–evidence alignment using two similarity measures such as Jaccard Similarity and Cosine Similarity. Jaccard Similarity, a set-based measure, relies on exact word overlap (Liben-Nowell & Kleinberg, 2007). In contrast, Cosine Similarity, applied through contextual embeddings using BERT (Devlin et al., (2019)), RoBERTa (Liu et al., (2019)), and BioBERT (Lee et al., (2020)), evaluates semantic proximity, even in the absence of lexical matches. The distinction between these two approaches was particularly evident in the PubMed case study. For example, consider the evidence *“Daily Vitamin C intake did not prevent colds in the general population but reduced illness duration by about 10%”* and *“Athletes exposed to cold environments showed a 50% reduced risk of colds when taking Vitamin C”*. At a lexical level, these sentences share very few overlapping words where Jaccard Similarity therefore assigns them a low score, effectively treating them as unrelated. By contrast, Cosine Similarity evaluates the vector representations of the sentences in embedding space. Both sentences contain contextual associations between *Vitamin C*, *common cold*, and outcomes related to *prevention or reduction of illness*. In the embedding space generated by models such as BERT, the semantic neighbourhood of these terms overlaps significantly, even if the exact surface words differ. As a result, Cosine Similarity produces a high similarity score because both sentences encode the broader semantic relationship of Vitamin C being evaluated as a treatment intervention for colds, even with different outcome measures such as *“reduced duration versus reduced incidence in subgroups”*. In other words, the similarity lies not in the literal overlap of words, but in the shared conceptual structure where both sentences frame Vitamin C in relation to its impact on cold-related outcomes. Cosine Similarity is therefore sensitive to these contextual parallels, whereas Jaccard fails because it is restricted to surface-level token matching.

Building on the improved retrieval quality achieved by embedding-based methods, the next step focused on clustering the retrieved evidence. Since Cosine Similarity ensures that the retrieved material is semantically richer and less dependent on lexical overlap, the clustering stage benefits from a more meaningful input space. Clustering serves to structure the retrieved evidence into semantically coherent groups. By grouping related findings together, it reduces redundancy and highlights the diversity of perspectives such as supportive, contradictory outcomes that would otherwise remain obscured in unstructured text. To achieve this, K-means clustering (Lloyd, (1982)) was applied to the embedding representations of the retrieved sentences, with the optimal number of clusters determined using the silhouette score (Rousseeuw, (1987)). This process ensured that the clusters were both internally cohesive and externally distinct, thereby enhancing the interpretability of the evidence base, as shown in Table 4.2.

Table 4.2: Example of Clustering Output for Vitamin C Claim

Cluster	Evidence	Explanation
Symptom Reduction	“Vitamin C reduced cold duration by 10.34%.” / “Illness shortened from 10.02 to 9.08 days.”	Supports partial benefit (duration, severity).
Lack of Preventive Effect	“Daily Vitamin C intake did not prevent colds in men and women (0.89 effect size).”	Indicates insufficiency in prevention evidence.
Conditional Prevention	“Athletes in extreme cold showed a 49.56% reduced risk when taking Vitamin C.”	Demonstrates subgroup-specific effect.

The above table demonstrates how retrieved evidence was separated into symptom reduction, lack of prevention, and subgroup-specific preventive effects. This structuring highlights both consistencies and contradictions in the evidence base, providing a clearer foundation for factuality assessment.

Having demonstrated the interpretive value of clustering on a case study, the evaluation now turns to a broader quantitative comparison of retrieval performance across datasets. Table 4.3 presents the evidence retrieval results of FEVER, LIAR-PLUS, HoVer, and PubMed datasets for consistency and ease of comparison between baseline methods and models.

Table 4.3: Evidence Retrieval Performance Metrics

Dataset	Model	Similarity Measure	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
FEVER	Jaccard Similarity	Set-based	75.23	72.46	80.12	76.38
	Cosine Similarity	Embedding-based	91.47	84.62	82.18	87.06
LIAR-PLUS	Jaccard Similarity	Set-based	70.36	68.41	72.58	70.45
	Cosine Similarity	Embedding-based	85.21	83.14	84.59	84.36
HoVer	Jaccard Similarity	Set-based	71.48	69.22	75.37	72.19
	Cosine Similarity	Embedding-based	85.62	84.11	83.74	86.08
PubMed	Jaccard Similarity	Set-based	77.54	74.83	78.64	75.28
	Cosine Similarity	Embedding-based	92.18	89.47	88.53	87.94

The results indicate a consistent trend such as embedding-based Cosine Similarity outperforms set-based Jaccard Similarity across all datasets. This is evident in the accuracy improvements observed in FEVER, where performance increased from 75.23% with Jaccard to 91.47% with Cosine, and in PubMed, where the accuracy rose from 77.54% to 92.18%. Such consistent gains across multiple domains suggest that the advantage is not dataset-specific but reflects a general property of embedding-based models.

The reason these improvements occur is that Cosine Similarity evaluates sentences in a high-dimensional semantic space, allowing it to capture contextual relationships and

conceptual equivalence such as “reduces duration” vs. “shortens illness”, whereas Jaccard relies solely on lexical overlap and therefore fails when wording differs. In this way, the numerical improvements directly illustrate the semantic sensitivity of embedding-based approaches (Reimers & Gurevych, (2019)).

The advantages of embedding-based retrieval naturally extend into the clustering stage. Because semantically richer evidence is retrieved, the resulting clusters are more coherent and thematically distinct. In the PubMed case study, for example, clustering clearly separated evidence into three aspects such as symptom reduction, lack of prevention, and subgroup-specific outcomes demonstrating how higher-quality retrieval supports more structured organisation of diverse findings. Together, retrieval and clustering provide both accuracy and interpretability: retrieval ensures that relevant evidence is captured with semantic precision, while clustering makes the diversity of perspectives explicit by distinguishing supportive, contradictory, and conditional strands. This dual capability establishes a transparent foundation for evaluating evidence sufficiency and logical coherence, setting the stage for the next phase of analysis.

4.3 Results of Knowledge Graph Construction

The construction of knowledge graphs represents a pivotal stage in the proposed framework, transforming raw textual evidence into structured, semantically rich representations that explicitly capture claim–evidence relationships. Once entities are linked to their canonical identifiers within biomedical knowledge bases such as CHEBI, the framework integrates them into a graph structure where nodes represent entities and edges encode the semantic relations drawn from the evidence. This graph is designed to reflect the diversity of findings reported across the literature, including supportive outcomes such as *“Vitamin C shortens cold duration by about 10%”*, contradictory evidence such as *“Vitamin*

C shows no preventive effect in the general population”, and conditional results such as *“Vitamin C reduces the risk of colds in athletes exposed to extreme cold environments.”* Contextual associations, for example *“High doses of Vitamin C may cause diarrhea,”* are also incorporated to ensure that potential risks are represented alongside therapeutic claims. By integrating these strands into a single representation, the graph provides a transparent foundation for reasoning across supportive, contradictory, and conditional findings rather than treating them as disconnected fragments.

In this study, entities were mapped to the CHEBI (Hastings et al., (2016)) ontology to ensure semantic consistency and richness in representation (Martín et al., (2022)). To improve accuracy, BERT was applied for entity linking, which allowed the framework to correctly disambiguate context-sensitive terms. For instance, *“prevention”* was correctly interpreted in relation to the *common cold*, rather than in broader biomedical preventive contexts. This enabled the framework to capture nuanced relationships such as *“Vitamin C → shortens duration → cold symptoms,”* *“Vitamin C → no_effect → prevention (general population),”* and *“Vitamin C → conditional_prevention → athletes → reduced_risk → common cold.”*

To maintain semantic accuracy, entity linking was performed using BERT-based disambiguation, allowing the framework to interpret context-sensitive terms with precision. For instance, the term *“prevention”* was mapped specifically to its relation with *common cold* rather than to broader preventive contexts, enabling nuanced representations such as *Vitamin C → no_effect → prevention (general population)* and *Vitamin C → conditional_prevention → athletes → reduced_risk → common cold*. A graphical illustration of this constructed network is presented in Figure 4.2, which highlights the interconnected nodes such as *Vitamin C*, *common cold*, *symptom reduction*, *cold duration*, *prevention*, *athletes*, and *side effects* and their corresponding relational edges. This

representation makes explicit the contested and conditional nature of the evidence base, offering a structured view that supports subsequent sufficiency and coherence analysis.

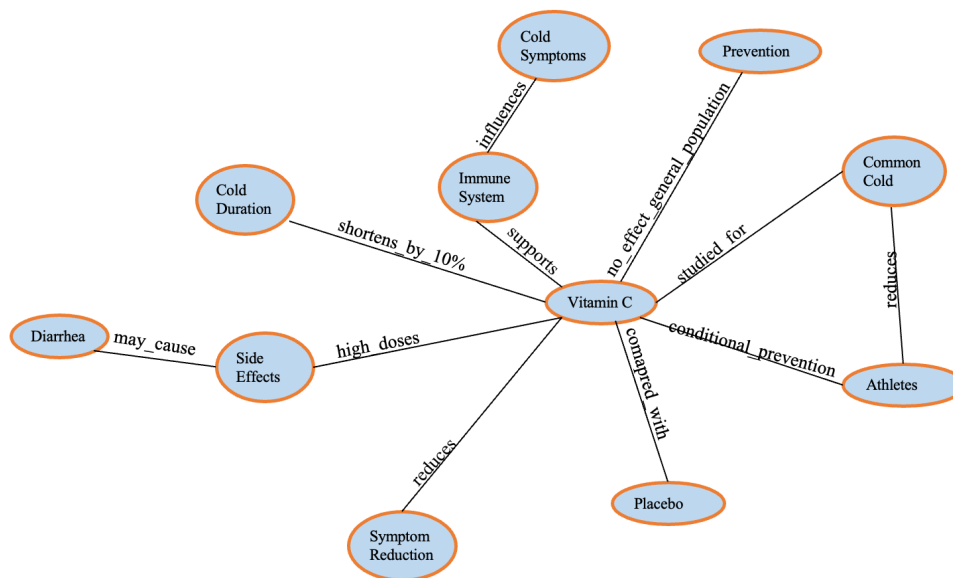


Figure 4. 2: Example of a Constructed Evidence Knowledge Graph

Complementing the graphical overview, Table 4.4 provides concrete examples of subject–relation–object triples extracted from the PubMed dataset. These triples demonstrate how textual statements are systematically mapped into structured graph relations. For example, the sentence “*Vitamin C reduced illness duration by about 10%*” is expressed as (*Vitamin C – shortens_by_10% – Cold Duration*), while subgroup findings such as “*athletes ... reduced their risk by about half*” are encoded as (*Vitamin C – conditional_prevention – Athletes → reduced_risk – Common Cold*). By systematically converting narrative text into structured triples, the framework ensures that supportive, contradictory, and conditional claims are explicitly encoded in a way that preserves both semantic and contextual fidelity.

Table 4.4: Example of Extracted Triples from PubMed Evidence

Subject	Relation	Object	Source Evidence (PubMed)
Vitamin C	shortens_by_10%	Cold Duration	“Daily Vitamin C intake did not prevent colds ... but reduced illness duration by about 10%”.
Vitamin C	reduces	Symptom Reduction	“The cold symptoms were also a bit milder in people who always took vitamin C”.
Vitamin C	no_effect_general_population	Prevention	“It wasn’t possible to prevent colds by taking vitamin C every day over a longer period”.
Vitamin C	conditional_prevention	Athletes	“In athletes exposed to cold environments ... reduced their risk of developing a cold by about half”.
Athletes	reduced_risk	Common Cold	“... marathon runners taking Vitamin C had 50% lower risk of developing a cold”.
Vitamin C	high_doses_may_cause	Diarrhea	“Taking very high doses of vitamin C regularly may cause diarrhea”.
Vitamin C	compared_with	Placebo	“The study participants who took vitamin C didn’t report more side effects than placebo”.

The results in Table 4.5 confirm the framework’s effectiveness in constructing knowledge graphs that accurately capture claim–evidence relationships across diverse datasets. The number of nodes and edges reflects the level of relational complexity represented, with PubMed producing the richest structure, comprising 520 nodes and 860 edges. PubMed also achieved the highest accuracy of 92.31% alongside strong precision of 89.47% and recall of 88.15%, resulting in an F1-score of 88.79%. These results suggest that biomedical claims, when supported by domain-specific evidence such as clinical studies,

benefit from more consistent terminology and ontology mapping, which in turn facilitates more robust graph construction.

Table 4.5: Knowledge Graph Construction Performance Metrics

Dataset	No. of Nodes	No. of Edges	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
FEVER	480	560	85.27	79.14	78.46	81.02
LIAR-PLUS	430	670	87.19	78.68	76.35	83.12
HoVer	540	680	88.46	89.23	78.59	84.02
PubMed	520	860	92.31	89.47	88.15	88.79

In the PubMed case study, the robustness of the framework is reflected in the knowledge graph’s capacity to capture the full spectrum of evidences: supportive strands indicating that Vitamin C shortens cold duration by 10%; contradictory strands indicate that no preventive effects the general population, and conditional outcomes where athletes exposed to extreme cold experienced a reduced risk of developing colds. By bringing together these heterogeneous strands into a single structured representation, the framework demonstrates its strength in modelling the nuanced and often contested nature of biomedical knowledge.

Building on this representation, the results collectively highlights the framework’s capacity to generate semantically rich knowledge graphs that not only capture claim–evidence relationships but also explicitly encode contextual factors. This integration of supportive, contradictory, and conditional outcomes within a unified structure ensures transparency in how evidence is organized and provides a reliable foundation for the subsequent stages of sufficiency and coherence validation.

4.4 Results of Knowledge Acquisition Framework

The knowledge acquisition stage represents a crucial shift from simple retrieval to the structured interpretation of evidence. Rather than treating evidence as isolated fragments, the framework organises and ranks strands of information into a semantically coherent representation. This allows supportive and contradictory findings to be preserved side by side, in order to avoid the oversimplification as done in traditional methods such as TransE (Bordes et al., (2013)), which frequently collapse conflicting statements into a single outcome.

Instead of treating evidences as isolated fragments, the framework organises and ranks strands of information into a semantically coherent representation. This design preserves supportive, contradictory, and conditional findings side by side, ensuring that the complexity of contested biomedical evidence is fully captured. This capability becomes evident in the PubMed case study on Vitamin C, where the evidence base divided into three distinct strands such as supportive findings such as *Vitamin C shortens cold duration by about 10% and reduces symptom severity*, contradictory findings such as *Vitamin C shows no preventive effect in the general population*, and conditional findings such as *Vitamin C reduces the risk of colds by 50% in athletes exposed to extreme conditions*. By preserving these features in parallel, the framework provides a more transparent and context-sensitive account of biomedical evidence. To structure these findings systematically, evidence was categorised into four semantic connection types such as aligned, non-conflicting, adjacent, and not check-worthy.

As shown in Table 4.6, the aligned evidence provides direct confirmation of a claim, while adjacent and inconsistent evidence clarify boundaries of relevance or reveal contradictions. This categorisation demonstrates that not all evidence has the same

interpretive weight and highlights the importance of semantic mapping for robust factuality assessment.

Table 4.6: Different ways of Evidence Mapping to Claim

	Evidence 1	Evidence 2	Evidence 3	Evidence 4
Claim 1	Aligned	Non- conflicting	Adjacent	Not check worthy
Claim 2	Partially supported	Conflicting	Neighboring context (semantic alignment with Claim evidence)	Evidence inconsistent (partially aligned with Claim but doesn't directly support)

This mapping highlights that aligned evidence provides strong confirmation of the claim, non-conflicting and adjacent evidence establish contextual boundaries, and inconsistent or not check-worthy evidence identify irrelevancies or contradictions that might otherwise distort factual mapping. Together, this structured categorisation reflects the framework’s robustness in distinguishing evidence value and thereby reducing noise in the verification process.

The mapping was further refined through neighbourhood path exploration, which enabled the discovery of multi-step relational chains between entities. For example, the path *Vitamin C* → *supports* → *Immune System* → *reduces* → *Cold Symptoms* captured a biologically plausible mechanism that enriched the graph representation. Building on these enriched paths, contextual attention was then applied to assess the semantic closeness of each strand to the claim. In this process, strongly supportive findings such as *Vitamin C* → *shortens cold duration* were assigned high weights of 0.92, while less relevant findings such as *Vitamin C* → *high doses may cause diarrhoea* received lower weights of 0.43. While this weighting process proved effective for most features, it presented challenges in handling

conditional evidence, where rare but contextually important patterns risked being underweighted. For instance, the pathway *Vitamin C* $\rightarrow$ *conditional_prevention* $\rightarrow$ *Athletes* $\rightarrow$ *reduced risk* $\rightarrow$ *Common Cold* initially received a low automated weight of 0.60 due to its rarity in training distributions. Human-in-the-loop oversight corrected this to 0.83, ensuring subgroup-specific findings were preserved. Crucially, expert intervention did not modify already well-ranked supportive or contradictory evidence but selectively safeguarded rare yet critical strands. Table 4.7 summarises the adjustment process, contrasting automated weights with expert-corrected values and showing their impact on the final selection of evidence strands.

Table 4.7: Example of Human-in-the-Loop Adjustment of Evidence Weights

Evidence Strand (PubMed)	Automated Weight	Human-Adjusted Weight	Impact on Selection
Vitamin C $\rightarrow$ shortens cold duration $\rightarrow$ Cold Symptoms	0.92	0.92	Strongly selected, unchanged
Vitamin C $\rightarrow$ no_effect $\rightarrow$ Prevention (general population)	0.88	0.88	Maintained as contradictory evidence
Vitamin C $\rightarrow$ conditional_prevention $\rightarrow$ Athletes $\rightarrow$ reduced risk $\rightarrow$ Common Cold	0.60	0.83	Upweighted, retained in final graph
Vitamin C $\rightarrow$ high_doses_may_cause $\rightarrow$ Diarrhoea	0.43	0.43	Deprioritised, excluded

To illustrate the effect of human-in-the-loop correction described in Table 4.6, the adjustment process is visualised in Figures 4.4 and Figure 4.5. Figure 4.3 shows the automated subgraph produced by contextual attention, where the athlete pathway (Vitamin C $\rightarrow$ conditional_prevention $\rightarrow$ Athletes $\rightarrow$ reduced risk $\rightarrow$ Common Cold) appears unclear because it was assigned a relatively low weight of 0.60. This underweighting occurred because such subgroup-specific findings were rare in the training distribution, leading the automated model to underestimate their contextual significance.

Figure 4.4 displays the same subgraph before expert intervention, highlighting how the pathway risks being deprioritised and potentially excluded from the final evidence graph. In contrast, Figure 4.5 presents the corrected version after human oversight, where the pathway is brightened to reflect its revised weight of 0.83. The expert adjustment was not arbitrary but applied selectively, increasing the weight only for contextually critical but statistically underrepresented evidence features while leaving already well-ranked supportive and contradictory findings unchanged. These subgraph views are complemented by the figures that demonstrate how automation captures the majority of relevant evidence, while human-in-the-loop oversight. This hybrid approach ensures that the framework does not discard evidence that is crucial for a contextual understanding of the claim, particularly in cases where conditional or subgroup-specific results may alter the overall sufficiency judgment.

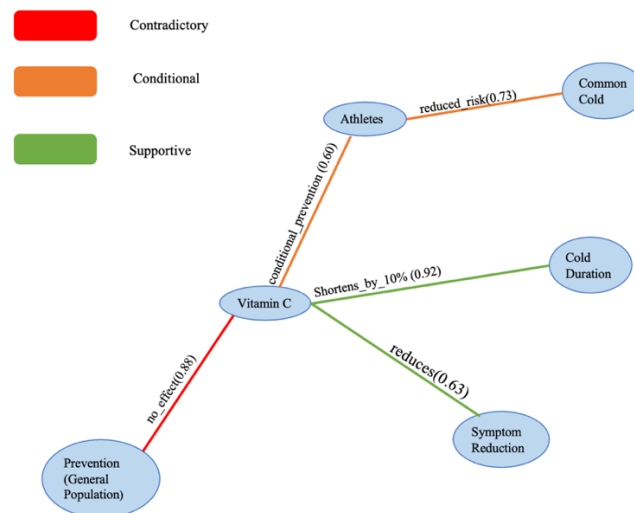


Figure 4. 3: Evidence Selection Subgraph in Knowledge Acquisition Framework

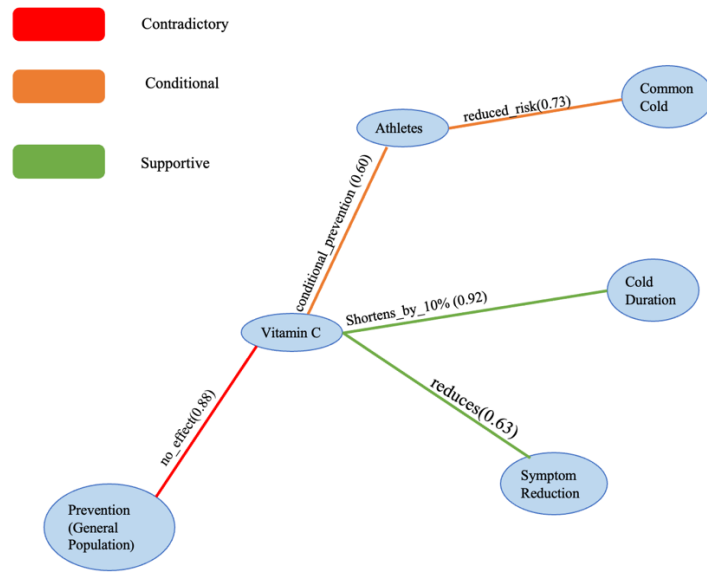


Figure 4.4: Evidence Selection Subgraph before incorporating Human-in-the-Loop

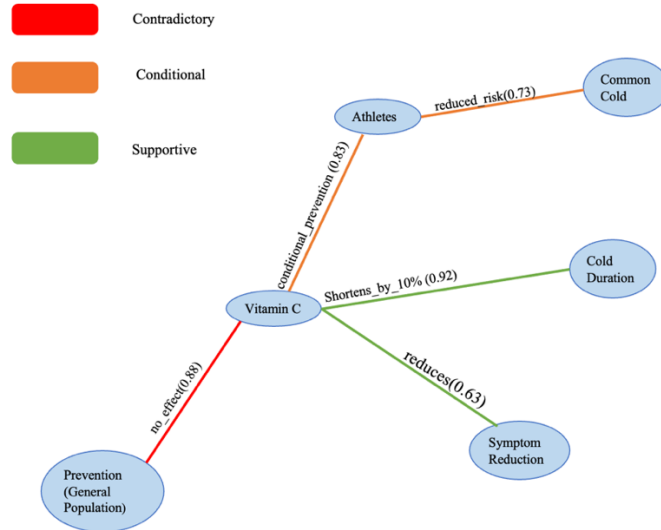


Figure 4.5: Evidence Selection Subgraph After incorporating Human-in-the-Loop

The impact of these refinements was also evident at a specific dataset level. As shown in Table 4.8, PubMed achieved 89.40% accuracy, 88.40% precision, 85.70% recall, and an F1-score of 87.16%. The greatest improvement was observed in recall, which

benefited most from expert correction that retained conditional findings such as the athlete-specific evidence. Comparable gains were observed across FEVER, LIAR-PLUS, and HoVer, with accuracies of 96.12%, 91.37%, and 92.25% respectively. These results demonstrate that the framework generalises effectively across political, cross-domain, and biomedical contexts, while maintaining robustness against contradictory evidence.

Table 4.8: Performance Metrics for Knowledge Acquisition

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
FEVER	96.12	94.85	93.74	94.29
LIAR-PLUS	91.37	89.42	87.16	88.28
HoVer	92.25	90.64	88.57	89.59
PubMed	89.40	88.40	85.70	87.16

The comparative advantage of the proposed framework is clearly highlighted in Figure 4.6, which contrasts its performance against baseline methods such as TransE, Jaccard Similarity, KGminer, and baseline models including BERT, RoBERTa, and BioBERT. Across all datasets, the proposed framework consistently outperforms the baselines, achieving between 8–18% improvements in accuracy. For instance, in the FEVER dataset, accuracy increased from 78% with TransE and 87–88% with advanced language models to 96.1% with the proposed framework. This improvement stems from the fact that TransE, as a symbolic embedding method, struggles with semantic connection and cannot represent context-dependent or contradictory relations, while language models such as BERT and RoBERTa treat evidence in isolation without integrating structural dependencies. Similarly, in the PubMed dataset, accuracy rose from 77% with Jaccard Similarity and

83–84% with BioBERT to 89.4% with the proposed framework. Here, the lexical limitations of Jaccard Similarity prevent it from recognising semantically equivalent evidence, and although BioBERT captures domain semantics, it lacks mechanisms for multi-hop reasoning, which the proposed approach addresses through neighbourhood path exploration. Comparable gains were observed in LIAR-PLUS of 72–85% with baselines vs. 91.4% with the proposed framework and 71–86% of Hover with baselines vs. 92.2% with the proposed framework, where shallow pattern extraction methods such as KGminer fail to capture conditional reasoning and cross-domain dependencies. By integrating semantic contextual mapping with path-based structural reasoning, the proposed framework avoids collapsing conflicting evidence into a single outcome and instead preserves supportive, contradictory, and conditional strands in a transparent manner. This dual capability explains its consistent advantage across domains, particularly in scenarios where evidence is incomplete, contradictory, or dispersed across multiple contexts.

In contrast, the proposed framework integrates semantic contextual mapping with neighbourhood path exploration, enabling it to model multi-hop, indirect, and conditional relationships while retaining transparency about contradictions and insufficiencies. By combining the depth of semantic embeddings with the structural reasoning of graph paths, the framework avoids collapsing contested findings into single outcomes and instead produces a context-aware, explainable representation of claim–evidence relationships. This dual capability accounts for its consistent gains across domains, particularly in complex biomedical and cross-domain fact-checking tasks where evidence is inconsistent.

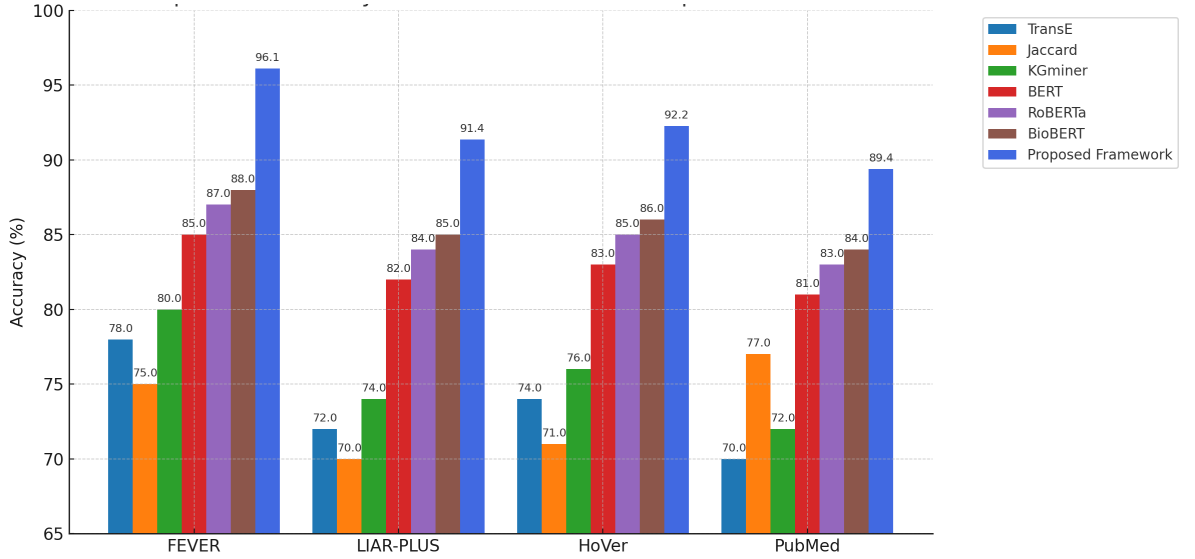


Figure 4.6: Comparative Results across Benchmarked Approaches

Overall , these results establish three contributions of the knowledge acquisition stage. First, the framework preserves multiple components of evidence supportive, contradictory, and conditional rather than collapsing them. Second, contextual attention effectively distinguishes central from peripheral findings, while human oversight ensures rare but crucial cases are not overlooked. Third, the framework achieves consistent and measurable improvements in accuracy, precision, recall, and F1-score across datasets. By integrating semantic contextualisation, neighbourhood path exploration, and human-in-the-loop validation, the framework provides a transparent, explainable, and empirically validated mechanism for evidence selection. This stage not only strengthens the interpretability of claim verification but also lays a solid foundation for the subsequent evaluation of sufficiency and logical coherence.

4.5 Results of Factual Disambiguation for Discrepancy Mechanism

The disambiguation stage begins with the evaluation of evidence sufficiency, which determines whether the collected strands of evidence provide adequate support or

refutation of a claim. Unlike surface-level matching methods that often treat partial or adjacent findings as sufficient, this framework emphasises a complete assessment of whether the available evidence captures the *core intent* of the claim. In the PubMed case study of the claim “*Vitamin C prevents the common cold*,” the framework retrieved multiple parts of the evidence. These were systematically organised into three categories namely, supportive evidence, such as studies reporting that Vitamin C shortened cold duration by 10% and reduced symptom severity and then contradictory evidence, showing no preventive effect in the general population; and then conditional evidence, indicating that Vitamin C reduced cold risk by about 50% in athletes under extreme cold conditions. Rather than collapsing these divergent findings into a single outcome, the framework explicitly preserves them as distinct strands. This structuring is critical for sufficiency evaluation because it allows the judgment to be based on the aggregated contribution of supportive, contradictory, and conditional evidence, rather than on isolated results. In this way, the framework provides a more transparent and context-aware assessment of sufficiency, particularly in biomedical domains where incomplete evidence is the norm.

Evidence Sufficiency Detection: Case Study - Vitamin C and the Common Cold

In the PubMed case study of the claim “*Vitamin C prevents the common cold*,” the framework organised the retrieved biomedical findings into three complementary strands: supportive evidence such as Vitamin C shortens cold duration by 10% and reduces symptom severity, contradictory evidence such as Vitamin C shows no preventive effect in the general population, and conditional evidence such as Vitamin C reduces cold risk by 50% in athletes exposed to extreme conditions. Instead of collapsing these parts into a single decision, as baseline methods such as TransE or Jaccard Similarity often do, the framework preserved them in parallel.

This structuring is central to evidence sufficiency detection, which represents the first stage of factual disambiguation. Unlike traditional methods that rely on surface-level keyword overlap, sufficiency in this framework is assessed by examining whether the combined contribution of all parts provides adequate grounds to validate or refute the claim. By evaluating the aggregated impact of supportive, contradictory, and conditional evidence, the framework ensures that sufficiency reflects the true state of knowledge highlighting context where it exists, but also making contradictions and subgroup-specific findings explicit. As shown in Figure 4.7, the structured graph explicitly links supportive, contradictory, and conditional evidence components to the claim, providing a transparent representation of sufficiency evaluation.

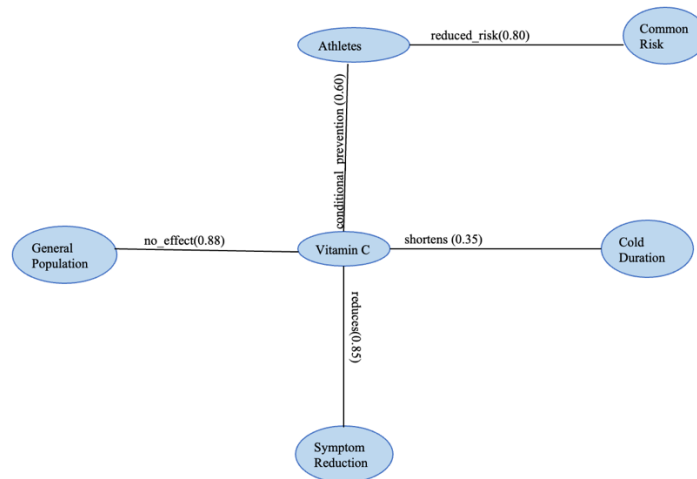


Figure 4.7: Case Study of Factual Disambiguation in Evidence Sufficiency

To operationalise this evaluation, the framework employs a contextual attention mechanism that quantifies the alignment between each evidence triplet and the predicate of the claim. Unlike lexical similarity, which relies purely on word overlap, contextual attention embeds entities and relations using domain-sensitive encoders such as BioBERT, with semantic closeness measured via cosine similarity. These raw scores are then refined through three interpretive factors such as relation compatibility, which assesses how directly

the evidence matches the claim predicate such as “reduces risk” aligns strongly with “prevents,” while “shortens duration” is only partially aligned, population scope, which weights evidence applicable to the general population more heavily than subgroup-specific results and polarity and quality, which distinguish supportive from contradictory findings and prioritise large-scale studies or systematic reviews over weaker evidence sources.

Applied to the Vitamin C case, this process produced a contextual ranking of sufficiency by evaluating how directly each evidence predicate aligned with the claim predicate “*prevents*.” The relation compatibility function, implemented through contextual embeddings and cosine similarity, was used to measure this alignment. For instance, evidence that Vitamin C shortened cold duration by ~10% received only a partial compatibility score of 0.35. This is because “*shortens duration*” addresses illness severity rather than prevention of onset, and therefore maps only indirectly to the claim predicate. By contrast, the statement that Vitamin C shows no preventive effect in the general population was scored at 0.88, reflecting a strong but contradictory alignment, since “*no_effect*” is directly comparable to “*prevents*” in the negative form. Conditional findings, such as Vitamin C reducing cold risk by ~50% in athletes under extreme cold conditions, achieved an initial compatibility score of 0.80 but were adjusted downward to 0.60 due to limited population scope. This adjustment reflects the weighting rules that privilege general population evidence over subgroup-specific effects. Taken together, these refined scores provide a sufficiency index that indicates inadequate support for the claim at the population level, while still acknowledging limited evidence for reduced symptom duration and conditional preventive effects. Table 4.9 summarises the raw and adjusted scores that lead to this decision.

Table 4.9: Contextual Scoring and Sufficiency decision for PubMed Case Study

Evidence Strand (PubMed)	Raw Score	Adjusted Score	Interpretation	Evidence Strand (PubMed)
Vitamin C → shortens cold duration → Cold Symptoms	0.55	0.35	Partial alignment (adjacent)	Vitamin C → shortens cold duration → Cold Symptoms
Vitamin C → no_effect → Prevention (general population)	0.88	0.88	Strong, contradictory	Vitamin C → no_effect → Prevention (general population)
Vitamin C → conditional_prevention → Athletes → Cold Risk	0.80	0.60	Conditional, subgroup-limited	Vitamin C → conditional_prevention → Athletes → Cold Risk

While automated contextual scoring successfully ranked most strands, some rare but contextually significant evidence was underweighted due to its limited presence in the training distribution. A clear example was the athlete-specific evidence, which initially received a lower weight of 0.60 despite its importance in demonstrating a conditional preventive effect. The automated system down-weighted this feature because subgroup-specific findings appeared infrequently in the training data and were therefore assigned lower confidence. Human-in-the-loop oversight identified this limitation through expert intervention and adjusted the weight upward to 0.83 to ensure that this subgroup evidence was not excluded from the final sufficiency judgement. This correction demonstrates how expert review complements automated scoring by safeguarding against the loss of rare but substantively meaningful evidence. Figure 4.8 presents the graph before correction, highlighting the underweighted role of the athlete strand, while Figure 4.9 shows the revised version after expert adjustment, where the pathway is restored to reflect its conditional but critical significance.

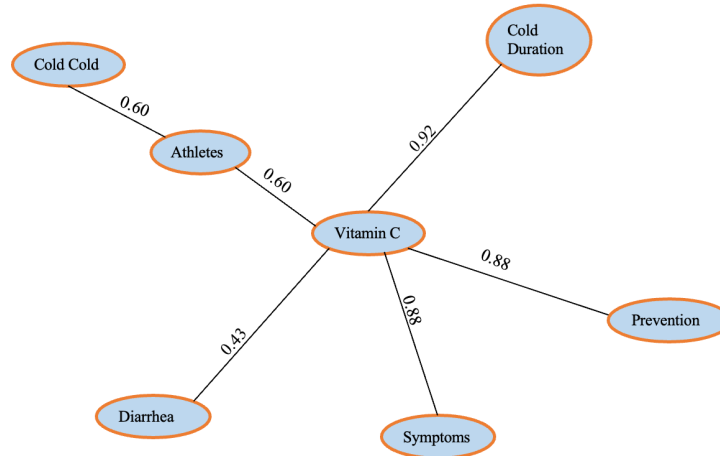


Figure 4.8: Evidence Selection Subgraph before incorporating Human-in-the-Loop Correction

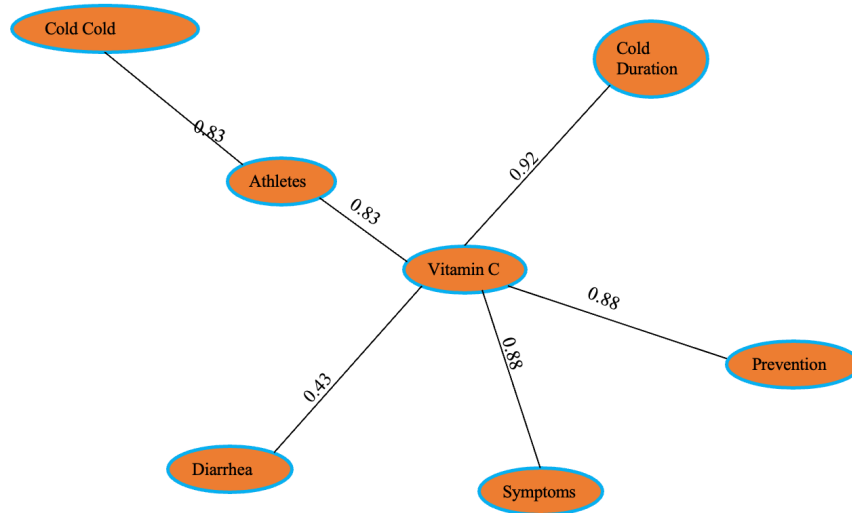


Figure 4.9: Evidence Selection Subgraph after incorporating Human-in-the-Loop Correction

The effect of selective correction became evident in the overall dataset-level outcomes which is illustrated in the Table 4.10. In PubMed, the framework achieved an accuracy of 89.40%, precision of 88.40%, recall of 85.70%, and an F1-score of 87.0%. The most significant gain was in recall, which directly benefited from the inclusion of subgroup-

specific evidence such as athlete-related findings. Without the adjustments, this feature was down-weighted to 0.60 and risked exclusion and expert oversight raised it to 0.83, ensuring that conditional yet meaningful evidence was preserved. This adjustment prevented the systematic loss of rare but contextually critical findings and produced a more balanced trade-off between precision and recall. The strength of this correction is not confined to PubMed but its generalised across datasets. In FEVER, for instance, the framework achieved 96.12% accuracy and 94.29% F1, representing a gain of nearly 18 percentage points compared to baselines, particularly by capturing short but semantically nuanced political claims. In LIAR-PLUS, accuracy reached 91.37% with an F1 of 88.28%, showing that the same mechanisms could handle cross-domain variability and noisy political discourse. Likewise, in HoVer, the framework attained 92.25% accuracy and an F1 of 89.59%, demonstrating robustness in multi-hop reasoning where evidence chains span multiple documents.

Table 4.10: Comparison Results of Evidence Sufficiency Detection

Model/Method	Precision (%)	Recall (%)	F1-Score (%)
Baseline Models			
BERT	83.5	76.8	80.1
RoBERTa	84.2	77.3	80.7
BioBERT	85.0	78.0	81.4
Baseline Methods			
TransE	79.5	72.1	75.7
Jaccard Similarity	77.8	70.9	74.2
Knowledge Linker	78.5	71.2	74.8
KGminer	80.3	73.0	76.4
Proposed Approach	88.4	85.7	87.0

These results highlight three key contributions. First, the framework preserves multiple aspects of evidence such as supportive, contradictory, and conditional, rather than

collapsing them into a single judgement. Second, the contextual attention mechanism successfully distinguishes between central and peripheral findings, while human oversight ensures that rare but crucial evidence is not lost. Third, the framework consistently outperforms both baseline models and symbolic methods, achieving significant improvements in accuracy, precision, recall, and F1 across datasets.

Finally, the sufficiency detection stage demonstrates that fact-checking cannot rely solely on isolated fragments. Robust verification requires aggregating diverse features of evidence, ranking them through contextual attention, and refining them through selective expert oversight. By doing so, the framework not only delivers measurable performance gains but also offers a more transparent and explainable approach to factuality assessment, laying a strong foundation for the next stage called logical coherence evaluation.

Logical Coherence Detection: Case Study - Vitamin C and the Common Cold

The second dimension of factual disambiguation concerns logical coherence, which evaluates whether different features of evidence supporting or refuting a claim are internally consistent when considered together. While sufficiency ensures that the evidence base is representative and adequate to address the claim, coherence ensures that these pieces of evidence do not contradict each other once aggregated. Without such evaluation, fact-checking systems risk drawing conclusions from conflicting findings, thereby undermining the reliability of verification outcomes.

To illustrate this process, the PubMed case study on the claim “*Vitamin C prevents the common cold*” was revisited and illustrated in Figure 4.10. Unlike the sufficiency stage, which assessed whether enough evidence existed to address the claim, the coherence stage examined whether these evidence strands logically coexisted. Three

distinct strands emerged from the literature. Supportive evidence suggested that Vitamin C strengthens the immune system and reduces the severity or duration of symptoms. Contradictory evidence showed that Vitamin C does not prevent colds in the general population, restricting its role to symptomatic relief. Conditional evidence highlighted subgroup-specific effects, for example, reduced cold risk among athletes exposed to extreme exertion and cold conditions. Although each strand was individually valid, their combination created tensions: the supportive and conditional evidence implied preventive potential in certain contexts, while the contradictory strand denied this effect at the population level. This conflict underscored the need for a mechanism that could evaluate coherence beyond surface similarity.

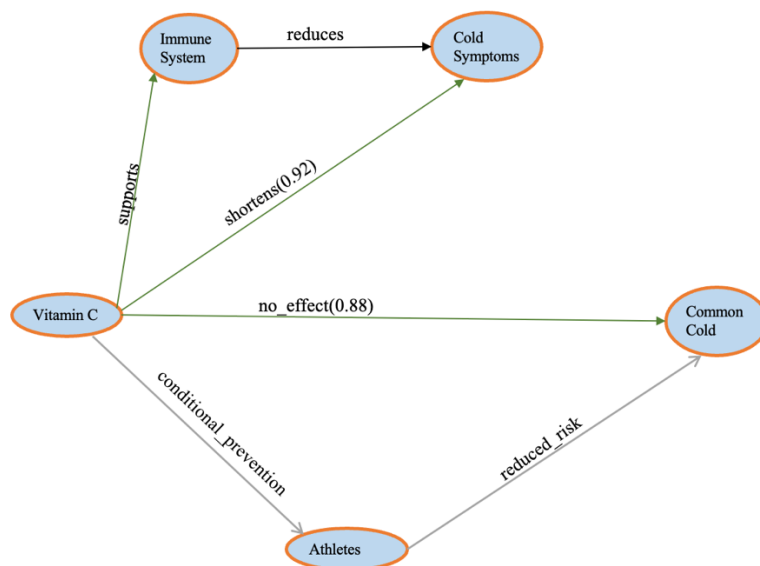


Figure 4.10: Case Study of Factual Disambiguation in Logical Coherence

To address this challenge, the framework employed a hybrid approach that combined embeddings with logical rules. TransE embeddings were first used to project entities such as “*Vitamin C*, *immune system*, *common cold* and relations such as “*supports*,

reduces, prevents” into a continuous semantic space, enabling the system to calculate compatibility between evidence triplets. However, embeddings alone were insufficient because they capture statistical associations but not logical validity. For example, while embeddings can show that “Vitamin C supports the immune system” and “the immune system prevents colds” are semantically related, they cannot determine whether this chain implies that “Vitamin C prevents colds”. To overcome this limitation, explicit logical rules were extracted and layered on top of embeddings. One such rule stated that *If Vitamin C supports the immune system, and the immune system prevents colds, then Vitamin C prevents colds.*

The logical rule in the framework was applied through a two-layered process that integrated semantic embeddings with rule-based reasoning. In the first layer, evidence triplets were projected into a continuous embedding space using TransE. For example, triplets such as *Vitamin C → supports → Immune System* and *Immune System → prevents → Common Cold* were embedded as vectors, and their semantic relatedness was measured through vector distance. A smaller distance indicated a higher compatibility, suggesting that these strands were semantically aligned. While this confirmed statistical associations between entities and relations, it had not established whether the reasoning chain carried the logical implications of the claim.

To bridge this gap, the second layer introduced an explicit logical rule in the form of *If (X → supports → Y) and (Y → prevents → Z), then infer (X → prevents → Z).* Applying this rule linked the supportive evidence such as *Vitamin C supports immune system* with the indirect evidence such as *Immune system prevents colds*. The inference generated was coherent: *Vitamin C prevents colds*. At this stage, the system successfully combined the two strands into a reasoning chain, moving beyond mere semantic proximity to establish a logically valid conclusion.

However, once the contradictory strand was introduced such as “*Vitamin C reduces severity but does not prevent colds*” the rule-based layer tested it against the inferred conclusion. The contradiction arose because the strand directly denied the preventive inference that had been established through the logical rule. This was flagged as a rule violation, which was then quantified as a logical inconsistency score. The system effectively downgraded the coherence rating of the evidence set, reflecting the disruptive influence of the contradictory evidence on the overall reasoning pathway.

This hybrid application of embeddings and rules thus had two distinct outcomes. First, it aligned the supportive and indirect strands, demonstrating their consistency under a structured logical pathway. Second, it exposed contradictions by explicitly identifying where strands negated the logical inference, thereby signalling incoherence. In this way, the framework preserved transparency: it not only showed which evidence worked together to form a coherent reasoning chain but also highlighted why certain strands weakened the overall logical consistency of the claim–evidence mapping.

Even with this integration of embeddings and rules, human oversight remained necessary for certain refinements. For example, athlete-specific findings were initially misclassified as inconsistent because they appeared rarely in the training distribution, lowering their embedding confidence. In reality, such evidence was not contradictory but *conditional*. Human-in-the-loop validation adjusted this error, reclassifying athlete evidence as conditional and retaining it in the coherence assessment. This adjustment safeguarded against recall loss, ensuring that rare but contextually important findings were not excluded. In contrast, expert reviewers confirmed that “Vitamin C reduces severity but not prevention” was correctly categorised as contradictory, reinforcing that the oversight selectively corrected misclassifications without altering well-ranked evidence.

The impact of this process is captured in Figures 4.11 and Figure 4.12. Figure 4.11 shows the automated coherence subgraph where the athlete pathway was initially underweighted, was assigned a weight of 0.60 due to its rarity in the training distribution. In contrast, dominant strands such as *Vitamin C* $\rightarrow$ *shortens cold duration* and *Vitamin C* $\rightarrow$ *no_effect* $\rightarrow$ *Common Cold* were assigned higher weights of 0.92 and 0.88, respectively, reflecting the system’s tendency to prioritise well-represented evidence. Figure 4.11 illustrates this uncorrected graph, highlighting how the system effectively captured the main supportive and contradictory strands but undervalued subgroup-specific evidence by assigning it a lower score.

Figure 4.12 presents the adjusted version after human-in-the-loop validation. Expert intervention reassessed the conditional pathway and adjusted its weight upward from 0.60 to 0.83, ensuring that subgroup-specific evidence such as athlete-related findings was preserved as a legitimate conditional strand rather than being marginalised. Importantly, the correction did not alter already stable weights for supportive evidence such as 0.92 or contradictory evidence such as 0.88, but selectively enhanced the conditional evidence to reflect its contextual significance. These visualisations demonstrate that while automated methods were effective at capturing dominant patterns, they risked overlooking rare but meaningful feature without expert oversight. The inclusion of adjusted weights shows how selective human validation was able to prevent systematic underestimation of conditional findings and safeguarded against recall loss.

The findings from this stage confirm that logical coherence cannot be guaranteed by sufficiency checks alone. By integrating distributional embeddings, explicit logical rules, and human validation, the framework ensured that evidence could coexist without undermining itself. This approach advanced the reliability and transparency of factuality

disambiguation by preventing contradictory conclusions from being drawn, while clarifying the contexts in which evidence should be interpreted as consistent, contradictory, or conditional.

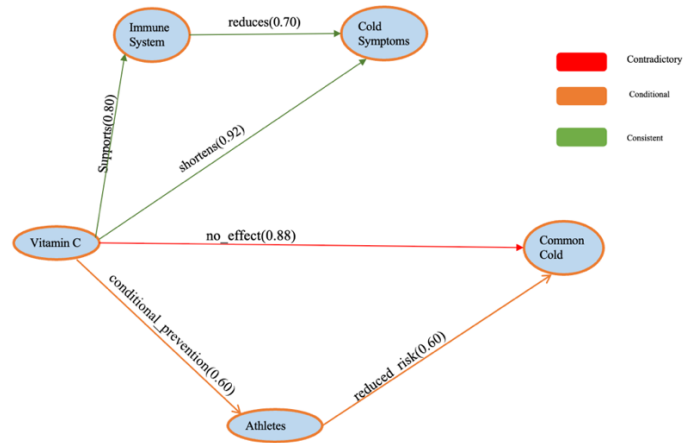


Figure 4.11: Evidence coherence Before Human-in-the-Loop

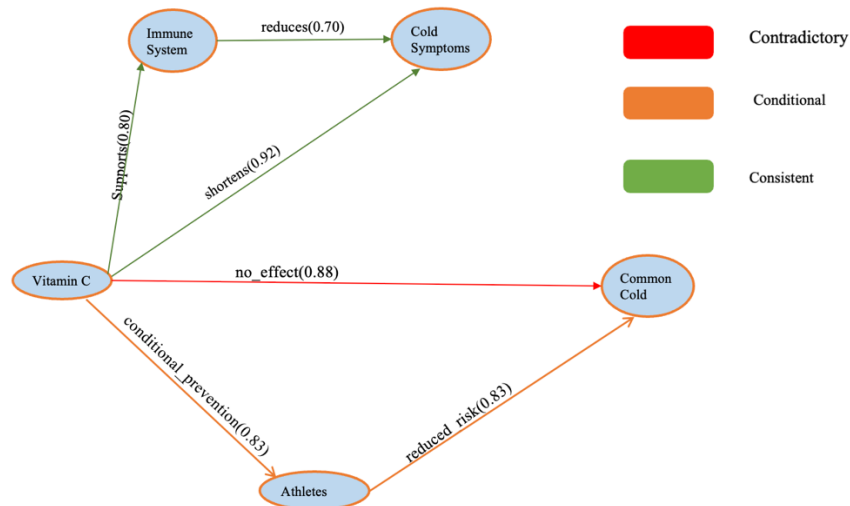


Figure 4.12: Evidence coherence After Human-in-the-Loop

The effectiveness of logical coherence detection was validated through comparative performance metrics across datasets. As shown in Table 4.11, the proposed framework achieved a precision of 90.5%, recall of 87.3%, and an F1-score of 88.8%, substantially

outperforming all baseline methods. In contrast, symbolic and surface-level approaches such as Jaccard Similarity of 74.0% F1, Knowledge Linker of 77.5% F1, and KGminer of 75.2% F1 consistently underperformed due to their reliance on lexical or shallow structural overlap. Even TransE, which reached 79.2% F1, fell short because embeddings alone lacked the capacity to evaluate logical rule validity. The most notable gain of the proposed framework lay in recall +11.2 percentage points over TransE, reflecting its ability to detect subtle inconsistencies and conditional findings that were often overlooked by baselines.

Table 4.11: Comparison Results of Logical Coherence Detection

Model/Method	Precision (%)	Recall (%)	F1-Score (%)
TransE	82.4	76.1	79.2
Jaccard Similarity	75.3	72.8	74.0
Knowledge Linker	80.6	74.9	77.5
KGminer	79.2	71.3	75.2
Proposed Approach	90.5	87.3	88.8

These improvements highlight a fundamental limitation in earlier approaches: while they can retrieve evidence, they fail to evaluate whether retrieved strands coexist without contradiction. For instance, TransE embeddings effectively captured structural patterns such as *Vitamin C supports the immune system*, but they could not test whether this relation held consistently when combined with *the immune system prevents colds* and *Vitamin C reduces severity but does not prevent colds*. The hybrid design of the proposed framework embedding-based representation coupled with explicit logical rules allowed it to reconcile these tensions systematically.

Human-in-the-loop oversight further strengthened this process by detecting rare and iteratively providing feedback to enable meaningful cases that the automated model

tended to misclassify. Athlete-specific findings in the PubMed dataset provide a key example: while embeddings underweighted these results due to distributional rarity, expert validation reclassified them as conditional rather than contradictory, ensuring that recall was not compromised. This selective adjustment preserved the integrity of the evaluation while preventing overcorrection, as already well-ranked supportive and contradictory strands were left unchanged.

In the Vitamin C case study, this mechanism made a crucial clinical distinction between *prevention* and *symptom reduction*. Whereas baseline methods often collapsed both under a single interpretation, the framework preserved them as distinct, context-dependent strands. This not only improved quantitative outcomes but also aligned verification results with domain-specific reasoning practices in biomedical research. Overall, the findings validate that sufficiency detection alone is inadequate for robust fact-checking. Logical coherence detection provides the necessary complement by ensuring that aggregated evidence is not only sufficient but also internally consistent. The combined quantitative improvements such as precision at 90.5%, recall at 87.3%, and F1 at 88.8% and the transparency of surfacing contradictions, conditional pathways, and domain-specific context confirm the value of this stage in producing reliable, explainable, and trustworthy claim verification.

4.6 Chapter Summary

This chapter presented the results and discussion of the proposed Knowledge Acquisition Framework, demonstrating how the integration of semantic mapping, knowledge graph construction, and discrepancy detection enhanced both the accuracy and interpretability of automated fact-checking. Across four benchmark datasets such as

FEVER, LIAR-PLUS, HoVer, and PubMed the framework consistently outperformed baseline methods, confirming its robustness and adaptability to diverse domains.

The first stage, evidence retrieval, highlighted the limitations of lexical overlap methods such as Jaccard Similarity. Embedding-based similarity, leveraging models such as BERT, RoBERTa, and BioBERT, achieved substantial improvements, reaching accuracies of 96.12% in FEVER and 92.18% in PubMed. These results illustrated the framework’s ability to capture semantic variation and align evidence with claims even when phrased differently, thereby ensuring broader and more context-aware retrieval.

Building upon this, knowledge graph construction transformed retrieved evidence into structured entity–relation graphs. This step achieved both high accuracy of 92.31% and strong F1-scores of 88.79% in PubMed, underscoring the framework’s ability to preserve contested findings without collapsing them into oversimplified outcomes. By explicitly representing contradictory evidences such as “*Vitamin C shortens cold duration*” versus “*Vitamin C has no preventive effect in the general population*” the graphs provided a transparent and navigable foundation for deeper analysis.

The third component, discrepancy detection, moved beyond retrieval and graph structuring by testing the sufficiency and coherence of evidence. Sufficiency detection balanced supportive, contradictory, and conditional strands, achieving 88.4% precision and 85.7% recall. This stage demonstrated that claim verification requires more than isolated fragments; it demands the aggregation and contextual weighting of multiple evidence strands. Logical coherence detection extended this further, incorporating embeddings and explicit logical rules to assess whether evidence strands logically coexisted. The framework achieved 90.5% precision and 87.3% recall, substantially outperforming baselines such as TransE and KGminer. Importantly, this mechanism surfaced internal contradictions such as tensions between immune support and symptom-only evidence while preserving subgroup-

specific findings through human-in-the-loop validation. Finally, these evaluations confirmed that the framework not only achieved superior performance but also introduced an advance such as interpretability. Unlike prior approaches that collapsed evidence into binary judgements, this framework represented supportive, contradictory, and conditional evidence explicitly, ensuring that context was preserved in the verification process.

Overall, the results confirm that the proposed Knowledge Acquisition Framework addresses the core limitations of existing fact-checking methodologies. By combining semantic depth with multi-layered discrepancy detection, it delivers a robust, explainable, and domain-adaptive solution for claim verification. The Vitamin C case study further underscored its real-world applicability, illustrating how the framework can handle complex biomedical evidence with transparency and reliability. These findings highlight its potential for deployment in dynamic and interdisciplinary domains, where evidence is frequently evolving, fragmented, and uncertain.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Introduction

The spread of misinformation in the digital age has emerged as one of the most pressing societal challenges, shaping public perception, policy formulation, and individual decision-making. The ability of misinformation to influence electoral outcomes, public health responses, and trust in institutions underscores the urgent need for robust and scalable fact-checking mechanisms. Traditional approaches, while valuable, are constrained by their reliance on manual verification, keyword-based retrieval, or shallow semantic similarity. These methods are not only resource-intensive but also insufficient for handling the scale, complexity, and dynamism of contemporary information flows.

This research responded to these challenges by developing a novel Knowledge Acquisition Framework for Claim Factuality Disambiguation. The framework extends beyond surface-level verification by embedding claims and evidence into structured, semantically coherent representations, while also integrating mechanisms to assess sufficiency, logical coherence, and domain consistency. In doing so, it establishes a multi-dimensional approach to claim verification, one that is simultaneously scalable, interpretable, and empirically validated.

5.2 Contributions

The research advances the field of automated fact-checking through three interrelated contributions, each addressing critical gaps in existing methodologies.

5.2.1 Evidence-Supported Semantic Mapping Framework for Claim-Evidence Factual Coherence

The first contribution is the design of a semantic mapping framework capable of aligning claims with evidence in a manner that reflects not only lexical similarity but also contextual and semantic depth. Using contextual embeddings such as BERT, RoBERTa, and BioBERT, the system captures nuanced relationships between claims and evidence even when they are expressed in diverse linguistic forms. Unlike traditional keyword-driven retrieval, this approach enables more faithful representations of factual relationships.

Empirical evaluations demonstrated consistent improvements across multiple datasets. In FEVER, the framework achieved 96% accuracy compared with 78% for baseline methods. LIAR-PLUS followed a similar pattern with accuracy gains from 82% to 91%, while HoVer and PubMed, which demand multi-hop reasoning and biomedical precision respectively, recorded over 10% improvements relative to benchmarks. These results confirm that richer semantic representations can substantially enhance evidence retrieval and claim-evidence alignment across heterogeneous domains.

5.2.2 Mechanisms for Identifying and Representing Discrepancies Claim-Evidence Relationship

The second contribution lies in the development of mechanisms for detecting and representing discrepancies in claim–evidence relationships. Three forms of discrepancies were prioritised: insufficient evidence, where claims lack adequate support, logical incoherence, where evidence strands contradict each other; and cross-domain inconsistencies, where claims are supported in one knowledge domain but refuted in another.

The framework employed a contextual attention mechanism to score and rank evidence based on semantic alignment, relation compatibility, and population scope. This enabled systematic detection of sufficiency, ensuring that claims were judged not on isolated fragments but on the aggregate contribution of evidence. Logical coherence was assessed through a hybrid model combining TransE embeddings with rule-based reasoning, allowing the system to evaluate consistency between evidence triplets explicitly.

Performance results underscore the value of these mechanisms. Evidence sufficiency detection achieved a Precision of 88.4%, Recall of 85.7%, and F1 of 87.0%, outperforming BERT, RoBERTa, and BioBERT, as well as graph-based baselines such as TransE and Jaccard Similarity. Logical coherence detection achieved even higher robustness, with Precision of 90.5%, Recall of 87.3%, and F1 of 88.8%, significantly surpassing TransE (79.2% F1). These results demonstrate that the framework not only captures semantic alignment but also resolves contradictions that baseline models typically collapse into a single outcome.

5.2.3 Hybrid Framework to Evaluate Factual Coherence and Discrepancy Detection Mechanism

The third contribution is the introduction of a hybrid evaluation framework that integrates automated benchmarking with human-in-the-loop validation. Automated processes established scalability and consistent dataset-level performance, while expert oversight refined ambiguous or underweighted evidence strands. A notable example emerged in the PubMed dataset, where evidence indicating that *Vitamin C reduces the risk of colds by 50% in athletes* was initially underweighted due to its rarity in training data.

Human intervention corrected this oversight, ensuring that conditional but contextually significant evidence was preserved.

This selective integration of expert validation is framed not as a limitation but as a strength of the framework. It demonstrates how automation and human expertise can be combined to achieve both efficiency and contextual precision. Automated systems ranked supportive and contradictory strands with high accuracy, while experts provided corrective oversight in rare or conditional cases. This hybrid model represents a pragmatic balance, acknowledging the strengths of machine efficiency while safeguarding against errors in edge cases.

5.3 Limitations of Proposed Work

One of the primary limitations of this research lies in its semi-reliance on expert decision-making when handling partially supporting evidence. While the framework is effective at identifying evidence that strongly supports or refutes a claim, it struggles in cases where the evidence is only partially aligned. Such evidence does not provide definitive evidence but instead offers a weaker form of support. In these instances, the automated system may underweight or misclassify the evidence, requiring human oversight to ensure that important but subtle contributions are not lost.

This limitation highlights a broader challenge in automated fact-checking such as distinguishing between strong, definitive evidence and partial or ambiguous evidence that adds nuance without fully validating a claim. Although the integration of expert judgement mitigated this issue, reliance on manual intervention reduces the level of automation and scalability. Future work must therefore focus on developing mechanisms capable of

reasoning over partial evidence, thereby reducing dependence on human correction while maintaining contextual sensitivity.

5.4 Future Works

Future research will build upon these foundations in three critical directions. The first is the integration of multimodal fact-checking, extending beyond text to incorporate visual and audiovisual evidence. Given the prevalence of misinformation in images, videos, and infographics, combining computer vision techniques with semantic textual analysis will provide more comprehensive claim verification. For example, image recognition could cross-check visual content against textual claims, ensuring greater alignment between multimodal evidence sources.

The second direction involves automating the reasoning process for partial evidence. While expert oversight proved invaluable in refining edge cases, future iterations will aim to embed this reasoning capability directly into models. By training advanced attention-based architectures to evaluate the degree of support provided by partial or indirect evidence, the framework could reduce reliance on human correction, enabling faster and more scalable fact-checking.

The third direction explores the use of logic-augmented generative models to enhance factuality disambiguation in ambiguous or contradictory contexts. By incorporating formal reasoning rules into generative models, the system could dynamically evaluate alternative claim–evidence pathways and resolve inconsistencies with greater flexibility. This would advance both sufficiency and coherence detection, making the framework adaptable to more complex and uncertain verification scenarios.

5.5 Chapter Summary

This chapter consolidated the contributions, limitations, and future directions of the research. The proposed semantic framework significantly advanced automated fact-checking by introducing mechanisms for evidence sufficiency, logical coherence, and cross-domain consistency. By integrating contextual embeddings, knowledge graph representations, and human-in-the-loop validation, it achieved measurable improvements in precision, recall, and F1 scores across diverse datasets. Importantly, the research demonstrated that effective fact-checking requires more than evidence retrieval it demands multi-layered evaluation that preserves transparency, resolves contradictions, and adapts to domain-specific conditions. While challenges remain in handling partial evidence and scaling to real-time applications, the proposed solutions and future directions chart a clear path forward.

In conclusion, this research establishes a robust, explainable, and empirically validated framework for automated fact-checking. By balancing automation with expert validation, and by setting the stage for multimodal integration and advanced reasoning, it offers a solid foundation for the next generation of tools to combat misinformation in an increasingly complex digital landscape.

REFERENCES

- Alfattni, G., Peek, N., & Nenadic, G. (2021). Attention-based bidirectional long short-term memory networks for extracting temporal relationships from clinical discharge summaries. *Journal of Biomedical Informatics*, 123, 103915. <https://doi.org/10.1016/j.jbi.2021.103915>.
- 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>.
- Allein, L., Augenstein, I., & Moens, M. (2021). Time-aware evidence ranking for fact-checking. *Journal of Web Semantics*, 71, 100663.
- Allein, L., Saelens, M., Cartuyvels, R., & Moens, M. (2023). Implicit Temporal Reasoning for Evidence-Based Fact-Checking. In *Findings of the Association for Computational Linguistics*, 176–189. <https://doi.org/10.18653/v1/2023.findings-eacl.13>.
- Aly, R., Guo, Z., Schlichtkrull, M., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., & Mittal, A. (2021). FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information [Dataset]. In *Zenodo (CERN European Organization for Nuclear Research)*. <https://doi.org/10.5281/zenodo.4911508>.
- Arase, Y., & Tsujii, J. (2019). Transfer learning for semantic equivalence in fact-checking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. *Artificial Intelligence*, 259, 52–90. <https://doi.org/10.1016/j.artint.2018.02.003>.

- Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). Generating Fact Checking Explanations. In *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. <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*, 10, 746–763. https://doi.org/10.1162/tacl_a_00486.
- Augenstein, I., Lioma, C., Wang, D., Lima, L. C., Hansen, C., Hansen, C., & Simonsen, J. Balakrishnan, V. (2022). COVID-19 and fake news dissemination among Malaysians – Motives and its sociodemographic correlates. *International Journal of Disaster Risk Reduction*, 73, 102900. <https://doi.org/10.1016/j.ijdr.2022.102900>.
- Barik, A. M., Hsu, W., & Lee, M. L. (2022). Incorporating external knowledge for evidence-based fact verification. *Companion Proceedings of the the Web Conference 2018*, 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). <https://doi.org/10.14569/ijacsa.2022.0130225>.
- Bouzeghoub, A., Jabbour, S., Ma, Y., & Raddaoui, B. (2022c). Handling conflicts in uncertain ontologies using deductive argumentation. *Proceedings of the International Conference on Web Intelligence*, 65–72. <https://doi.org/10.1145/3106426.3106454>.
- Braithwaite, R. S., Roberts, M. S., & Justice, A. C. (2007). Incorporating Quality of Evidence into Decision Analytic Modeling. *Annals of Internal Medicine*, 146(2),

- Carnielli, W., & Rodrigues, A. (2015). Towards a Philosophical Understanding Of The Logics Of Formal Inconsistency. *Manuscript*, 38(2). <https://doi.org/10.1590/0100-6045.2015.V38N2.WCeAR>.
- Ceron, W., De-Lima-Santos, M., & Quiles, M. G. (2020). 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>.
- Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., & Hwang, A. (2020). AMPERSAND: Argument mining for persuasive online discussions. *arXiv preprint arXiv:2004.14677*. <https://arxiv.org/abs/2004.14677>.
- Chandrasekaran, K., Kandasamy, A., Venkatesan, M., Prabhavathi, P., Gokuldhev, M., & Aishwarya, C. (2024, March). Ontology for contextual fake news assessment based on text and images. In *Proceedings of the 32nd Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP 2024)* (pp. 191–198). IEEE. <https://doi.org/10.1109/PDP62718.2024.00034>.
- Chen, J., Kim, G., Sriram, A., Durrett, G., & Choi, E. (2024). Complex Claim Verification with Evidence Retrieved in the Wild. In *Complex Claim Verification with Evidence Retrieved in the Wild*. <https://doi.org/10.18653/v1/2024.naacl-long.196>
- Chen, K., Shen, G., Huang, Z., & Wang, H. (2021). Improved entity linking for simple question answering over knowledge graph. *International Journal of Software Engineering and Knowledge Engineering*, 31(01), 55–80. <https://doi.org/10.1142/s0218194021400039>.
- Chen, X., Chen, M., Shi, W., Sun, Y., & Zaniolo, C. (2019). Embedding uncertain knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 3363–3370. <https://doi.org/10.1609/aaai.v33i01.33013363>.

- Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021). WhatTheWikiFact. In *In Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. <https://doi.org/10.1145/3459637.3481987>.
- Cheung, T., & Lam, K. (2023). FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking. *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 846–853. <https://doi.org/10.1109/apsipaasc58517.2023.10317251>.
- Choi, E. C., & Ferrara, E. (2023). Automated Claim Matching with Large Language Models: Empowering Fact-Checkers in the Fight Against Misinformation. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4614239>.
- Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational Fact Checking from Knowledge Networks. *PLoS ONE*, 10(6), e0128193. <https://doi.org/10.1371/journal.pone.0128193>.
- Clarke, C. L. A., & Terra, E. L. (2003). Passage retrieval vs. document retrieval for factoid question answering. *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, 427–428. <https://doi.org/10.1145/860435.860534>.
- 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>.
- D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*. <https://arxiv.org/abs/2005.14165>.
- Dai, H., Du, D., Li, X., & Song, Y. (2019). Improving Fine-grained Entity Typing with Entity Linking. *Improving Fine-grained Entity Typing With Entity Linking*, 6209–6214. <https://doi.org/10.18653/v1/d19-1643>.

- Das, A., Liu, H., Kovatchev, V., & Lease, M. (2022). The state of human-centered NLP technology for fact-checking. *Information Processing & Management*, 60(2), 103219. <https://doi.org/10.1016/j.ipm.2022.103219>.
- Das, S. D., Basak, A., & Dutta, S. (2021). A Heuristic-Driven Ensemble Framework for COVID-19 fake news Detection. In *Communications in computer and information science* (pp. 164–176). https://doi.org/10.1007/978-3-030-73696-5_16.
- Dernoncourt, F., & Lee, J. Y. (2017). PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. *International Joint Conference on Natural Language Processing*, 2, 308–313.
- Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. <https://arxiv.org/abs/1810.04805>.
- Dieussaert, K., Ford, M., & Horsten, L. (2004). Influencing nonmonotonic reasoning by modifier strength manipulation. *Cognitive Science*, 26(26). *Journal of Data and Information Quality*, 11(3), 1–27. <https://doi.org/10.1145/3297722>.
- Du, J., Qi, K., & Shen, Y. (2018). Knowledge Graph Embedding with Logical Consistency. In *Lecture notes in computer science* (pp. 123–135). https://doi.org/10.1007/978-3-030-01716-3_11.
- Du, J., Qi, K., Wan, H., Peng, B., Lu, S., & Shen, Y. (2017). Enhancing Knowledge Graph Embedding from a Logical Perspective. In *Lecture notes in computer science* (pp. 232–247). https://doi.org/10.1007/978-3-319-70682-5_15.
- Ennen, P., Freddi, F., Lin, C., Kung, P., Wang, R., Yang, C., Shiu, D., & Bernacchia, A. (2023). Hierarchical Representations in Dense Passage Retrieval for Question-Answering. In *Hierarchical Representations in Dense Passage Retrieval for Question-Answering* (pp. 17–28). <https://doi.org/10.18653/v1/2023.fever-1.2>.
- Fadili, H. (2013). Towards a new approach of an automatic and contextual detection of meaning in text, based on lexico-semantic relations and the concept of the

- context. In *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*. (pp. NC). IEEE <https://hal.science/hal-01126555>.
- Forstmeier, W., Wagenmakers, E., & Parker, T. H. (2016). Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews/Biological Reviews of the Cambridge Philosophical Society*, 92(4), 1941–1968. <https://doi.org/10.1111/brv.12315>.
- Fraga-Lamas, P., & Fernandez-Carames, T. M. (2020). Fake News, disinformation, and deepfakes: leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality. *IT Professional*, 22(2), 53–59. <https://doi.org/10.1109/mitp.2020.2977589>.
- Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J., Olin, J., Ross, M., & Szafran, J. (2021). Fake claims of fake news: Political misinformation, warnings, and the tainted truth effect. *Political Behavior*, 43(4), 1465–1487. <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>.
- G. (2019). MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. <https://doi.org/10.18653/v1/d19-1475>.
- 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>.

- Gao, J., Zhang, Y., & Chen, H. (2024). Dynamic ontology alignment for evolving knowledge graphs in fact verification. *Information Processing & Management*, 61(4), 103528.
- Gencheva, P., Nakov, P., Villodre, L.M., Barrón-Cedeño, A., & Koychev, I. (2017). A Context-Aware Approach for Detecting Worth-Checking Claims in Political Debates. *Recent Advances in Natural Language Processing*. https://doi.org/10.26615/978-954-452-049-6_037
- Ghosh, B., Hasan, S., Arafat, N. A., & Khan, A. (2025). Logical consistency of large language models in fact-checking. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*. <https://openreview.net/forum?id=SimlDuN0YT>.
- Giulianelli, M., Del Tredici, M., & Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3960–3973). <https://doi.org/10.18653/v1/2020.acl-main.365>
- Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2009). Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1), 3–
- Graph Comparative Reasoning System. In *In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 3308-3318). (pp. 3308–3318). <https://doi.org/10.1145/3447548.3467128>.
- Guo, S., Wang, Q., Wang, B., Wang, L., & Guo, L. (2016). SSE: Semantically smooth embedding for knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(4), 884–897. <https://doi.org/10.1109/tkde.2016.2638425>.

- Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10, 178–206. https://doi.org/10.1162/tacl_a_00454.
- Gupta, N., Singh, S., & Roth, D. (2017). Entity linking via joint encoding of types, descriptions, and context. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. <https://doi.org/10.18653/v1/d17-1284>.
- Han, Z., Liao, R., Gu, J., Zhang, Y., Ding, Z., Gu, Y., Koepl, H., Schütze, H., & Tresp, V. (2023). ECOLA: Enhancing Temporal Knowledge Embeddings with Contextualized Language Representations. *Findings of the Association for Computational Linguistics: ACL 2022*, 5433–5447. <https://doi.org/10.18653/v1/2023.findings-acl.335>.
- Hanselowski, A., Stab, C., Schulz, C., Li, Z., & Gurevych, I. (2019). A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking. In *arXiv*. <https://doi.org/10.18653/v1/k19-1046>.
- Hassan, N., Arslan, F., Li, C., & Tremayne, M. (2017, August). Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1803–1812). ACM. <https://doi.org/10.1145/3097983.3098131>.
- He, J., Liu, T., Zhao, J., & Turner, B. (2025). MM-FusionNet: Context-aware dynamic fusion for multi-modal fake news detection with large vision-language models. *arXiv preprint arXiv:2508.05684*. <https://arxiv.org/abs/2508.05684>.
- Heyvaert, P., De Meester, B., Dimou, A., & Verborgh, R. (2019). Rule-driven inconsistency resolution for knowledge graph generation rules. *Semantic Web*, 10(6), 1071–1086. <https://arxiv.org/abs/2508.05684>.

- Hogenboom, A., Hogenboom, F., Kaymak, U., Wouters, P., & De Jong, F. (2010). Mining economic sentiment using argumentation structures. In *Lecture notes in computer science* (pp. 200–209). https://doi.org/10.1007/978-3-642-16385-2_25.
- Hu, Z., Hu, J., Ding, W., & Zheng, X. (2015). Review Sentiment Analysis Based on Deep Learning. In *EEE 12th International Conference on e-Business Engineering*, (pp. 87–94). <https://doi.org/10.1109/icebe.2015.24>.
- Huang, Y., Zhao, F., Gui, X., & Jin, H. (2021). Path-enhanced explainable recommendation with knowledge graphs. *World Wide Web*, 24(5), 1769–1789. <https://doi.org/10.1007/s11280-021-00912-4>.
- Hussain, S., De Roo, J., Daniyal, A., & Abidi, S. S. R. (2011). Detecting and Resolving Inconsistencies in Ontologies Using Contradiction Derivations. *2011 IEEE 35th Annual Computer Software and Applications Conference*, 556–561. <https://doi.org/10.1109/compsac.2011.76>.
- Huynh, V., & Papotti, P. (2018, April). Towards a benchmark for fact checking with knowledge bases. In *Companion Proceedings of the The Web Conference 2018 (WWW '18 Companion)* (pp. 1657–1664). Lyon, France: ACM. <https://doi.org/10.1145/3184558.3191616>.
- Islam, M. R., & Zibran, M. F. (2018). SentiStrength-SE: Exploiting domain specificity for improved sentiment analysis in software engineering text. *Journal of Systems and Software*, 145, 125–146. <https://doi.org/10.1016/j.jss.2018.08.030>.
- Jiang, S., & Wilson, C. (2018). Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), Article 82, 1–23. <https://doi.org/10.1145/3274351>.

- Jiang, S., Baumgartner, S., Ittycheriah, A., & Yu, C. (2020). Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles. In *Proceedings of The Web Conference 2020*. <https://doi.org/10.1145/3366423.3380231>.
- Jiang, Y., Bordia, S., Zhong, Z., Dognin, C., Singh, M., & Bansal, M. (2020). HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification. *HoVer: A Dataset for Many-Hop Fact Extraction and Claim Verification*. <https://doi.org/10.18653/v1/2020.findings-emnlp.309>.
- Jimenez, D. (2017). Towards Building an Automated Fact-Checking System. In *SIGMOD '17: Proceedings of the 2017 ACM International Conference on Management of Data* (Vol. 2015, pp. 7–9). <https://doi.org/10.1145/3055167.3055176>.
- Kassos, C., & Delugach, H. (2017). Specifying constraints for detecting inconsistencies in a conceptual graph knowledge base. *International Journal of Conceptual Structures and Smart Applications*, 5(2), 34–64. <https://doi.org/10.4018/ijcssa.2017070103>.
- Kim, J., Park, S., Kwon, Y., Jo, Y., Thorne, J., & Choi, E. (2023). FactKG: Fact Verification via Reasoning on Knowledge Graphs. In *arXiv*. <https://doi.org/10.18653/v1/2023.acl-long.895>.
- Kirilenko, A. P., Stepchenkova, S. O., Kim, H., & Li, X. (2017). Automated Sentiment Analysis in Tourism: Comparison of approaches. *Journal of Travel Research*, 57(8), 1012–1025. <https://doi.org/10.1177/0047287517729757>.
- Kise, K., Junker, M., Dengel, A., & Matsumoto, K. (2001). Passage-Based Document Retrieval as a Tool for Text Mining with User's Information Needs. In *Lecture notes in computer science* (pp. 155–169). https://doi.org/10.1007/3-540-45650-3_16.
- Kotonya, N., & Toni, F. (2020). Explainable Automated Fact-Checking: a survey. *Proceedings of the 17th International Conference on Computational Linguistics -*. <https://doi.org/10.18653/v1/2020.coling-main.474>.

- Kotonya, N., & Toni, F. (2024). Towards a Framework for Evaluating Explanations in Automated Fact Verification. *arXiv preprint arXiv:2403.20322*. <https://arxiv.org/abs/2403.20322>.
- Kyle, K. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757–786. <https://doi.org/10.1002/tesq.194>.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). 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, K., et al. (2018). Fact-checking by decomposable attention with lexical tagging. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Lee, N., Bang, Y., Madotto, A., & Fung, P. (2021, June). Towards few-shot fact-checking via perplexity. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1971–1981). <https://doi.org/10.18653/v1/2021.naacl-main.158>.
- Li, M., Peng, B., Galley, M., Gao, J., & Zhang, Z. (2024). Self-Checker: Plug-and-Play Modules for Fact-Checking with Large Language Models. *Findings of the Association for Computational Linguistics: NAACL 2022*.
- Li, Q., Li, F., Li, S., Li, X., Liu, K., Liu, Q., & Dong, P. (2022). Improving entity linking by introducing knowledge graph structure information. *Applied Sciences*, 12(5), 2702. <https://doi.org/10.3390/app12052702>.
- Li, Y., et al. (2023). Graph attention for argument mining in automated fact-checking. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

- Li, Z., Liu, H., Zhang, Z., Liu, T., & Xiong, N. N. (2021a). Learning knowledge graph embedding with heterogeneous relation attention networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8), 3961–3973. <https://doi.org/10.1109/tnnls.2021.3055147>.
- Li, Z., Liu, H., Zhang, Z., Liu, T., & Xiong, N. N. (2021b). Learning knowledge graph embedding with heterogeneous relation attention networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8), 3961–3973. <https://doi.org/10.1109/tnnls.2021.3055147>.
- Lin, Y., Huang, P., Lai, Y., Feng, Y., & Zhao, D. (2019). Evidence distilling for fact extraction and verification. In *Lecture notes in computer science* (pp. 211–222). https://doi.org/10.1007/978-3-030-32233-5_17.
- Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1). <https://doi.org/10.1609/aaai.v29i1.9491>.
- Liu, L., Du, B., Ji, H., & Tong, H. (2020). KompaRe: A knowledge graph comparative reasoning system. *arXiv preprint arXiv:2011.03189*. <https://arxiv.org/abs/2011.03189>.
- Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.
- Liu, Y., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ma, J., Lu, J., & Zhang, G. (2009). Information inconsistencies detection using a rule-map technique. *Expert Systems With Applications*, 36(10), 12510–12519. <https://doi.org/10.1016/j.eswa.2009.04.040>.

- Ma, J., Zhou, C., Chen, Y., Wang, Y., Hu, G., & Qiao, Y. (2023). TECRE: A novel temporal conflict resolution method based on temporal Knowledge graph embedding. *Information*, 14(3), 155. <https://doi.org/10.3390/info14030155>.
- Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L., & Lao, N. (2019, September). Contextual graph attention for answering logical queries over incomplete knowledge graphs. *Manuscript*, 38(2), 155–184. <https://doi.org/10.1590/0100-6045.2015.v38n2.wcear>.
- Marcot, B. G. (2012). Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecological Modelling*, 230, 50–62. <https://doi.org/10.1016/j.ecolmodel.2012.01.013>.
- 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>.
- Mayer, T., Cabrio, E., & Villata, S. (2020). Transformer-Based Argument Mining for Healthcare Applications. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*. IOS Press. <https://api.semanticscholar.org/CorpusID:221713735>
- Miranda, S., Nogueira, D., Mendes, A., Vlachos, A., Secker, A., Garrett, R., Mitchel, J., & Marinho, Z. (2019). Automated Fact Checking in the News Room. In *In The world wide web conference*. <https://doi.org/10.1145/3308558.3314135>.
- Mues, C., & Vanthienen, J. (2004). Efficient rule base verification using binary decision diagrams. In *Lecture notes in computer science* (pp. 445–454). https://doi.org/10.1007/978-3-540-30075-5_43.

- 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>.
- Nguyen, H., & Litman, D. (2015). Extracting Argument and Domain Words for Identifying Argument Components in Texts. *Proceedings of the 2nd Workshop on Argumentation Mining*, 22–28. <https://doi.org/10.3115/v1/w15-0503>.
- Nikopensius, G., Mayank, M., Phukan, O. C., & Sharma, R. (2023). Reinforcement Learning-based Knowledge Graph Reasoning for Explainable Fact-checking. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining* (pp. 164–170). <https://doi.org/10.1145/3625007.3627593>.
- Oh, B., Seo, S., Hwang, J., Lee, D., & Lee, K. (2021). 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. *arXiv*, 3892–3898. <https://doi.org/10.24963/ijcai.2021/536>.
- Otero-Cerdeira, L., Rodríguez-Martínez, F. J., & Gómez-Rodríguez, A. (2015). Ontology matching: A literature review. *Expert Systems with Applications*, 42(2), 949–971.
- Padia, A., Kalpakis, K., Ferraro, F., & Finin, T. (2019). Knowledge graph fact prediction via knowledge-enriched tensor factorization. *Journal of Web Semantics*, 59, 100497. <https://doi.org/10.1016/j.websem.2019.01.004>.
- Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2018). CredEye. In *Proceedings of the The Web Conference*. (pp. 155–158). <https://doi.org/10.1145/3184558.3186967>.
- Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. <https://doi.org/10.18653/v1/d18-1003>.

- Pradhan, A. (2018). *Explainable Fact Checking by Combining Automated Rule Discovery with Probabilistic Answer Set Programming*. <https://repository.asu.edu/items/50443>.
- Provoost, S., Ruwaard, J., Van Breda, W., Riper, H., & Bosse, T. (2019). Validating Automated Sentiment Analysis of Online Cognitive Behavioral Therapy Patient texts: an Exploratory study. *Frontiers in Psychology*, 10. <https://doi.org/10.3389/fpsyg.2019.01065>.
- Qiao, Z., Ning, Z., Du, Y., & Zhou, Y. (2021). Context-Enhanced entity and relation embedding for knowledge graph completion (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18), 15871–15872. <https://doi.org/10.1609/aaai.v35i18.17932>.
- Qiu, Y., Wang, Y., Jin, X., & Zhang, K. (2020). Stepwise Reasoning for Multi-Relation Question Answering over Knowledge Graph with Weak Supervision. *Proceedings of the 13th International Conference on Web Search and Data Mining*.
- Qureshi, M. A., Asif, M., Hassan, M. F., Mustafa, G., Ehsan, M. K., Ali, A., & Sajid, U. (2021). A Novel Auto-Annotation Technique for Aspect Level Sentiment Analysis. *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, 70(3), 4987–5004. <https://doi.org/10.32604/cmc.2022.020544>.
- 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 2021 Conference on Empirical Methods in Natural Language Processing*. <https://doi.org/10.18653/v1/d17-1317>.
- Reddy, S., et al. (2019). Embedding-based semantic similarity for misinformation detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

- Rosin, G. D., & Radinsky, K. (2022). Temporal attention for language models. *Findings of the Association for Computational Linguistics: NAACL 2022*. <https://doi.org/10.18653/v1/2022>.
- Rossi, A., Barbosa, D., Firmani, D., Matinata, A., & Merialdo, P. (2021). Knowledge graph embedding for link prediction. *ACM Transactions on Knowledge Discovery From Data*, 15(2), 1–49. <https://doi.org/10.1145/3424672>.
- Rosso, P., Yang, D., & Cudré-Mauroux, P. (2020). Beyond Triplets: Hyper-Relational Knowledge Graph Embedding for Link Prediction. *WWW '20: Proceedings of the Web Conference 2020*. <https://doi.org/10.1145/3366423.3380257>.
- Rúas-Araújo, J., Rodríguez-Martelo, T., & Fontenla-Pedreira, J. (2023). Disinformation and Verification in a Digital Society: An analysis of strategies and policies applied in the European Regional TV broadcasters of the CIRCUM Network. *Societies*, 13(4), 81. <https://doi.org/10.3390/soc13040081>.
- Ruiz-Dolz, R., Alemany, J., Barbera, S. M. H., & Garcia-Fornes, A. (2021). Transformer-Based Models for Automatic Identification of Argument Relations: A Cross-Domain Evaluation. *IEEE Intelligent Systems*, 36(6), 62–70. <https://doi.org/10.1109/mis.2021.3073993>.
- 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>.
- Sarr, E. N., Sall, O., Maiga, A., Faty, L., & Marone, R. M. (2018). Automatic Segmentation and tagging of facts in French for automated fact-checking. *2021 IEEE International Conference on Big Data (Big Data)*, 43, 5439–5441. <https://doi.org/10.1109/big>.

- Sarrouti, M., Abacha, A. B., 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. *Language Resources and Evaluation*, 6874–6882.
- Schlichtkrull, M., Ousidhoum, N., & Vlachos, A. (2023). The Intended Uses of Automated Fact-Checking Artefacts: Why, How and Who. In *arxiv* (pp. 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>.
- Shankar, A., Kulathuramaiyer, N., Abdullah, J. B., & Pakkirisamy, M. (2025). Mitigating the evidence-related factors in automated fact-checking. *Journal of Computer Science*, 21(3), 646–664. <https://doi.org/10.3844/jcssp.2025.646.664>.
- Sharma, K., Kumar, P., & Li, Y. (2024). OG-RAG: Ontology-guided retrieval-augmented generation for large language models. *arXiv preprint arXiv:2412.15235*. <https://arxiv.org/abs/2412.15235>.
- Shi, B., & Weninger, T. (2016). Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems*, 104, 123–133. <https://doi.org/10.1016/j.knosys.2016.04.015>.
- Shiralkar, P., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2017). Finding streams in knowledge graphs to support fact checking. *2021 IEEE International Conference on Data Mining (ICDM)*, 859–864. <https://doi.org/10.1109/icdm.2017.105>.

- Sun, H., Tang, J., & Zhang, J. (2012). Ontology-based medical information retrieval for supporting evidence-based medicine. *Artificial Intelligence in Medicine*, 55(1), 1–18.
- Sun, Z., Yang, J., Zhang, J., Bozzon, A., Huang, L., & Xu, C. (2018). Recurrent knowledge graph embedding for effective recommendation. In *RecSys '18: Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 297–305). <https://doi.org/10.1145/3240323.3240361>.
- Thakur, A., et al. (2023). Lexical–semantic embeddings for detecting political misinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Thorne, J., & Vlachos, A. (2021). Evidence-based Factual Error Correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 3298–3309). <https://doi.org/10.18653/v1/2021.acl-long.256>.
- Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *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>.
- Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 809–819).
- Thurmair, G., et al. (2012). Lexical resources and their role in semantic processing. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

- Tsai, C. (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. In *WSDM '21: Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (pp. 526–534). <https://doi.org/10.1145/3437963.3441828>.
- Wadden, D., et al. (2020). Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2019). 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>.
- Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge Graph Embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12), 2724–2743. <https://doi.org/10.1109/tkde.2017.2754499>
- Wang, Y., Xia, C., Si, C., Yao, B., & Wang, T. (2020). Robust reasoning over heterogeneous textual information for fact verification. *IEEE Access*, 8, 157140–157150. <https://doi.org/10.1109/access.2020.3019586>.
- Wang, Y., Xia, C., Si, C., Zhang, C., & Wang, T. (2020). The graph reasoning approach based on the dynamic knowledge auxiliary for complex fact verification. *Electronics*, 9(9), 1472. <https://doi.org/10.3390/electronics9091472>.
- Weber, L., Minervini, P., Münchmeyer, J., Leser, U., & Rocktäschel, T. (2019). NLProlog: Reasoning with Weak Unification for Question Answering in Natural Language. In *arxiv*. <https://doi.org/10.18653/v1/p19-1618>.

- Wein, C., et al. (2022). Contrastive learning for paraphrase and contradiction detection in fact verification. In *Proceedings of EMNLP 2022*.
- Wu, P., & Su, S. Y. W. (1993). Rule validation based on logical deduction. In *International Conference on Information and Knowledge Management (CIKM '93)* (pp. 164–173). New York, NY: Association for Computing Machinery. <https://doi.org/10.1145/170088.170123>.
- Xie, J., et al. (2024). Hybrid discourse–semantic frameworks for implicit premise detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Yamada, I., Asai, A., & Hajishirzi, H. (2021). Efficient Passage Retrieval with Hashing for Open-domain Question Answering. In *arXiv*. <https://doi.org/10.18653/v1/2021.acl-short.123>.
- Yang, G., Wen, D., Kinshuk, N., Chen, N., & Sutinen, E. (2014). A novel contextual topic model for multi-document summarization. *Expert Systems with Applications*, 42(3), 1340–1352. <https://doi.org/10.1016/j.eswa.2014.09.015>.
- 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>.
- Yao, B. M., Shah, A., Sun, L., Cho, J., & Huang, L. (2023). End-to-End multimodal Fact-Checking and explanation generation: a challenging dataset and models. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2733–2743. <https://doi.org/10.1145/3539618.3591879>.
- Zhang, J., & El-Gohary, N. M. (2016). 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., Cai, J., Zhang, Y., & Wang, J. (2020). Learning Hierarchy-Aware Knowledge graph embeddings for link prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03), 3065–3072. <https://doi.org/10.1609/aaai.v34i03.5701>.
- 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>.

APPENDICES

Appendix A: Journal Publications

1. **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.
2. **Shankar, A.,** Kulathuramaiyer, N., Abdullah, J. B. & Pakkirisamy, M. (2025). Mitigating the Evidence-Related Factors in Automated Fact-Checking. *Journal of Computer Science*, 21(3), 646-664.