



UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

**Ph.D. DEGREE IN
ELECTRONIC AND COMPUTER ENGINEERING**

Cycle XXXVII

TITLE OF THE Ph.D. THESIS

Building Human Digital Twins through Natural Language Processing: From

Conversation to Knowledge

Scientific Disciplinary Sector(s)

ING-INF/03

Ph.D. Student: Luigi Serreli

Supervisor Prof. Michele Nitti

Final exam. Academic Year 2023/2024

Thesis defence session: March 2025

Foreword

First, I would like to express my deepest gratitude to my thesis supervisor, Prof. Michele Nitti for his guidance, unwavering support, and insightful feedback throughout this research journey. His expertise and encouragement have been instrumental in shaping this thesis and my growth as a researcher.

I extend my heartfelt thanks to my parents, whose unconditional love and support have been my foundation throughout my academic journey. They have always believed in me, encouraged my pursuits, and provided the stability I needed to focus on my studies. Their sacrifices and dedication to my education have made all of this possible.

To my girlfriend, Gulnaziye Bingol, who has been my constant source of motivation and emotional support, thank you for your patience and understanding and for bringing joy and balance to my life during the challenges of life. Your presence has made this journey infinitely more meaningful.

Finally, I would like to thank all my friends who have supported me throughout this journey, providing much-needed encouragement and moments of levity when the path seemed challenging.

Contents

List of Figures	5
List of Tables	7
Abstract	7
1 Introduction	9
1.1 Research Challenges	10
1.2 Thesis Objectives and Contributions	10
1.3 List of Publications	12
1.4 Thesis Structure	12
2 Literature Review and Background	14
2.1 Historical Background - From Digital Twin to Human Digital Twin	14
2.1.1 Paradigm Shifts in Digital Twin Technology	15
2.1.2 Human Centric Interaction in Digital Twin Evolution	17
2.1.3 Technological Drivers	18
2.2 Applications and Emerging Challenges in Human Digital Twins	20
2.2.1 Privacy and Ethical Considerations	21
2.2.2 Technical Challenges	23
2.3 Knowledge Representation in Human Digital Twins	24
2.3.1 Knowledge Graph Architectures for HDT	25
2.3.2 Temporal Aspects and Evolution	27
2.3.3 Knowledge Integration and Embeddings	28
2.3.4 From Knowledge Representation to Content Analysis	29
2.4 Topic Classification for Knowledge Representation	31
2.4.1 Text Classification Approaches and Methods	32
3 Human Digital Twin Model	35
3.1 Reference Scenario and System Overview	36
3.2 Enabling Paradigms	38

3.2.1	From Digital Twin to Human Digital Twin: An Enhanced Perspective	38
3.2.2	Evolution Toward Proactive Assistance	39
3.3	The Envisioned IoT-PDA Framework	40
3.4	Digital Brain Architecture	42
3.4.1	Acquiring User Knowledge	42
3.4.2	Cumulative Knowledge Processing	43
3.4.3	Insights Utility Evaluation	44
3.5	Temporal Dynamics and Adaptive Learning	45
3.6	Preliminary Evaluation	46
3.7	Privacy and Security Considerations	50
4	Topic Modelling	53
4.1	Aims and Scope of the Research	54
4.2	Proposed Solution	56
4.2.1	Problem Definition	56
4.2.2	Key Sentence Extractor	57
4.2.3	Word Embedding	59
4.2.4	Topic Classifier	60
4.3	Experimental Setup	63
4.3.1	Global News 60K: A Novel Dataset for Topic Classification	63
4.3.2	Dataset Analysis	65
4.3.3	Simulation Setup	66
4.4	Experimental Evaluation	68
4.4.1	Evaluation Metrics	69
4.4.2	Classification Results	70
4.4.3	Performance Analysis of the Key Sentence Extractor	72
4.4.4	Comparative Analysis with State-of-the-Art Methods	74
4.4.5	Comparative Analysis with Large Language Models	77
4.4.6	Conclusions and Future Directions	79
5	From LLM Reliability to Fake News Detection: A Case Study	81
5.1	Reliability in Large Language Model	81
5.2	From Hallucinations to Disinformation: A Framework for Analysis	82
5.3	State of the Art on Disinformation and Trustworthiness	83
5.3.1	Trustworthiness on News Providers	86
5.3.2	State of the Art in Fallacy Classification	87
5.4	A Framework for Fake News Detection and Trust Management	88
5.4.1	Feedback Evaluation Model	89
5.4.2	Context-based Analysis	95
5.4.3	Trust Model Framework	98

5.4.4	Experimental Evaluation	105
5.4.5	Performance Analysis of the Feedback Model	107
5.4.6	Trust Model Performance Analysis	113
5.4.7	System Performance Analysis in Real-World Conditions	119
5.4.8	Alternative Approaches to Data Management: A Blockchain Perspective	121
5.4.9	Conclusions and Future Directions	123
5.5	A Framework for Fallacy Detection	124
5.5.1	Challenges in Dataset Availability	125
5.5.2	Implementation Framework and Methodology	126
5.5.3	Dataset Construction and Model Adaptation	126
5.5.4	Dataset Augmentation Strategy	127
5.5.5	Model Selection and Fine-tuning Approach	130
5.5.6	Inference Phase	131
5.5.7	Performance Analysis and Results	132
6	Conclusions and Future Work	135
6.1	Key Research Outcomes	136
6.1.1	Advancements in Topic Classification	136
6.1.2	Human Digital Twin Framework Evolution	136
6.1.3	Information Reliability Framework	137
6.2	Future Research Directions	137
6.2.1	Technical Enhancement Opportunities	137
6.2.2	Knowledge Representation and Processing	138
6.2.3	Information Reliability and Trust	138
6.2.4	Application Domains and Scaling	139
6.2.5	Theoretical Foundations	139
	Appendices	141
	A Acronyms	143

List of Figures

3.1	High level reference scenario.	40
3.2	Digital Brain structure.	42
3.3	Initial knowledge graph constructed from the first analysed scene, showing detected entities (nodes) and their relationships (edges) with confidence scores.	48
3.4	Evolved knowledge graph showing the enriched understanding of Sheldon's character after processing additional scenes.	48
3.5	Privacy Protection Stages	50
4.1	News Topic Classification System Overview	54
4.2	System Architecture Flow	57
4.3	Proposed architecture with BiLSTM, Attention Mechanism, and a fully connected layer.	63
4.4	Confusion Matrix for AGNews dataset classification. The classes are mapped as follows: 1 - World, 2 - Sports, 3 - Business, 4 - Sci/Tech. The diagonal elements represent correct classifications, while off-diagonal elements indicate misclassifications between categories.	70
4.5	Sentence percentage distribution	75
4.6	Accuracy and Efficiency Trade-off	75
5.1	Overall system architecture.	89
5.2	Sequence diagrams for the fake news detector and the trust management model.	90
5.3	Analysis of the number of supported claims and news length for the Kaggle dataset.	94
5.4	ROC curves for the proposed machine learning classifiers.	109
5.5	XGBoost model - Test Confusion Matrix	109
5.6	FEVER Score distribution	111
5.7	Fact-based overall accuracy over different thresholds of feedback and similarity	111
5.8	Impact of advertisements (ads)	113
5.9	Final trust score: Three different behaviours.	116

5.10 Trust score with and without the relevance algorithm.	116
5.11 Trust model performance through different fake news detector accuracy systems.	118
5.12 Fallacy Extraction and Detection	125
5.13 System Overview	126
5.14 Fallacy Extraction and Detection	129

List of Tables

2.1	Technological Drivers of Human Digital Twin Development	20
2.2	Technical Challenges and Solutions in HDT Systems	24
2.3	Knowledge Graph Capabilities and Limitations in Human Digital Twin (HDT) Systems	30
3.1	Technical Overview of Digital Brain Components	45
4.1	Main Parameters of the Proposed Framework	58
4.2	List of sources employed to construct the dataset	65
4.3	Comparative Analysis of Dataset Characteristics	66
4.4	Hardware and Software Configuration for Experimental Evaluation	68
4.5	Classification Results	72
4.6	Works comparison on BBC News and AGNews	77
4.7	Performance Comparison between Our Framework and Large Language Models	79
5.1	Analysis of existing fake news detectors.	85
5.2	Message-based factor features.	92
5.3	Simulation Parameters and Their Optimal Values	106
5.4	Investigation for the sentiment analysis.	113
5.5	Trust and average feedback scores related to the most evident sources	117
5.6	Machine Specifications	119
5.7	Average impact of the plugin over 100 users.	121
5.8	Database and Blockchain comparison.	121
5.9	Categories and Descriptions of Logical Fallacies in News Content	128
5.10	Statistical Analysis of Augmented Dataset	129
5.11	Fine-tuning Hyperparameters	131
5.12	Performance comparison of base models	133
5.13	Performance improvement through fine-tuning	134

Abstract

Human Digital Twins (HDTs) are a revolutionary advancement in human-computer interaction, offering personalised digital representations of individual knowledge, behaviour, and preferences. These sophisticated systems face fundamental challenges in ensuring the reliable organisation and processing of extensive conversational data. This thesis presents an integrated framework that addresses these challenges through advanced topic modelling, natural language processing, and information verification techniques.

A key objective is to enhance HDTs' efficiency in capturing and organizing human knowledge from diverse conversational sources. This involves developing novel methods for reducing the dimensionality of conversational data while preserving essential semantic relationships. By improving the scalability of knowledge organisation, HDTs can effectively process large-scale text data without compromising information quality.

Recognising the potential for errors in information retrieval, similar to the issue of "hallucinations" in Large Language Models, this thesis also explores robust verification mechanisms. Drawing parallels with detecting misinformation in human-generated content, such as fake news, the research investigates underlying patterns of inaccurate information dissemination. This analysis leads to the development of innovative approaches for ensuring response reliability in HDTs.

The experimental validation of the framework encompasses various scales of text analysis, from short-form content like social media posts to long-form articles and technical papers. The results demonstrate the effectiveness of the integrated framework in organising and verifying information while maintaining semantic coherence across different textual contexts. This research contributes to the broader understanding of how HDTs can evolve into reliable partners in human-AI interaction, capable of handling complex information processing tasks while maintaining high standards of accuracy and trustworthiness. The methodologies and insights presented open new avenues for research into proactive HDT systems that can effectively serve as intelligent assistants while preserving the critical balance between capability and reliability.

Chapter 1

Introduction

The rapid evolution of digital technologies has fundamentally transformed our approach to human-machine interaction, marking a significant shift from traditional computing paradigms toward more sophisticated, human-centric systems. This transformation has been particularly evident in the development of Digital Twin (DT) technology, which has progressed from its origins in manufacturing and product lifecycle management to become a cornerstone of modern digital transformation strategies. Initially conceived as digital representations of physical assets, DTs have evolved to encompass increasingly complex systems, leading to the emergence of HDT, digital representations that capture not only physical characteristics but also human knowledge, behaviours, and interactions. This evolution from DTs to HDTs represents more than a simple technological advancement; it marks a fundamental shift in how we conceptualise the relationship between humans and digital systems. Traditional DTs, focused primarily on physical systems and operational data, provided valuable insights into system performance and maintenance. However, the extension of these concepts to human subjects introduces unprecedented complexity, requiring systems to process and understand not just structured data but also the nuanced, context-dependent nature of human communication and behaviour. The intersection of HDT technology with recent advances in artificial intelligence, particularly Large Language Models (LLMs), has created new possibilities for sophisticated human-machine interaction. These models offer unprecedented capabilities in natural language understanding and generation, potentially enabling HDTs to engage in more natural and meaningful interactions with users. However, this integration also introduces significant challenges regarding information reliability and trustworthiness, particularly given the tendency of LLMs to produce hallucinations - plausible but factually incorrect information.

1.1 Research Challenges

The primary challenge in implementing effective HDT systems lies in the processing and organising of complex conversational knowledge. Current Natural Language Processing (NLP) approaches often fail to capture the full semantic richness of human communication, mainly when dealing with domain-specific knowledge and context-dependent meanings. This limitation becomes critical in HDT systems that must maintain a comprehensive understanding of user interests, preferences, and expertise across multiple domains and interaction contexts. Traditional approaches to text processing and topic modelling often need to be revised for HDT applications due to their computational intensity and lack of scalability. These methods typically require processing entire documents to extract meaningful information, making them impractical for real-time applications where rapid response and efficient resource utilisation are crucial. The need for more efficient processing mechanisms to maintain accuracy while reducing computational overhead represents a significant technical challenge. This efficiency requirement becomes particularly critical when considering information verification tasks, where multiple sources must be processed and cross-referenced in real-time to establish content reliability. As HDT systems scale up their processing of information from various sources, the challenge of maintaining both computational efficiency and information trustworthiness becomes increasingly interlinked.

The problem of information reliability in HDT systems extends beyond simple accuracy to encompass issues of trust and credibility. The parallel between AI hallucinations and human-generated misinformation suggests deeper patterns in how false information maintains apparent plausibility while deviating from truth. Understanding and addressing these patterns requires new frameworks for analysing and verifying information, particularly in contexts where multiple sources of information must be integrated and evaluated. The development of practical solutions to these challenges requires not only theoretical advances but also innovative architectural approaches that can balance competing requirements effectively. Current systems often address individual aspects of these challenges in isolation, leading to fragmented solutions that fail to provide comprehensive frameworks for HDT implementation. This fragmentation creates a crucial gap between theoretical capabilities and practical implementation requirements.

1.2 Thesis Objectives and Contributions

Building upon the identified challenges, this thesis aims to develop an integrated framework for HDT that effectively addresses the organisation of conversational knowledge, ensures efficient information processing, and maintains response reliability. These objectives are pursued through different approaches that contribute to both theoretical understanding and practical implementation of HDT systems. The primary objective of this research is to

create a comprehensive framework that bridges the gap between theoretical capabilities and practical implementation requirements in HDT systems. This framework must balance the competing demands of processing efficiency, semantic accuracy, and information reliability while maintaining practical feasibility for real-world deployment. By carefully integrating topic modelling, natural language processing, and information verification techniques, we seek to establish new paradigms for human-AI interaction that can serve as foundations for future HDT development. The key contributions of this thesis can be summarised as follows:

1. **Novel Topic Modelling Framework:** We introduce an innovative approach to topic classification that significantly reduces computational overhead while maintaining high accuracy. The framework implements a sophisticated key sentence extractor algorithm that intelligently identifies and preserves semantically significant content, coupled with an attention-based neural architecture optimised for processing selected content. Through comprehensive empirical validation across multiple datasets, we demonstrate superior performance compared to both traditional classification approaches and modern large language models.
2. **Scalable Knowledge Organization:** We develop new methodologies for organising and processing conversational knowledge across different scales. This contribution encompasses advanced techniques for reducing the dimensionality of conversational data while preserving essential semantic relationships alongside efficient mechanisms for processing large-scale text data without compromising information quality. We also introduce novel approaches to maintaining semantic coherence across different textual contexts.
3. **Information Reliability Framework:** We present an innovative approach to ensuring response reliability in HDT systems by developing sophisticated verification mechanisms that draw parallels between Artificial Intelligence (AI) hallucinations and human-generated misinformation. This framework incorporates the implementation of robust trust evaluation metrics for assessing information sources, complemented by the creation of adaptive feedback mechanisms that enable continuous refinement of system responses.
4. **Novel Dataset Development:** We contribute to the research community through the creation of the Global News 60K dataset, comprising 60,000 carefully curated news articles. This contribution includes the implementation of comprehensive annotation schemes for topic classification.

These contributions collectively advance the state-of-the-art in HDT technology while providing practical solutions to key challenges in human-AI interaction. The

framework's effectiveness is demonstrated through extensive experimental validation, including both controlled laboratory testing and real-world application scenarios.

1.3 List of Publications

The following publications are related to the research presented in this thesis:

1. L. Serreli, C. Marche and M. Nitti, "Reducing Data Volume in News Topic Classification: Deep Learning Framework and Dataset," in IEEE Open Journal of the Computer Society, doi: 10.1109/OJCS.2024.3519747 [1].
2. C. Marche, I. Cabiddu, C. G. Castangia, L. Serreli and M. Nitti, "Implementation of a Multi-Approach Fake News Detector and of a Trust Management Model for News Sources," in IEEE Transactions on Services Computing, vol. 16, no. 6, pp. 4288-4301, Nov.-Dec. 2023, doi: 10.1109/TSC.2023.3311629 [2].
3. C. Marche, I. Cabiddu, C. G. Castangia, L. Serreli and M. Nitti, "Fake News Detection based on Blockchain Technology," 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Kyoto, Japan, 2022, pp. 654-659, doi: 10.1109/PIMRC54779.2022.9978074 [3].

1.4 Thesis Structure

This thesis is organised into Six chapters, each addressing specific aspects of Human Digital Twin development and implementation. The structure progressively builds from fundamental concepts through specific technical contributions to broader implications for human-AI interaction.

Chapter 2 establishes the theoretical foundation and background necessary for understanding Human Digital Twins. Beginning with a historical perspective on Digital Twin technology, the chapter traces the evolution from traditional DTs to modern HDTs, examining crucial paradigm shifts and technological drivers. Particular attention is paid to emerging challenges in HDT implementation, including privacy considerations and technical complexities in system development. The chapter concludes with an examination of knowledge representation approaches, establishing the groundwork for subsequent technical discussions.

Chapter 3 introduces our comprehensive HDT model, building upon the foundations established in previous chapters. The chapter presents a detailed examination of the Digital Brain architecture and our innovative approach to processing and organising user knowledge. Particular emphasis is placed on the system's temporal dynamics and adaptive learning

capabilities, demonstrating how the framework maintains an evolving understanding of user interests and preferences. The discussion includes preliminary evaluation results that validate the framework’s effectiveness in real-world scenarios.

Chapter 4 presents our framework for topic modelling and classification. This chapter addresses the fundamental challenge of organising and processing conversational data efficiently. The discussion begins with a formal problem definition and proceeds through the development of our Key Sentence Extractor algorithm, highlighting its role in reducing computational overhead while maintaining semantic accuracy. The chapter presents comprehensive experimental validation using multiple datasets, including our novel Global News 60K collection, demonstrating the framework’s effectiveness across different textual content types.

Chapter 5 examines the critical challenge of information reliability in HDT systems through an innovative case study of fake news detection. This chapter establishes significant parallels between Large Language Model (LLM) hallucinations and human-generated misinformation, developing new frameworks for ensuring information reliability. The discussion extends beyond simple detection to encompass broader questions of trust and credibility in human-AI interaction. The chapter presents extensive experimental results demonstrating the effectiveness of our proposed approaches in maintaining information reliability across different contexts.

Chapter 6 concludes the thesis by summarising the key contributions and discussing their implications for the broader field of HDT development. The chapter revisits the core themes explored throughout the thesis, synthesizing the technical advancements and theoretical insights into a cohesive narrative. The discussion extends to future research directions, outlining potential avenues for further exploration and development within the field. This concluding chapter provides a comprehensive overview of the thesis’s impact on the current state of HDT research while offering valuable perspectives on future challenges and opportunities.

Chapter 2

Literature Review and Background

The digital transformation of industrial and technological systems has fundamentally altered how we approach human-machine interaction and knowledge representation. This chapter examines the theoretical foundations and technological developments that have shaped the evolution from traditional Digital Twins to more sophisticated Human Digital Twin systems. Through a systematic analysis of existing literature, we explore how advances in artificial intelligence, changes in industrial paradigms, and growing recognition of the need for human-centric approaches in digital systems have driven this transformation. The examination begins with a historical perspective on DT evolution, tracing the development from early manufacturing applications to contemporary human-centric implementations. We then analyse the paradigm shifts that have characterised this evolution, mainly focusing on emerging new interaction modalities and integrating human factors in system design. The discussion extends to technical challenges and privacy considerations when implementing these systems, providing a comprehensive view of opportunities and constraints in HDT development. The latter sections of this chapter focus on knowledge representation approaches and topic classification methods, examining how these technical foundations enable the sophisticated modelling of human expertise and behaviour patterns in digital environments. This progression from historical context through technical implementation provides essential background for understanding the novel contributions presented in subsequent chapters. Through this structured examination of existing research, we establish the theoretical framework necessary for advancing HDT capabilities while highlighting crucial gaps in current approaches that our work aims to address.

2.1 Historical Background - From Digital Twin to Human Digital Twin

The evolution of DT technology represents a significant transformation in digital representation systems, progressing from essential product lifecycle management tools to

sophisticated human-centric platforms. This evolution can be characterised by several distinct phases, each marked by technological breakthroughs and paradigm shifts in application domains.

The conceptual foundation of DT technology emerged in 2002 and was published in 2005 at the University of Michigan, where the concept was first introduced [1]. In their retrospective analysis, the authors describe how this initial conceptualisation established the fundamental framework of a digital informational construct capable of mirroring physical systems, laying the groundwork for future DT implementations [2]. A significant advancement occurred through NASA’s adoption and expansion of the DT concept. Their technology roadmap formalised DT as “an integrated multi-physics, multi-scale simulation capable of mirroring the life of physical systems” [3]. This definition expanded the scope of DT applications beyond simple representation to include complex system simulation and analysis capabilities.

A crucial phase in DT evolution was marked by the integration of Internet of Things (IoT) technologies. [4] documented how IoT integration transformed DTs from static representations to dynamic, data-driven models capable of real-time adaptation. This transformation fundamentally altered the nature of digital-physical system interaction, enabling continuous synchronisation and responsive system behaviour.

The incorporation of AI and Machine Learning (ML) capabilities represented another key advancement in DT functionality. Liu et al. [5] discussed how these technologies enhanced the analytical and predictive capabilities of DTs, enabling autonomous decision-making and complex pattern recognition. This integration significantly expanded the application scope of DT technology across various industrial sectors.

The most recent evolution in DT technology has been driven by the emergence of Industry 5.0, which emphasises human-centricity as a core value in manufacturing systems. This shift has led to the development of HDT, defined as a critical method to realise human-centricity in smart manufacturing systems [6]. HDTs are proposed as digital representations of humans that aim to supplement human strengths, amplify productivity, develop human potential, and prioritise physical and mental well-being in smart manufacturing systems. Unlike traditional DTs focused on physical assets, HDTs create explicit connections between humans and technologies to complement their respective strengths, representing a fundamental advancement in human-system integration [7].

2.1.1 Paradigm Shifts in Digital Twin Technology

The evolution from traditional DTs to HDTs can be described through two fundamental categories that capture the complexity of this transformation. These categories reflect not only technological advancement but also our deepening understanding of human-system integration in industrial environments [6].

Human-centric interaction emerges as a critical paradigm shift that places human needs, capabilities, and limitations at the centre of digital twin development. This category represents a departure from traditional system-centred approaches, where humans were expected to adapt to system requirements. Instead, it encompasses the sophisticated evolution of how humans interact with digital systems and how they adapt to human characteristics. This shift reflects a growing recognition that effective digital twins must do more than represent human activities; they must actively support and enhance human capabilities while maintaining natural and intuitive interaction patterns.

Technological Advancements, as the second fundamental category, captures the technical capabilities and infrastructural developments that enable increasingly sophisticated human-digital integration. This category extends beyond mere improvements in processing power or sensor capabilities, encompassing fundamental changes in how we represent human characteristics in digital environments and how we enable meaningful collaboration between humans and machines. These advancements have created new possibilities for representing human knowledge, behaviour, and expertise in digital forms while simultaneously enabling more natural and effective human-machine collaboration.

The first category, Human-Centric Interaction, encompasses two dimensions: the evolution of interactive interfaces and the integration of human factors.

Evolution of Interactive Interfaces: This dimension concerns how humans engage with digital systems. Initially, simple control panels and physical interfaces offered basic command inputs. As technology progressed, these interfaces evolved into sophisticated visualisation systems on desktop computers and touch panels. The most recent advancement in this dimension has been the emergence of extended reality technologies, which have revolutionised human-machine interaction by creating immersive experiences. This progression has enabled increasingly natural and intuitive interactions between humans and digital systems [8].

Integration of Human Factors: This dimension involves the shift from basic physical indicators to complex cognitive and psychological metrics. Early systems primarily considered simple physical parameters such as response times and basic ergonomic factors. Modern systems, however, have expanded to incorporate a comprehensive range of human characteristics, including physical ergonomics, cognitive load, emotional states, and psychological well-being. This evolution reflects a growing recognition that effective human-system integration must account for the full spectrum of human experience in the workplace [9].

The second category, Technological Advancements, encompasses two additional dimensions: digital representation advancement and the evolution of human-machine collaboration.

Digital Representation Advancement: This dimension marks the progression from simple numerical models to sophisticated multi-dimensional representations. Initial

digital representations were limited to basic 2D visualisations and static models. Contemporary systems now incorporate dynamic, adaptive models capable of capturing the complexity of human behavior and characteristics in real-time. This advancement has been particularly significant in enabling predictive capabilities and personalized adaptations based on individual user characteristics [10].

Evolution of Human-Machine Collaboration: This dimension concerns the transformation of how humans and machines interact within digital systems. Initially, humans were often considered external entities, merely observing or providing basic inputs. Modern systems recognise humans as integral components capable of complex decision-making and strategic coordination. This shift reflects a fundamental change in how we conceptualise human-machine relationships, moving from a replacement paradigm to augmentation and collaboration [11].

These categories are deeply interconnected, with developments in one area often catalysing advances in others. For instance, improvements in interactive interfaces have enabled more sophisticated human factor integration, while advances in digital representation have supported the evolution of human-machine collaboration. This interconnectedness highlights the systemic nature of the transformation toward HDTs.

The convergence of these categories has created unprecedented opportunities for human-machine integration in industrial settings. Modern HDTs can now adapt to individual user characteristics, support natural interaction patterns, and facilitate complex collaborative tasks. This evolution has been particularly significant in the context of Industry 5.0, where human-centricity is a core principle guiding technological development.

Understanding these categories is crucial not only for comprehending the historical development of HDTs but also for anticipating future developments. As technology continues to advance, we can expect further evolution along each of these dimensions, leading to even more sophisticated and human-centric digital systems. This ongoing evolution will likely continue to reshape our understanding of human-machine interaction and collaboration in industrial environments.

2.1.2 Human Centric Interaction in Digital Twin Evolution

The evolution of adaptive response capabilities represents another crucial advancement in human-digital integration. Modern systems leverage sophisticated real-time monitoring to create responsive operational environments that dynamically adjust to changing human states and requirements. This adaptation manifests through multiple interrelated mechanisms that work together to optimise human-system interaction. At the core of this advancement lies the development of predictive modelling capabilities that anticipate user needs before performance degradation occurs. As demonstrated by Castro et al. [12],

these systems continuously analyse patterns of human-machine interaction to identify early indicators of potential challenges or opportunities for optimisation. This proactive approach enables systems to implement subtle adjustments that maintain optimal working conditions while avoiding abrupt changes that might disrupt user workflow. Integrating these adaptive capabilities has transformed the nature of human-system interaction, creating environments that actively support human performance rather than simply responding to explicit commands. These systems learn from accumulated interaction patterns, progressively refining their responses based on individual user preferences and demonstrated effectiveness. This learning capability enables the creation of increasingly personalized operational environments that align with each user's unique working patterns and preferences. This shift toward adaptive, learning-enabled systems represents a fundamental advancement in human-digital integration, establishing new paradigms for how digital systems can enhance human capabilities while maintaining natural and effective interaction patterns. The synthesis of these developments has produced interaction paradigms that actively support and enhance human capabilities rather than merely accommodating them. These systems learn from user behaviour patterns, anticipating needs and adapting responses based on accumulated experience. This learning capability enables progressive refinement of interaction patterns, creating increasingly personalised and efficient operational environments.

This evolution in human-centric interaction has established the foundation for more sophisticated human-digital twin implementations, where the digital system serves as an active partner in task execution rather than a passive tool. The resulting interaction paradigms represent a significant advance in human-machine integration, enabling levels of collaboration previously unattainable in industrial environments.

2.1.3 Technological Drivers

The advancement of digital representation technologies has fundamentally transformed the capability to model and simulate human characteristics in virtual environments. As Lu et al. [13] note, this evolution has been characterised by increasingly sophisticated approaches to capturing and representing human behaviour patterns in real-time digital environments. The progression from basic geometric models to complex multi-dimensional representations marks a significant leap in system capabilities. Wang et al. [14] demonstrate how modern systems can now simultaneously capture and process multiple aspects of human behaviour, including physical movements, cognitive states, and social interactions. This multi-dimensional modelling enables a more comprehensive understanding of human users within their operational context. Real-time data integration has emerged as a critical enabler of advanced digital representation. Bourechak et al. [15] highlight how integrating sensor networks and edge computing has enabled continuous monitoring and adaptation of

digital models. These systems can process multiple data streams simultaneously, updating digital representations with millisecond precision to maintain synchronisation with physical reality. The evolution of ML algorithms has particularly enhanced the adaptive capabilities of these systems. As documented by Buerkle et al. [16], modern systems can now learn from accumulated experience, recognising patterns in user behaviour and automatically adjusting their responses to optimise interaction. This learning capability extends beyond simple pattern recognition to include predictive modelling of operator intentions and needs. Contextual awareness represents another crucial advancement in digital representation. Research by Montini et al. [17] demonstrates how modern systems can integrate environmental conditions, task requirements, and social factors into their operational models. This broader awareness enables more nuanced and appropriate system responses to complex operational situations. The convergence of these technological capabilities has enabled unprecedented advances in human-machine collaboration. As Kolade and Owoseni [18] observe, modern systems can dynamically adjust their level of automation and support based on real-time assessment of user needs and capabilities. This adaptive collaboration represents a fundamental shift from traditional fixed-role automation to flexible, context-aware support systems. Knowledge representation and transfer have also evolved significantly. Recent work by Shangguan et al. [19] demonstrates how modern systems can capture and codify complex operational knowledge, enabling more effective transfer of expertise between users and across operational contexts. This capability is particularly crucial in maintaining operational continuity and supporting skill development.

The integration of these technological advances has enabled the development of systems that not only represent human user but actively support their development and performance optimisation. Modern implementations can maintain detailed models of user capabilities, preferences, and learning patterns, enabling personalised support and training interventions. This technological foundation continues to evolve, driven by advances in sensing technologies, computational capabilities, and artificial intelligence algorithms. Table 2.1 summarises the key technological drivers enabling HDT development and their respective impacts.

2.2 Applications and Emerging Challenges in Human Digital Twins

The implementation of HDT systems has demonstrated significant potential across various industrial domains while simultaneously revealing critical challenges in human-system interaction. Analysis of current applications provides insights into both the capabilities and limitations of existing approaches, particularly in the domain of human-system communication and knowledge exchange. Manufacturing environments have seen notable

Table 2.1: Technological Drivers of Human Digital Twin Development

Technology Category	Key Capabilities	Impact on HDT	Key Studies
Sensing & Data Collection	Real-time monitoring; Multi-modal sensing; Wearable technologies	Enables continuous capture of physical, physiological, and behavioural data for accurate human representation	Bourechak et al. [15]
Processing & Computing	Edge computing; Real-time processing; Distributed systems	Facilitates immediate data processing and model updates, ensuring synchronization between physical and digital representations	Montini et al. [17]
AI & ML	Pattern recognition; Predictive modelling; Adaptive learning	Enables system learning from accumulated experience and prediction of operator intentions and needs	Buerkle et al. [16]
Knowledge Management	Knowledge codification; Expertise transfer; Learning patterns	Supports capture and transfer of operational knowledge, enabling personalised training and development	Shangguan et al. [19]
Human-Machine Interfaces	Contextual awareness; Adaptive interaction; Multi-modal feedback	Creates intuitive and natural interaction channels between humans and digital systems	Wang et al. [14]

implementations of HDT systems for process optimisation and worker support. Maruyama et al. [20] demonstrate how HDTs, named Digital Human (DH)v patterns. However, these systems primarily rely on physical sensor data, lacking sophisticated interpretation of human intent and contextual understanding. In maintenance operations, HDT applications have shown promise in knowledge capture and transfer. Naqvi et al. [21] describe systems that attempt to codify expert knowledge for maintenance decision support. While these implementations successfully capture procedural knowledge, they struggle with the nuanced aspects of human expertise expressed through natural communication. Training and skill development represent another significant application domain. Lee et al. [22] showcase HDT systems that support operator training through real-time feedback and guidance. These systems excel at physical task instruction but face limitations in understanding and responding to learner queries and concerns expressed through natural language. The emergence of collaborative industrial environments has particularly highlighted the importance of natural interaction capabilities. Zhang et al. present advanced systems for human-robot collaboration [23], yet note significant challenges in interpreting human intent expressed through verbal communication. These applications reveal a consistent pattern:

while current HDT implementations excel at processing structured data from physical sensors, they often struggle with the complexity of human communication. This limitation becomes particularly apparent in scenarios requiring nuanced understanding of human intent, expertise sharing, or contextual adaptation. As Wang et al. [24] observe, the ability to process and understand human communication represents a critical frontier in HDT development. The analysis of conversational data presents unique challenges distinct from traditional sensor data processing. Unlike physical measurements, conversational data contains implicit context, ambiguous references, and complex semantic relationships. Traditional analysis approaches optimized for structured data prove inadequate for capturing the rich information embedded in human communication. This limitation has significant implications for the evolution of HDT systems. As industrial environments become increasingly collaborative, the ability to understand and respond to natural human communication becomes crucial. The need for sophisticated conversational analysis capabilities represents not merely a technical challenge but a fundamental requirement for realising the full potential of human-digital twin integration.

2.2.1 Privacy and Ethical Considerations

The implementation of HDT systems inherently generates vast amounts of personal data, raising significant privacy concerns that extend beyond traditional data protection frameworks. The intimate nature of HDT data collection, which captures not only physical behaviours but also cognitive patterns and physiological states, necessitates a fundamental rethinking of privacy preservation.

Data anonymisation emerges as a critical technical response to these privacy challenges. However, as Katsoulakis et al. [25] discuss, traditional anonymisation techniques often prove insufficient for HDT applications due to the complex, interconnected nature of human behavioral data. The challenge lies in maintaining data utility while ensuring meaningful privacy protection. Recent work by Faliagka et al. highlights how advanced anonymization techniques, including differential privacy and federated learning approaches, can help balance these competing demands [26]. These techniques represent a significant advancement in privacy preservation by enabling granular control over data exposure while maintaining analytical utility. Differential privacy, in particular, provides mathematical guarantees about the maximum amount of information that can be inferred about any individual from the aggregated data, while federated learning allows models to be trained on distributed datasets without centralizing sensitive information.

The regulatory landscape governing HDT implementations continues to evolve, with the European Union’s AI Act [27] representing a significant advancement in the governance of AI-driven systems. This regulation, building upon General Data Protection Regulation (GDPR) foundations, introduces specific requirements for high-risk AI sys-

tems, a category that encompasses many HDT applications. Park et al. [28] examine how these regulatory frameworks influence system design and implementation choices in collaborative industrial environments. The requirements for transparency, accountability, and human oversight present both technical and organisational challenges for HDT deployment. The emergence of the AI Act particularly impacts HDT systems through its emphasis on algorithmic transparency and fairness. As analysed by Kobayashi et al. [29], the regulation’s requirements for eXplainable AI (XAI) decisions and algorithmic accountability necessitate fundamental changes in how HDT systems process and utilise personal data. These requirements extend beyond simple data protection to encompass the entire decision-making pipeline within HDT systems.

The ethical implications of HDT deployment extend to broader contexts where continuous monitoring and predictive analytics play a role. Recent research by Montini et al. [30] highlights how such systems can influence behaviour and decision-making within monitored environments. These findings underscore the importance of addressing privacy and autonomy concerns in the design and deployment of HDT systems, ensuring they align with ethical standards and societal values. The “right to be forgotten”, a key principle in data protection, presents significant challenges in the context of HDT. Due to the interconnected and distributed nature of HDT data, achieving complete data deletion is technically complex and can target the integrity of the broader system. This issue becomes particularly pertinent when sensitive or outdated information must be erased while maintaining seamless system operations [31]. In addition, algorithmic bias, another critical concern in HDT implementations, arises from the training datasets or the structural design of algorithms. Such biases can inadvertently permeate decision-making systems, influencing outcomes based on skewed or incomplete information. This underscores the need for rigorous checks to mitigate biases that may affect the fairness and reliability of HDT-based analyses and operations [32].

2.2.2 Technical Challenges

The implementation of Human Digital Twin systems presents significant technical challenges that extend beyond traditional data processing paradigms. The inherent complexity of human behaviour and interaction patterns necessitates sophisticated approaches to data integration and analysis that current systems struggle to fully address. A fundamental challenge lies in the heterogeneous nature of HDT data sources. As Rabah et al. [33] shown, HDT systems must process diverse data types ranging from structured sensor readings to unstructured conversational data. This multimodal data landscape includes physiological measurements, motion capture data, environmental sensors, and human-generated content such as text, speech, and gestural information. The integration of these disparate data streams presents significant technical hurdles in terms of synchronization, normalization,

and meaningful fusion. The temporal aspects of data processing add another layer of complexity. Isah et al. [34] emphasize the challenges posed by varying temporal resolutions and latencies in different data streams. To address this, they implemented a Long Short Term Memory (LSTM) network to model long-term dependencies in time series data. However, their approach introduces complexity to the system. In fact, the increased complexity introduced leads to slower processing times, necessitating asynchronous techniques to avoid data loss from high-frequency sensor streams. However, the limitations of conversational data differ. Real-time constraints are less stringent, but the need for contextual understanding and semantic coherence imposes unique challenges. This temporal heterogeneity further complicates the creation of coherent, real-time digital representations. For this reason, knowledge representation and processing present particularly significant challenges. Traditional approaches to data analysis, optimised for structured data, need to be revised when dealing with the complexity of human knowledge and expertise. As Schneider et al. [35] observe, capturing and operationalising human expertise requires sophisticated approaches to knowledge extraction and representation. Current systems struggle to bridge the gap between explicit, easily codifiable knowledge and tacit understanding that emerges through experience and practice. The challenge of multimodal integration becomes particularly acute in the context of human-machine interaction. Zhang et al. [36] describe how current systems struggle to integrate physical actions with verbal communication and contextual understanding. This limitation significantly impacts the system's ability to interpret human intent and provide appropriate responses in complex collaborative scenarios. Natural language processing in industrial contexts presents unique challenges beyond traditional applications. As highlighted by Montini et al. [17], industrial communication often contains domain-specific terminology, abbreviated expressions, and context-dependent meanings that standard NLP approaches fail to capture adequately. Integrating this linguistic data with other sensor streams further complicates the processing pipeline. These technical challenges point toward the need for more sophisticated approaches to knowledge processing and representation. Current limitations in handling multimodal data and extracting meaningful insights from complex human interactions suggest the necessity for advanced analytical frameworks. The development of such frameworks requires careful consideration of both the theoretical foundations of knowledge processing and practical implementation strategies in industrial contexts.

2.3 Knowledge Representation in Human Digital Twins

The effective representation and management of knowledge is fundamental to the success of HDT systems. While sensor data and physical measurements provide the foundation for digital representation, the complex nature of human knowledge and expertise requires more sophisticated approaches to capture and process semantic relationships and contextual

Table 2.2: Technical Challenges and Solutions in HDT Systems

Challenge Domain	Technical Implications	Current Solutions and Limitations
Data Integration [33]	Heterogeneous data sources including physiological measurements, motion capture, environmental sensors, and human-generated content	Implementation of multimodal fusion techniques; Limited by synchronization issues and varying data formats
Temporal Processing [34]	Multiple temporal resolutions across data streams; Real-time processing requirements; Historical data analysis	LSTM networks for temporal modelling; Challenges in balancing processing speed with accuracy
Knowledge Representation [35]	Capturing explicit and tacit knowledge; Converting human expertise into digital format; Maintaining semantic relationships	Knowledge graph implementations; Limited by difficulty in formalizing implicit knowledge
Human-Machine Interaction [36]	Integration of physical actions with verbal communication; Context understanding; Intent recognition	Multimodal interaction systems; Struggles with nuanced interpretation of human behavior
Natural Language Processing [17]	Domain-specific terminology; Context-dependent meanings; Integration with sensor data	Specialized NLP models; Limited by industrial vocabulary and context understanding

understanding. Knowledge Graphs (KGs) have emerged as a powerful tool for representing complex relationships in digital twin environments. As demonstrated by Wei et al. [37], KGs can capture the intricate connections between different aspects of human expertise, creating structured representations that support both human interpretation and machine processing. The graph-based architecture consists of entities (nodes), relationships (edges), and attributes characterising both nodes and relationships. This structure enables the representation of complex relationships between entities, making it particularly suitable for modelling the multifaceted nature of human knowledge in HDT systems. Recent advances in Knowledge Graph Embedding (KGE) have further enhanced the capability to process and utilise structured knowledge representations. Listl et al. [38] show how embedding techniques can transform discrete graph structures into continuous vector spaces, enabling more sophisticated analysis and inference capabilities. These embedded

representations facilitate the integration of structured knowledge with other forms of data, including unstructured text and sensor measurements. In HDT contexts, KGEs provide a powerful framework for combining different types of operator knowledge and expertise into unified representations that can support advanced analysis and decision-making. The temporal dynamics of human knowledge and expertise present particular challenges in HDT environments. Temporal Knowledge Graphs (TKGs) extend traditional KGs to capture how relationships and knowledge evolve over time. As demonstrated by Cao et al. [39], TKGs can model the progression of user capabilities, track the development of expertise, and capture the dynamic nature of human-system interactions. This temporal dimension is crucial for HDT systems that must adapt to evolving operator skills and changing operational requirements. The flexibility of KG architectures makes them particularly well-suited to HDT applications. The ability to represent multi-dimensional relationships allows HDTs to capture not only explicit knowledge but also tacit knowledge derived from operator experience. This comprehensive knowledge representation enables HDT systems to understand better and support human decision-making processes, leading to more effective human-system collaboration.

2.3.1 Knowledge Graph Architectures for HDT

The architecture of Knowledge Graphs in HDT contexts requires specific considerations to effectively capture the complexity of human-machine interactions and expertise. Unlike traditional KGs, HDT Knowledge Graphs must represent not only static relationships but also dynamic human characteristics, operational contexts, and evolving expertise patterns. The fundamental components of HDT Knowledge Graphs include entities representing users, physical systems, operational procedures, and environmental contexts. As Wang et al. [6] demonstrate, these entities must be modelled with sufficient granularity to capture both explicit operational characteristics and implicit human factors. For example, an operator entity, in a working setting, might include attributes ranging from basic qualifications and certifications to more nuanced characteristics such as preferred working methods or specific experiential knowledge. Relationships in HDT Knowledge Graphs serve multiple functions, representing not only operational connections but also expertise transfer paths and collaborative patterns. Zhou et al. [7] highlight how these relationships must capture both direct interactions (such as user-machine control sequences) and indirect influences (such as knowledge transfer between operators or the impact of environmental conditions on human performance). The directionality and attributes of these relationships provide crucial context for understanding how human expertise develops and propagates within the system. The attribute structure of HDT Knowledge Graphs extends beyond simple property-value pairs to include complex metadata that characterises both entities and relationships. For instance, operator attributes might include time-varying skill levels,

cognitive states, and performance metrics, while relationship attributes might capture the quality of interactions, frequency of engagement, and contextual factors affecting performance. This rich attribute structure enables the representation of nuanced human factors that traditional digital twin architectures struggle to capture. Particularly significant in HDT Knowledge Graphs is the representation of human expertise and capabilities. Naqvi et al. [19] explain how KG architectures can model the multi-layered nature of human knowledge, from explicit procedural knowledge to tacit understanding developed through experience. This modelling includes:

- **Explicit Knowledge:** Formalized procedures, documented best practices, and certified competencies
- **Experiential Knowledge:** Insights gained through practical experience and problem-solving
- **Collaborative Knowledge:** Understanding developed through interaction with other operators and systems
- **Contextual Knowledge:** Awareness of how environmental and operational factors influence performance

The integration of sensor data with semantic relationships presents unique architectural challenges in HDT Knowledge Graphs. Miller et al. [16] describe how KG architectures must bridge the gap between continuous sensor streams and discrete semantic representations. This integration requires sophisticated mapping mechanisms that can translate raw sensor data into meaningful semantic relationships while preserving temporal and contextual information.

2.3.2 Temporal Aspects and Evolution

The temporal dimension of knowledge representation in HDT systems introduces unique challenges in capturing the dynamic nature of human expertise and system interactions. Temporal Knowledge Graphs (TKGs) provide a framework for modeling these evolutionary aspects, extending beyond static representations to capture how human knowledge, skills, and interactions develop over time. The formal structure of TKGs in HDT contexts can be defined as $G = (E, R, T, F)$, where E represents entities (such as operators, machines, and procedures), R defines relationships, T denotes timestamps, and F represents temporal facts. This framework enables the representation of time-sensitive information critical to understanding human performance and development. As Wang et al. [6] demonstrate, this temporal modeling is essential for tracking the progression of operator capabilities and the evolution of human-machine interactions. The temporal granularity in HDT Knowledge

Graphs must accommodate multiple time scales, from immediate interactions to long-term skill development:

- Real-time Interactions: Capturing immediate operator-machine interactions and momentary performance metrics
- Short-term Patterns: Representing daily or weekly variations in performance and operational behaviors
- Medium-term Development: Tracking skill acquisition and procedural knowledge development
- Long-term Evolution: Modeling career progression and expertise development over years

A particular challenge in temporal modeling for HDTs is the representation of evolving human expertise. Zhou et al. [7] highlight how temporal patterns in operator performance and decision-making must be captured to understand skill development trajectories. This includes modeling both explicit skill advancement through training and implicit knowledge gained through experience. The TKG structure enables the tracking of these developmental patterns by maintaining temporal relationships between operators, tasks, and performance outcomes. The evolution of collaborative relationships presents another critical aspect of temporal modeling in HDT systems. Buerkle et al. [16] describe how human-machine collaboration patterns change over time as operators develop increased familiarity and expertise with systems. TKGs must capture these evolving interaction patterns to enable adaptive system responses that align with operator development. This includes modeling:

- Changes in operator confidence and system trust
- Evolution of preferred interaction methods
- Development of specialized operational strategies
- Refinement of collaborative problem-solving approaches

The integration of temporal information with other aspects of HDT knowledge representation requires sophisticated modeling approaches. Montini et al. [17] show how temporal relationships must be considered alongside other forms of semantic information to create comprehensive understanding of human-system interactions. This integration enables HDT systems to:

- Predict future performance based on historical patterns
- Adapt support strategies to operator development stages
- Identify optimal intervention points for training and assistance
- Track the effectiveness of learning and development initiatives

2.3.3 Knowledge Integration and Embeddings

Knowledge Graph Embeddings (KGE) in HDT systems transform complex human-related knowledge structures into continuous vector spaces, enabling sophisticated analysis and inference capabilities. This transformation is particularly crucial for HDTs, where diverse forms of knowledge, from explicit operational procedures to tacit experiential understanding, must be integrated into a unified computational framework. The embedding process in HDT contexts aims to preserve both the structural and semantic properties of human knowledge while enabling computational analysis. As described by Montini et al. [17], KGEs capture not only direct relationships between entities but also higher-order patterns that characterize human expertise and operational behaviors. The embedding space accommodates multiple dimensions of information, ranging from structured operational knowledge and protocols to unstructured experiential insights and problem-solving strategies. Temporal patterns of skill development and interaction histories are seamlessly integrated with social relationships and knowledge transfer pathways. The integration of different knowledge modalities presents unique challenges in HDT systems. Traditional embedding approaches, designed for static knowledge structures, must be adapted to handle the dynamic and multi-modal nature of human-system interactions. Wang et al. [6] demonstrate how advanced embedding techniques can bridge the gap between different knowledge representations. These techniques enable the fusion of sensor data with semantic knowledge while maintaining the temporal patterns and static relationships that characterize human-system interactions. The result is a comprehensive representation that aligns human expertise models with system capabilities. A particularly significant aspect of knowledge embeddings in HDT contexts is the representation of tacit knowledge. Unlike explicit knowledge that can be directly encoded, tacit knowledge emerges from experience and is often difficult to formalize. Zhou et al. [7] show how embedding techniques can capture these implicit patterns through analysis of operational behavior sequences and identification of successful problem-solving strategies. This capability is crucial for preserving and transferring expertise that might otherwise be lost in traditional knowledge management systems. The effectiveness of knowledge integration through embeddings depends critically on the choice of embedding architecture and training approach. Contemporary HDT systems employ a range of embedding techniques, from TransE variants for basic operational relationships to complex embeddings for temporal dynamics. Hyperbolic embeddings prove particularly effective for representing hierarchical knowledge structures, while multi-modal approaches enable the integration of diverse data sources into cohesive representations. The scalability of embedding approaches becomes particularly important as HDT systems accumulate knowledge over time. Buerkle et al. [16] describe how incremental learning techniques can be employed to update embeddings as new knowledge is acquired. These approaches maintain the relevance and accuracy of the knowledge

representation while managing computational resources effectively, ensuring the system remains responsive to evolving operational needs. Future directions in HDT knowledge embeddings point toward increased integration of contextual awareness and adaptive capabilities. Emerging research suggests the potential for embedding architectures that can dynamically adjust their representation based on operational context and user needs. This adaptivity promises to enable more nuanced and effective human-system collaboration, furthering the goal of seamless integration between human expertise and digital systems.

Table 2.3: Knowledge Graph Capabilities and Limitations in HDT Systems

Aspect	Current Capabilities	Limitations & Challenges
Structural Representation	Effective modeling of explicit relationships and hierarchies between operators, machines, and procedures	Difficulty in representing informal or implicit relationships that characterize human behavior
Temporal Modeling	Can track discrete changes and evolutionary patterns in expertise and interactions	Limited ability to handle continuous temporal changes and multiple time scales simultaneously
Knowledge Integration	Successfully combines structured data from multiple sources (sensors, documentation, protocols)	Struggles with integrating tacit knowledge and contextual understanding
Embedding Capabilities	Enables mathematical operations on knowledge representations for inference and analysis	Current embeddings may lose subtle semantic nuances crucial for human expertise modelling
Scalability	Handles large volumes of structured relationships and attributes	Computational overhead increases significantly with temporal and contextual complexity
Adaptability	Can update representations based on new data and relationships	Limited real-time adaptation to rapid changes in human behavior or context
Human-Centric Features	Represents explicit human characteristics and documented expertise	Difficulty capturing emotional states, intuition, and dynamic social interactions

2.3.4 From Knowledge Representation to Content Analysis

While Knowledge Graphs provide powerful mechanisms for representing structured relationships in HDT systems, a significant portion of human knowledge exists in unstructured textual form. Documentation, reports, communications, and operational notes contain valuable insights that must be effectively processed and integrated into the broader knowl-

edge framework. As Wang et al. [6] observe, these textual sources often contain critical contextual information and experiential knowledge that complement structured KG representations. The challenge of processing this unstructured content becomes particularly significant in HDT environments, where textual information spans multiple domains and contexts. Traditional knowledge representation approaches struggle to capture the nuanced insights embedded in natural language descriptions of operational procedures, problem-solving strategies, and expert observations. This limitation has led to increased interest in techniques that can automatically analyze and categorise textual content, bridging the gap between unstructured documentation and structured knowledge representations.

2.4 Topic Classification for Knowledge Representation

The convergence of advanced speech recognition technologies and sophisticated language understanding capabilities has fundamentally transformed the potential applications of topic classification in HDT systems. While traditional applications focus primarily on document analysis, modern HDT implementations must process and understand a broader spectrum of human communication, including real-time conversations, informal dialogues, and multi-participant discussions. This evolution in communication processing capabilities enables HDTs to maintain a more comprehensive and nuanced understanding of user interactions and knowledge patterns.

Recent advances in Automatic Speech Recognition (ASR) technology have achieved remarkable accuracy levels, with error rates approaching human performance in many contexts. As demonstrated by [park2024recent], modern ASR systems can achieve word error rates below 5% in controlled environments, while maintaining robust performance even in challenging acoustic conditions. This technical advancement removes a crucial barrier to processing verbal communications, enabling HDT systems to capture and analyse spoken interactions with unprecedented fidelity.

The recent advancement in NLP with LLMs has further enhanced the capability to process and understand human communication in its various forms. These models demonstrate sophisticated comprehension of linguistic nuance, contextual relationships, and conversational flow, enabling more natural and effective human-machine interaction. When integrated with topic classification capabilities, these advances enable HDT systems to process and categorize real-time conversations while maintaining contextual awareness throughout extended dialogues. The ability to track topic evolution and transitions within conversations, particularly in multi-participant discussions, represents a significant advancement in HDT functionality.

The systematic analysis of human communication through topic classification has

thus emerged as a fundamental capability for modern HDT systems. This capability extends beyond traditional document processing to encompass the full spectrum of human interaction. This broader application enables automatic organisation and categorisation of conversational knowledge, facilitating more effective information discovery and utilisation within HDT frameworks.

The integration of topic classification with conversational analysis presents unique challenges that distinguish it from traditional document processing. Conversations typically exhibit non-linear topic progression, often containing multiple concurrent themes and context-dependent meanings. The informal nature of spoken communication, combined with rapid topic transitions and returns, requires sophisticated approaches that can maintain coherent topic tracking while adapting to the dynamic nature of human communication.

Modern HDT systems must employ robust classification frameworks capable of processing streaming conversational data in real-time while maintaining context across multiple interaction sessions. These systems need to adapt to individual communication styles and patterns while integrating various communication signals. This sophisticated processing enables HDT systems to build more comprehensive and accurate models of human knowledge and interaction patterns, proving particularly crucial in professional environments where tracking and understanding evolving discussions can significantly impact decision-making processes and knowledge transfer.

Furthermore, the application of topic classification to conversational data enables HDT systems to identify and leverage patterns in human communication that might not be apparent in formal documentation. These patterns can reveal individual expertise and knowledge domains, along with communication preferences and professional interests. Such enhanced understanding of communication patterns and knowledge distribution enables HDT systems to provide more targeted and effective support, adapting their responses and recommendations to specific conversational contexts and user needs.

This evolution in topic classification capabilities represents a crucial advancement in HDT technology, enabling these systems to process and understand human communication in its most natural forms. The ability to effectively classify and organise conversational content enhances the HDT's capacity to serve as an intelligent partner in human-machine interaction, maintaining a comprehensive understanding of user context while delivering personalised and adaptive support.

2.4.1 Text Classification Approaches and Methods

Text classification represents a fundamental task in natural language processing, requiring an input text and returning the most likely category that summarises the text. This procedure can be referenced through various terms, such as topic analysis, topic extraction, or automatic labelling [40]. The classification process typically involves two main stages:

feature extraction/selection and the actual classification task. The effectiveness of text classification methods varies based on several parameters, including text length, writing style, and content type. As highlighted by Munir et al. [41], document length significantly impacts classification performance - short texts (under 40 words) provide limited information compared to medium (40-100 words) and long texts (exceeding 100 words). This consideration becomes particularly relevant when processing industrial documentation and technical reports, where document length can vary substantially.

Feature Extraction and Selection Methods Feature extraction represents a critical initial step, involving the selection of essential features and their transformation into numerical representations. This transformation is necessary since machine learning models require vector formats for processing [42]. The approaches range from basic techniques to more sophisticated embedding methods: Bag of Words (BoW) provides a straightforward approach by representing text as unique word lists, typically using one-hot encoding for numerical conversion. However, this method presents significant limitations for technical documentation analysis, as noted by Behera et al. [43], particularly regarding high computational requirements for large vocabularies. Term Frequency-Inverse Document Frequency (TF-IDF) offers improved results by weighting words based on their frequency and document occurrence. While more effective than BoW, TF-IDF still faces computational challenges due to sparse matrix creation [44]. N-grams capture local context by considering word sequences, though computational requirements increase with N-gram length [45]. Modern approaches increasingly utilise word embedding techniques, which capture both syntactic and semantic aspects of words [46]. These methods are particularly valuable for technical document analysis, as they can identify semantic similarities between technical terms and concepts, facilitating more nuanced classification.

Classification Methodologies Text classification approaches can be categorized into three primary paradigms: supervised, unsupervised, and semi-supervised learning. The supervised approach dominates current applications due to its effectiveness and validation simplicity. Among supervised methods, several algorithms have demonstrated particular effectiveness: Support Vector Machines (SVM) excel in text classification due to their ability to handle multiple features while mitigating overfitting issues. As demonstrated by Belkebir et al. [47], SVMs perform especially well with sparse vectors, which commonly represent technical documents. Deep Learning approaches have shown remarkable results in processing large volumes of textual data. Recurrent Neural Networks (RNN) [48] and Long Short-Term Memory (LSTM) networks [49] handle sequential data effectively, while Convolutional Neural Networks (CNN) [50] excel at extracting local features and word relationships. Recent advances in transformer models have further enhanced classification capabilities. Studies by Shi et al. [51] demonstrate the effectiveness of combining RoBERTa

with BiLSTM networks for topic classification across various datasets. The emergence of Large Language Models (LLMs) like GPT-3.5 and Llama has introduced new possibilities, as shown by Milios et al. [52], who achieved superior results using Llama for classification without fine-tuning.

Recent Developments and Applications Contemporary research has explored various combinations of feature extraction and classification methods. For instance, Shah et al. [53] utilized BERT for both feature extraction and classification in news topic analysis. Kavitha et al. [54] employed FastText for embedding generation before feeding data into a multichannel CNN. Kumar Velu et al. [55] focused on short text classification using a combined CNN-LSTM architecture optimised through Caledonian crow optimization. This hybrid approach demonstrates the potential of combining multiple techniques for improved classification performance. Transformer-based architectures continue to show promising results, as evidenced by Mandal et al. [56], who successfully applied ALBERT and RoBERTa for short text classification. The attention mechanisms in these models enable complex relationship capture between words, facilitating deeper semantic understanding. When applied to industrial contexts and human digital twin systems, these classification approaches must often be adapted to handle domain-specific terminology and varying document lengths. The choice of classification method typically depends on factors such as document characteristics, computational resources, and specific application requirements.

Chapter 3

Human Digital Twin Model

In Chapter 4, a comprehensive framework for topic modelling and classification has been introduced, showing how ML techniques can effectively process and categorise textual information. Through careful feature selection and advanced neural architectures, our system showed the potential of automated text analysis for understanding complex linguistic content. The development of HDT represents a significant evolution in how digital systems interact with and understand human users. While traditional IoT architectures have focused primarily on sensor data and explicit commands, modern HDTs must process and understand a broader range of human communication, including natural language. This integration of verbal communication capabilities marks a crucial step toward more intuitive and responsive digital systems. Our proposed framework introduces a novel architecture emphasising proactive user engagement while balancing system performance and resource utilisation. At its core lies the Digital Brain, a processing unit that integrates information from multiple sources to maintain an accurate model of user state and intentions. This component implements three primary processing pathways to capture different aspects of human-system interaction. The first pathway focuses on acquiring user knowledge through continuous monitoring of both verbal and non-verbal communication channels. This process builds upon established topic modelling techniques while introducing additional layers for context awareness. The system processes verbal input through advanced speech recognition models, converting spoken language into structured representations that can be analysed alongside other interaction data. This integration enables a more comprehensive understanding of user intent by combining explicit verbal commands with implicit behavioural signals. The second pathway implements cumulative knowledge processing, creating and updating dependencies between various user interests and behaviours. This component maintains a dynamic knowledge graph structure that evolves based on user interactions, enabling the system to recognise and adapt to changing patterns of behaviour and preference. The temporal nature of this processing ensures that the system maintains awareness of both immediate context and longer-term trends. This

accumulated knowledge serves as the foundation for interaction patterns and proactive assistance. The third pathway focuses on insights utility evaluation, representing the system’s ability to predict future user needs and anticipate requests. This component leverages the accumulated knowledge to generate contextually relevant responses, evaluating potential actions based on both immediate context and historical patterns. The evaluation process considers multiple factors including short-term activities, long-term behavioural patterns, and environmental context, creating a more nuanced understanding of user needs. A key innovation in our approach lies in the handling of always-listening capabilities. The framework implements an adaptive processing strategy that allows users to choose between two operational modes: one optimized for Quality of Experience (QoE) and another focused on sustainability and privacy preservation. This flexibility enables users to balance system responsiveness against resource utilization and privacy concerns based on their specific needs and preferences. The QoE-focused mode maintains comprehensive processing of user interactions, enabling more precise context understanding and proactive assistance. This mode implements data fusion techniques that combine information from multiple sources to maintain accurate user state representation. The sustainability-focused mode, in contrast, implements selective processing strategies that reduce computational overhead while maintaining essential system functionality. This dual-mode approach recognises the inherent trade-offs in HDT implementation between system capability and resource efficiency. Rather than imposing a single operating paradigm, our framework empowers users to make informed decisions about how their digital twin processes and responds to interactions. This flexibility proves particularly valuable in real-world applications where requirements may vary based on context or user preference. The proposed framework represents a significant advancement in HDT technology, particularly in its treatment of verbal communication and contextual understanding. By providing flexible operational modes and processing capabilities, the system enables more natural and effective human-machine interaction while addressing crucial concerns about privacy and resource utilization. This approach establishes a foundation for future developments in human-centric digital systems, where understanding and responding to natural communication becomes increasingly central to effective system operation.

3.1 Reference Scenario and System Overview

Our proposed framework is designed to proactively enhance user experiences by leveraging contextual understanding and data aggregation. This section presents a use case showing how our proposed framework delivers personalised assistance tailored to individual preferences and specific circumstances. In this scenario, we consider Alice, who has been feeling exhausted and frustrated from work. This feeling can be obtained from a combination of non-verbal language information, from her connected devices, and from Alice’s exclamation

about the need for a vacation. The HDT system begins to form an understanding of her current state through these various inputs, setting the stage for more meaningful interactions. One evening, while having dinner with friends, the conversation turns to travel plans. Her friends excitedly discuss the possibility of visiting London or Paris soon. Though intrigued by the idea, Alice regretfully mentions that she cannot join due to her heavy workload, expressing her disappointment at missing out on such an opportunity. During this conversation, Alice engages in detailed discussions about both destinations, sharing her interest in experiencing the charm of Paris and exploring London's rich history. Her friends describe the unique attributes of each city, from Paris's art and cuisine to London's cultural heritage and traditions. The conversation provides rich contextual information about Alice's interests and current constraints. Her verbal exchanges with friends reveal not only her travel preferences but also hint at the underlying tension between her desire for a vacation and her professional obligations. This natural dialogue creates a complex web of information that the HDT system can process and understand. A week later, Alice is at work, taking a short break with a colleague. She buys a croissant from the vending machine, and as she bites into it, she comments that the croissant is disappointing. Her colleague, noticing her reaction, suggests trying the croissants in Paris. This casual exchange leads to a broader discussion about French pastries and Parisian cafes, with Alice expressing her longstanding dream of visiting the city. In response to these interactions, the HDT system identifies an opportunity for meaningful engagement. Through her PDA, the platform communicates a tourism experience specifically crafted for Alice, suggesting that she visit Paris in the upcoming autumn. This timing consideration accounts for her current workload while offering the promise of experiencing the city during one of its most beautiful seasons. The suggestion encompasses the opportunity to enjoy authentic croissants while walking through Parisian streets lined with trees, connecting directly with her expressed interests and recent experiences. The innovation of the framework behind this scenario lies not only in its intelligent responses, as many AI-based solutions already provide them, but also in its proactivity. It offers continuous learning about the user, anticipating her needs, and acts on them without requiring a direct request. The system demonstrates its understanding of context by connecting seemingly unrelated events - a disappointing croissant, a previous conversation about travel, and signals of work-related stress - into a meaningful and personalized suggestion. Indeed, without detailed insights about Alice, the system could only offer basic assistance, such as comparing ticket prices or providing travel advisories for both cities. However, with a deeper understanding of her preferences, the platform makes specific, appealing recommendations based on real-time inputs and past interactions, significantly enhancing and personalising the user experience. The scenario illustrates how the combination of verbal communication analysis, contextual awareness, and proactive engagement can create more meaningful and relevant interactions between users and their digital twins. Indeed, without detailed insights about Alice, the

system could only offer basic assistance, such as comparing ticket prices or providing travel advisories for both cities. However, with a deeper understanding of her preferences, the platform makes specific, appealing recommendations based on real-time inputs and past interactions, significantly enhancing and personalising the user experience.

In the following sections, we present a detailed analysis of the framework that enables such interactions. We begin by examining the enabling paradigms that form the foundation of our approach, focusing particularly on the evolution from traditional Digital Twins to human-centric systems. This analysis is followed by a comprehensive examination of our IoT-PDA framework, detailing its architecture and core components. Special attention is given to the Digital Brain architecture, which implements the sophisticated processing capabilities demonstrated in our reference scenario. Through this systematic examination, we establish how our framework achieves the contextual awareness and proactive assistance capabilities illustrated in Alice’s case while maintaining privacy protection and system security.

3.2 Enabling Paradigms

Building upon our discussion in Section 4, where we explored the foundational evolution from Digital Twins to Human Digital Twins, this section extends that analysis to examine the critical paradigm shifts enabling next-generation IoT systems. While our previous discussion focused primarily on the technological transition from DT to HDT, we now explore a more comprehensive transformation: the evolution from reactive to proactive systems, fundamentally changing how digital twins interact with and serve their users.

3.2.1 From Digital Twin to Human Digital Twin: An Enhanced Perspective

The transition from traditional Digital Twins to Human Digital Twins represents more than a simple technological evolution—it marks a fundamental shift in how we conceptualise digital representations of human entities. This transformation is enabled by the convergence of advanced data integration technologies and representation techniques.

Modern HDTs exhibit two fundamental characteristics that distinguish them from traditional DTs: dynamic adaptation and multimodal integration. The dynamic adaptation capabilities of HDTs demonstrate remarkable flexibility in creating and maintaining virtual replicas of individual users across various domains. This flexibility extends beyond simple state representation to encompass the complex and fluid nature of human behavior and preference. Through machine learning algorithms, HDTs continuously adapt to users’ changing states and behaviours in real time, processing and integrating behavioural data streams to maintain comprehensive, evolving profiles. These profiles reflect current

user states while preserving crucial historical context, enabling the system to generate increasingly personalised services and recommendations that evolve alongside the user’s changing needs and preferences.

The multimodal integration capabilities of modern HDTs represent another significant advancement over traditional Digital Twin approaches. These systems implement perception capabilities that combine both quantitative and qualitative aspects of human life, creating holistic representations of behaviour, preference, and emotional state. The integration of emotional intelligence through facial expression analysis and voice tone interpretation enables systems to recognise subtle variations in user mood and intent. Additionally, the processing of physiological indicators provides deeper insight into the user state, while pattern recognition and synthesis capabilities enable the development of creative intelligence—allowing systems to generate novel solutions and suggestions based on learned patterns of user behaviour and preference.

3.2.2 Evolution Toward Proactive Assistance

The integration of LLMs with Personal Digital Assistants (PDAs) marks a crucial advancement in human-machine interaction capabilities, fundamentally transforming how digital systems understand and respond to user needs. This integration enables contextual interpretation and ubiquitous assistance capabilities that extend far beyond traditional reactive systems.

LLM-powered PDAs implement advanced contextual understanding that transcends simple command processing. These systems integrate multiple data streams to form a coherent understanding of user context, combining historical behaviour patterns with current environmental conditions and emotional state inference. This multi-faceted contextual awareness enables systems to maintain temporal context, ensuring interactions occur at appropriate times and in suitable ways. The synthesis of these various data streams creates a rich understanding of user state that enables more natural and effective assistance.

Modern PDAs exhibit pervasiveness across devices and geographical locations, offering contextually appropriate interactions wherever users require support. This ubiquity is achieved through seamless integration with IoT infrastructure and state management systems that maintain consistency across multiple devices and platforms. The systems implement context-aware interaction modality selection, choosing appropriate communication channels based on user situation and preference. Crucially, this pervasive assistance is governed by robust privacy-aware information sharing protocols that ensure sensitive data remains protected while enabling effective cross-device coordination.

As illustrated in Figure 5.1, these enabling paradigms create a comprehensive framework for human-digital interaction. The system maintains continuous awareness of user state and context while implementing privacy protections and adaptive response mechanisms. This

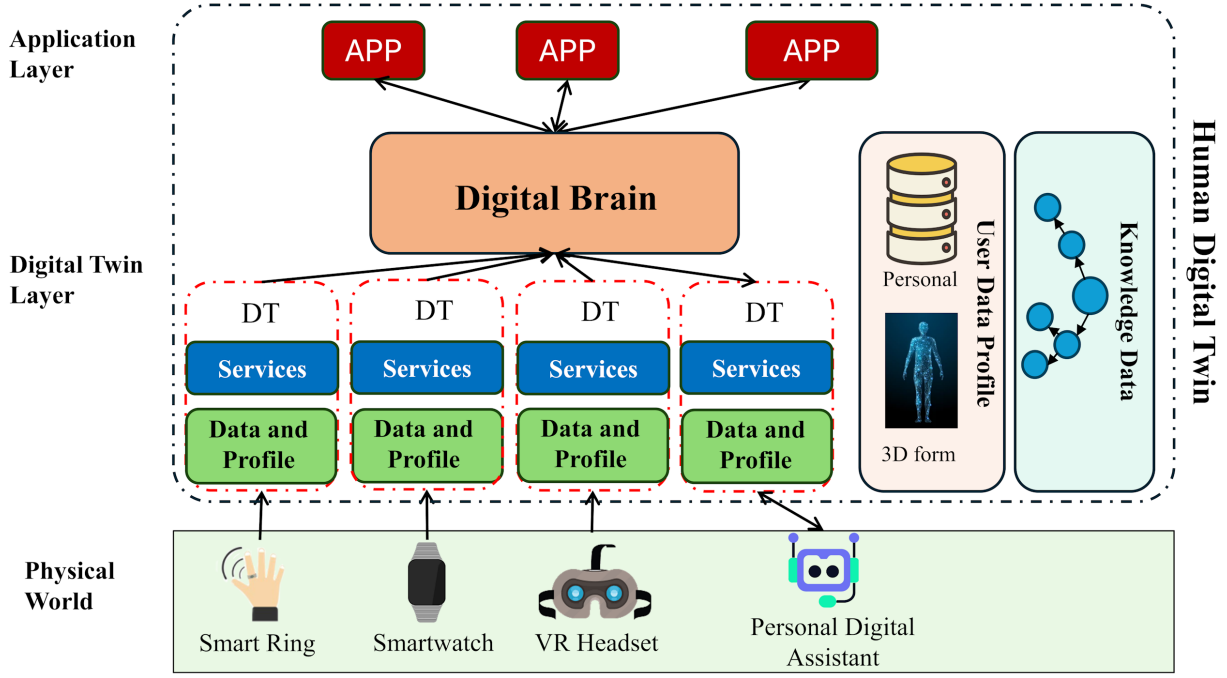


Figure 3.1: High level reference scenario.

architecture enables the development of truly proactive assistance systems that anticipate and respond to user needs before explicit requests are made, fundamentally transforming the nature of human-digital interaction.

This enhanced understanding of enabling paradigms provides the foundation for our subsequent discussion of the IoT-PDA framework implementation in Section 3.3. The integration of these concepts enables the development of systems that not only understand user needs but actively anticipate and address them, creating more natural and effective human-digital interaction environments.

3.3 The Envisioned IoT-PDA Framework

The evolution of IoT systems toward more personalised and proactive user experiences demands frameworks capable of interaction and adaptation. In this section, we present an IoT-PDA framework, which fundamentally presents how digital systems can deliver personalised and proactive user experiences by integrating verbal language processing. This framework builds upon the concept of DT, establishing a comprehensive digital counterpart for every real-world entity within the system. The reference user-centric scenario, illustrated in Figure 3.1, shows the relationships between physical entities and their digital representations. The framework's architecture implements three fundamental keystones that together enable human-digital interaction. The first of these, the HDT component, represents a significant advancement in how digital systems model and interact

with users. When users register on the platform, the cloud provider establishes a virtual user space, represented by the dashed black line in Figure 3.1, that serves as their digital counterpart. This space provides both storage and computing resources necessary for maintaining user information and executing DT applications. The implementation leverages modern containerization technologies and micro services principles, with orchestrators like Kubernetes managing resource allocation and system efficiency. Within this virtual space, the system maintains two distinct types of user information. Static information resides in the user data profile, encompassing fixed attributes such as personal data and three-dimensional representations of the user. This foundational layer provides context for more dynamic interactions. Simultaneously, the Digital Twin Layer maintains active representations of physical entities, exposing both their basic profiles and raw data streams. This layer extends beyond simple representation, offering additional services that users can selectively activate based on their needs and preferences. The resulting HDT transcends traditional digital representation, integrating information from diverse and heterogeneous sources. This integration encompasses raw sensor data, application states, and service interactions, creating a comprehensive digital model of the user's presence and activities within the system.

The architecture raises important research questions regarding the adaptation of orchestration systems for HDT-specific requirements, particularly in areas of scalability, modularity, and secure isolation while maintaining efficient resource utilization. The second keystone, Knowledge Data, implements mechanisms for capturing and utilizing user behaviour patterns. The system continuously monitors user interactions, whether through direct conversation with others or engagement with smart devices and applications. This observational data feeds into a learning process that builds detailed models of user habits, interests, preferences, and thought processes. The framework stores this accumulated knowledge within the user's HDT using graph-based representations, supporting multiple formalisms including Resource Description Framework (RDF), Knowledge Representation Learning (KRL), and KGE. These knowledge graphs serve a crucial function in maintaining individualised behavioural models. Rather than implementing a shared knowledge structure across users, the system constructs and maintains separate graphs for each individual, allowing for a highly personalised representation of behavioural patterns and preferences. This approach ensures that the system can capture and respond to the unique characteristics and needs of each user while maintaining appropriate isolation between user data. The Digital Brain represents the third keystone of the framework. This component enables learning and adaptation capabilities, allowing HDTs to evolve based on observed user behaviour. The Digital Brain implements knowledge-sharing mechanisms between users' digital twins while maintaining appropriate privacy boundaries. This sharing capability enables the system to learn from collective experience while preserving individual user autonomy and data protection.

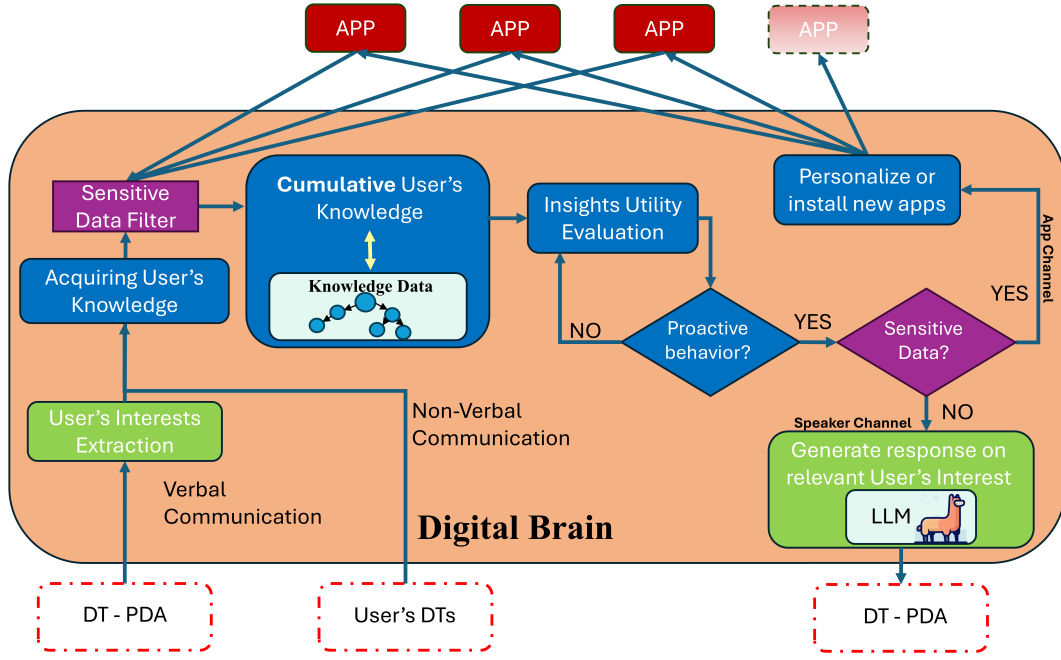


Figure 3.2: Digital Brain structure.

3.4 Digital Brain Architecture

The Digital Brain represents the core processing component of our framework, implementing mechanisms for integrating and analysing diverse user interaction data. Figure 3.2 illustrates the comprehensive architecture of this system, where blue modules represent the main information flow, integrating data from both verbal interactions through PDAs and non-verbal communication through IoT devices. The architecture incorporates specialised components for verbal communication processing (shown in green) and implements robust privacy protection mechanisms (shown in violet). The Digital Brain's architecture reflects the growing convergence of adaptive learning models, neural architectures, and LLM. This integration presents significant research challenges, particularly in developing effective collaborations between unsupervised methods such as Latent Dirichlet Allocation (LDA) and supervised approaches like Support Vector Machines (SVMs). The system must balance these different methodologies while maintaining coherent processing of user interactions across multiple modalities. The architecture implements three primary processing pathways, each addressing distinct aspects of user interaction analysis and knowledge management:

3.4.1 Acquiring User Knowledge

The knowledge acquisition component implements a multi-modal approach to understanding user interests and behaviours. At its core, the system employs ASR models, particularly

Whisper-based transformers¹, to process verbal communication. This speech processing capability enables robust conversion of spoken language into structured text representations suitable for further analysis. For IoT-specific data processing, the system could employ fine-tuned LLMs, adapted using Parameter-Efficient Fine-Tuning techniques. This adaptation enables the models to effectively process IoT environments' unique patterns and contexts. The integration of Low-Rank Adaptation (LoRA) techniques proves particularly valuable in maintaining model efficiency while enabling domain-specific optimisation. The true innovation in this component lies in its synthesis capabilities. By implementing LLMs as a unifying framework, the system can effectively process and integrate diverse data streams, including IoT sensor readings, application logs, and verbal communications. This integration enables a comprehensive interpretation of both explicit user commands and implicit behavioural patterns. The system employs specialised topic classification algorithms that work in concert with the fine-tuned LLMs, enabling the refinement of broad interaction patterns into specific, actionable insights.

3.4.2 Cumulative Knowledge Processing

The cumulative knowledge component implements mechanisms for maintaining and updating relationships between user interests and behaviours. This module maintains continuous communication with the knowledge data storage, enabling dynamic updates to the user's digital profile. The implementation leverages advanced graph database technologies, particularly Neo4j [57], to create dynamic knowledge structures capable of representing complex, interconnected user data. This graph-based approach enables the modelling of relationships between different aspects of user behaviour and preference. The system implements a dual-phase learning approach, combining immediate updates with long-term pattern recognition. The immediate update mechanism employs rule-based processing to adjust edge weights within the knowledge graph based on observed topic frequencies and interaction patterns. This real-time adaptation ensures responsive system behaviour while maintaining coherent knowledge representation. The system implements periodic model retraining processes for long-term learning that capture evolving user interests and behavioural patterns. This temporal aspect of knowledge accumulation enables the system to maintain awareness of both immediate context and broader behavioural trends. The integration of Graph Neural Networks (GNNs) and Graph Attention Networks (GANs) enhances the system's ability to capture subtle relationships within the knowledge structure. These networks excel at learning appropriate weightings for different node connections, enabling the system to infer complex dependencies that might not be immediately apparent through simpler analysis methods. This capability proves particularly valuable in understanding how different aspects of user behaviour and interest relate to each other over time.

¹<https://huggingface.co/openai/whisper-large>

3.4.3 Insights Utility Evaluation

The Insights Utility Evaluation module represents the culmination of the Digital Brain’s processing capabilities, implementing mechanisms for predicting user needs and generating proactive assistance. This critical component leverages the accumulated knowledge from both verbal and non-verbal interactions to anticipate user requirements and generate contextually appropriate responses. The implementation builds upon LLMs, extending beyond simple response generation to create a comprehensive system for understanding and predicting user needs. Traditional approaches to user assistance often rely on basic scoring mechanisms that fail to capture the complexity of human behaviour and preference. Our framework addresses this limitation through the implementation of a multi-faceted evaluation system that considers multiple dimensions of user context and behaviour. This scoring system integrates various aspects of user state and context. In the temporal domain, the system maintains awareness of both immediate activities and long-term behavioural patterns, enabling response generation that aligns with both current needs and established preferences. The evaluation process extends beyond simple activity tracking to include emotional state assessment and contextual factors such as location and time of day. This comprehensive approach enables the generation of suggestions that not only meet immediate needs but also align with broader patterns of user behaviour and preference. The system implements a robust feedback mechanism to evaluate and refine its suggestions. This evaluation process operates through both explicit and implicit channels, monitoring direct user responses while also interpreting subtle behavioural cues that might indicate the effectiveness of system interventions. The framework maintains continuous adaptation through a feedback loop that enables progressive refinement of its understanding of user preferences and needs. The implementation of this evaluation system represents a significant advancement in proactive digital assistance. Rather than waiting for explicit user requests, the system maintains constant awareness of user context and potential needs, enabling it to generate meaningful suggestions before they are explicitly requested. This proactive capability fundamentally transforms the nature of human-digital interaction, creating opportunities for more natural and intuitive system assistance. Table 3.1 presents a comprehensive overview of the technologies and approaches implemented across these three core modules. This analysis identifies both currently available building blocks and critical areas requiring further research and development. The table highlights the immediate possibilities offered by existing technologies while also identifying crucial gaps that must be addressed to realize the full potential of these systems. The integration of these three core modules, i.e., Acquiring User Knowledge, Cumulative Knowledge Processing, and Insights Utility Evaluation, creates a framework for understanding and responding to user needs. The system maintains continuous adaptation through its feedback mechanisms while implementing robust privacy protections to ensure secure and appropriate handling of user

Table 3.1: Technical Overview of Digital Brain Components

Component	Available Technologies	Still Needed	Potential Applications
Acquiring Knowledge	Speech recognition	IoT-specific embeddings	Speech-to-text conversion
	Fine-tuning techniques	Multi-modal fusion (text + IoT Data)	Interest extraction
	Unsupervised learning	Contextual understanding	Semantic analysis
Cumulative Knowledge	Graph databases	Dynamic weight updates	Knowledge representation
	Graph Neural Networks	Long-term learning	Relationship inference
Insights Evaluation	Multi-faceted scoring	Context-aware triggers	Proactive suggestions
	Feedback mechanisms	Adaptive thresholds	User engagement optimization

data. This comprehensive approach enables the development of digital assistance systems that can effectively anticipate and respond to user needs while maintaining appropriate boundaries and protections. The ongoing development of these systems presents several important research directions, particularly in the refinement of evaluation mechanisms and the enhancement of proactive response capabilities. Future work must address challenges in balancing system proactivity with user privacy and autonomy, while also developing mechanisms for understanding and responding to complex user needs and preferences.

3.5 Temporal Dynamics and Adaptive Learning

The Digital Brain's effectiveness fundamentally relies on its ability to understand and adapt to temporal patterns in user behavior and interests. This temporal awareness manifests across multiple time scales, from immediate interaction processing to long-term pattern recognition and knowledge evolution. The system implements time-dependent mechanisms that govern how topics and their connections evolve within the knowledge graph, ensuring that the digital representation remains current and relevant to the user's evolving interests and needs. The framework employs dynamic utility assessment for connections between topics, implementing continuous adjustment based on observed interaction patterns. When users demonstrate reduced engagement with particular topics, such as mentioning them less frequently in conversations or showing decreased interaction with related applications, the associated utility values undergo systematic decay. Conversely, topics that maintain a consistent presence in users' daily interactions see their utility values strengthen over time. This dynamic adjustment ensures that the knowledge graph maintains an accurate representation of current user interests and priorities while gracefully adapting to changing patterns of engagement. A crucial aspect of this temporal processing lies in the framework's ability to identify and respond to emerging patterns of interest. The system maintains continuous monitoring of topic utilities and interaction frequencies, enabling the detection of significant shifts in user behaviour or interest. When such shifts are identified, the system initiates appropriate responses through the PDA, ranging from simple information provision to more assistance. These responses might include notifying users of relevant information about topics maintaining high utility values or suggesting new activities aligned with emerging interests. The framework implements particularly

innovative mechanisms for application management based on temporal patterns. The system continuously evaluates the relevance and utility of installed applications, making recommendations for both additions and removals based on observed usage patterns and evolving user interests. New applications, represented as transparent elements in the system visualisation, serve dual purposes: they provide immediate utility to the user while simultaneously contributing to the ongoing refinement of the knowledge graph. This bidirectional relationship between application usage and knowledge representation creates a self-reinforcing cycle of system adaptation and improvement. To maintain effective temporal processing across different time scales, the system implements a multi-level caching mechanism that balances immediate responsiveness with long-term pattern recognition. Recent interactions can be maintained in a high-speed cache for immediate context awareness. At the same time, longer-term patterns are processed through more analysis pipelines that identify emerging trends and evolving interests. This hierarchical approach enables the system to maintain responsive interaction while building a deeper understanding of user behaviour patterns over time. The temporal dynamics of the system extend beyond simple pattern recognition to include predictive modelling of user interests and needs. By analysing the evolution of topic utilities and interaction patterns over time, the system develops increasing models of user behaviour. These models enable proactive suggestion generation, allowing the system to anticipate and prepare for likely user needs before they are explicitly expressed. The continuous refinement of these predictive models through feedback and observation enables increasingly accurate and relevant proactive assistance. The system’s ability to balance immediate responsiveness with long-term adaptation creates a foundation for personalised digital assistance. By maintaining awareness of both immediate context and historical patterns, the Digital Brain can deliver assistance that reflects both current user needs and established preferences. This temporal awareness enables the development of interaction patterns that adapt naturally to evolving user interests while maintaining consistent and reliable assistance across varying time scales.

3.6 Preliminary Evaluation

To validate the proposed framework and demonstrate its practical capabilities, we conducted a detailed evaluation using dialogue analysis from a well-known television series. This evaluation approach allows us to examine the framework’s performance in processing and understanding complex conversational interactions while maintaining controlled experimental conditions. Our evaluation focuses on analysing dialogues from “The Big Bang Theory” television series, specifically examining the character of Sheldon Cooper. This choice of subject provides several advantages for preliminary evaluation. First, the character exhibits consistent behavioural patterns and clear-cut areas of expertise, making it possible to verify the accuracy of our knowledge extraction and representation mecha-

nisms. Second, the availability of complete dialogue transcripts enables a thorough analysis of conversational patterns and topic relationships. For this evaluation, we selected the third scene of the first episode, which presents a complex social interaction involving multiple characters. The scene centres on Penny’s exploration of Leonard and Sheldon’s apartment, featuring interactions demonstrating various aspects of Sheldon’s personality and expertise, particularly when discussing his scientific work. While the scene involves multiple characters, our analysis focuses specifically on Sheldon’s dialogue, creating a more challenging test of our system’s ability to extract meaningful information from a partial conversational context. The evaluation implementation follows our framework’s processing pipeline, beginning with the conversion of textual data into vector representations. We employed the all-MiniLM-L6-v2² model to generate 384-dimensional embeddings for each dialogue segment. This embedding process captures the semantic meaning of conversations while enabling efficient storage and retrieval. These embeddings serve a crucial role in enabling future Retrieval-Augmented Generation (RAG) operations, allowing the system to maintain contextual awareness across multiple interactions. The Cumulative User’s Knowledge module implements a two-stage processing approach. The first stage employs *Llama 3.1 8B*, a LLM selected explicitly for its semantic analysis capabilities. This model performs deep analysis of the conversational content, identifying and extracting key entities, relationships, and attributes from the dialogue. The model’s ability to understand context and nuance proves particularly valuable in processing Sheldon’s often complex and technical statements. The second processing stage involves constructing and maintaining a dynamic knowledge graph structure. This graph evolves continuously as new information is processed, with entities becoming nodes and their relationships forming weighted edges within the structure. The weighting system implements an approach to relationship strength: Figure 3.3 presents the initial knowledge graph constructed from our analysis. The graph reveals several key aspects of the system’s performance in entity and relationship detection. The LLM successfully identifies three distinct categories of entities represented by different colours: persons (red), topics (turquoise), and activities (green). Each relationship carries a specific type indicator, such as “collaborates_with” or “expert_in”, along with an associated confidence score.

The initial analysis examined several key dialogue segments that demonstrate the system’s ability to extract and process complex conversational content. For example, consider the following segment:

- f1: “We can’t have Thai food, we had Indian for lunch.”
- f2: “They’re both curry based cuisines”

These statements, while seemingly straightforward, led to the creation of high-weight connections in the food domain, achieving a confidence score of 0.91. This high confidence

²<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

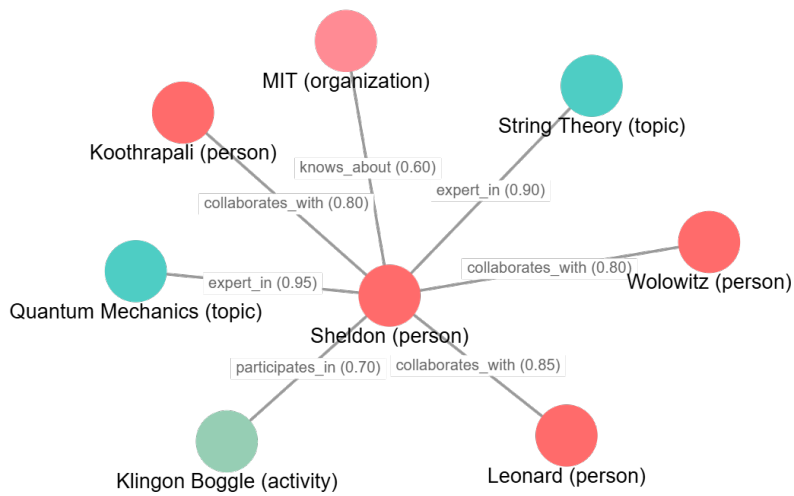


Figure 3.3: Initial knowledge graph constructed from the first analysed scene, showing detected entities (nodes) and their relationships (edges) with confidence scores.

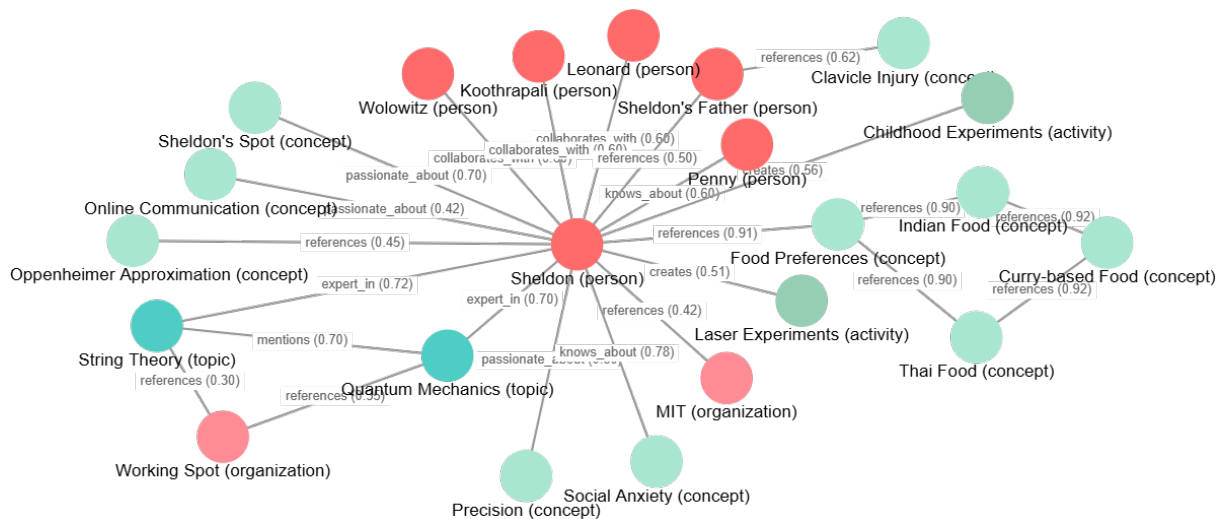


Figure 3.4: Evolved knowledge graph showing the enriched understanding of Sheldon's character after processing additional scenes.

score demonstrates the system’s ability to identify and process thematic relationships even in casual conversation. As more scenes were processed, the knowledge graph evolved to capture increasingly complex relationships and interactions. Figure 3.4 illustrates this evolution, showing the graph’s expansion from 8 initial nodes to 22 nodes after processing the complete first episode. This expanded graph demonstrates several aspects of the system’s knowledge representation capabilities, particularly in capturing nuanced relationships such as those between Sheldon’s workspace and his research interests.

The system shows particular attention to establishing weighted connections between different domains. For instance, it created connections between Sheldon’s workspace and his research areas with varying confidence scores, 0.30 for String Theory and 0.35 for Quantum Mechanics. These differing weights reflect the system’s ability to capture subtle variations in relationship strength based on contextual evidence. The framework’s proactive capabilities became evident when analyzing patterns in the evolved graph. The system demonstrated its ability to make contextually appropriate suggestions based on identified behavioural patterns. For example, when processing the food-related dialogue, the system generated the following suggestion:

- f3: “Given the rejection of curry-based cuisines, might I suggest either sushi or Italian food”

This suggestion’s validity was later confirmed in the episode when another character, Wolowitz, suggested visiting a sushi bar. This alignment between system prediction and actual character behaviour validates the framework’s ability to generate contextually appropriate recommendations. To enhance these proactive capabilities, the system implements RAG mechanisms that integrate historical interaction patterns with similar past scenarios. This integration enables the system to validate its suggestions against previous experiences and outcomes, creating more reliable and contextually appropriate recommendations. The evaluation also examined subjective metrics to assess user satisfaction and emotional response to system suggestions. These assessments employed established metrics including QoE measurements using Absolute Category Rating (ACR) scores and the Self-Assessment Manikin (SAM) scale for evaluating emotional aspects of user interaction. This preliminary evaluation demonstrates the framework’s ability to construct and maintain knowledge representations while generating appropriate proactive suggestions. The system successfully captures both explicit relationships and subtle contextual nuances, creating a foundation for more advanced human-digital twin interactions.

3.7 Privacy and Security Considerations

The implementation of Human Digital Twin systems that process conversational data introduces significant privacy and security challenges that must be carefully addressed. Our

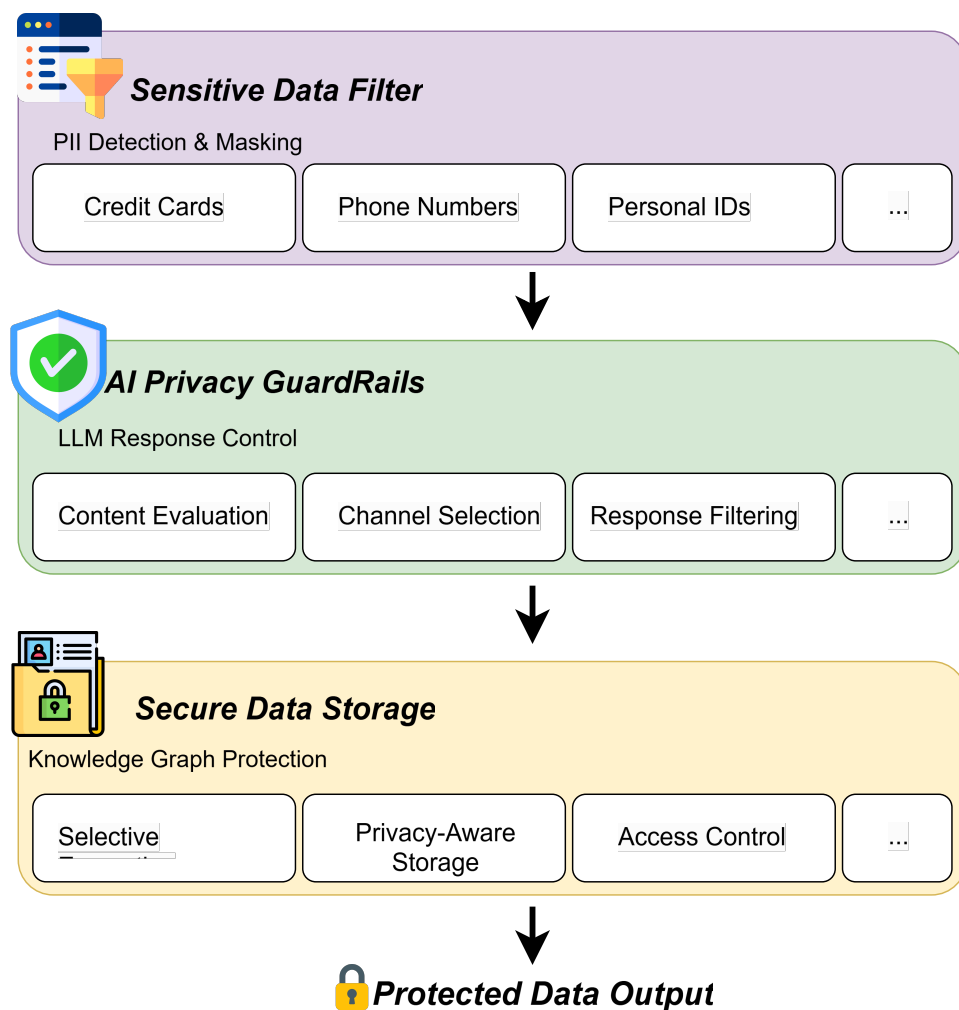


Figure 3.5: Privacy Protection Stages

framework implements comprehensive privacy protection mechanisms that maintain security without compromising system functionality or user experience. Figure 3.5 illustrates the multi-layered approach to privacy protection implemented within the proposed framework, demonstrating how data flows through successive stages of privacy enhancement.

At the foundation of our privacy protection strategy lies the Sensitive Data Filter, represented in the first layer of Figure 3.5. This initial defence mechanism employs pattern recognition and machine learning techniques to identify and protect sensitive information, such as credit card numbers and personal contact details, before it enters the processing pipeline. The visualisation emphasizes three primary categories of sensitive data detection: credit card information, phone numbers, and Personal Identifiable Information (PII). These categories represent the most common forms of explicitly sensitive information encountered in conversational data.

The automated PII detection system utilises advanced tools, including Microsoft Presidio, Amazon Comprehend, and Google Cloud Data Loss Prevention (DLP). When users mention sensitive information during interactions, the system automatically detects

and masks this data, replacing it with anonymous identifiers while preserving the semantic context necessary for meaningful interaction. This preservation of context while removing sensitive content enables the system to maintain a natural conversation flow without compromising privacy.

The second layer of protection, visualised in the middle tier of Figure 3.5, implements AI Privacy Guardrails. This component specifically addresses the challenges posed by Large Language Model responses, which might inadvertently reveal sensitive information through inappropriate channels. The visualisation highlights three critical functions of these guardrails: content evaluation, channel selection, and response filtering. These functions work in concert to ensure that system responses maintain privacy boundaries while delivering relevant information to users. The guardrails evaluate the sensitivity of the response and automatically select appropriate communication channels based on content classification, for example, preventing sensitive information from being shared through insecure channels like email or SMS. Tools like *Nemo Guardrails* [58]³ and *Llama Guard* [59]⁴ are examples of such guardrails, offering functionalities like content filtering and channel selection to prevent unintended data leakage.

The final layer of protection, shown in the bottom tier of Figure 3.5, focuses on secure storage and knowledge graph protection. This layer implements selective encryption through advanced cryptographic techniques, ensuring that sensitive information remains protected even within the system’s knowledge representations. The visualisation emphasizes three key aspects of this protection: selective encryption, privacy-aware storage, and access control mechanisms. These components work together to create a comprehensive security envelope around stored user information.

Beyond the technical protections visualised in Figure 3.5, the framework’s privacy management system adheres to several key principles. The system minimises data collection, gathering only information necessary for its intended functionality. It maintains contextual privacy assessment capabilities, evaluating privacy implications within the broader context of user interactions. Additionally, the protection mechanisms adapt based on the sensitivity of detected information and the context of its use, providing a dynamic response to varying privacy requirements. User control remains paramount; the framework provides transparent mechanisms for users to understand and control how their information is processed and protected. This transparency, combined with the multi-layered protection visualised in Figure 3.5, creates a comprehensive privacy solution that balances security requirements with system functionality.

This comprehensive approach, combining multi-layered protection with user-centric principles, ensures that the HDT system can deliver context-aware assistance while maintaining robust protection of user information. The system’s ability to balance these competing

³<https://github.com/NVIDIA/NeMo-Guardrails>

⁴<https://huggingface.co/meta-llama/Llama-Guard-3-8B>

requirements represents a significant advancement in human-digital interaction systems. The successful implementation of these privacy mechanisms establishes a foundation for future developments in privacy-aware digital twin systems, where maintaining the balance between functionality and privacy protection remains crucial for practical deployment.

Chapter 4

Topic Modelling

The development of HDT systems requires robust mechanisms for understanding and processing human communication. While previous chapters established the theoretical foundations of digital twin technology and examined the challenges of human-digital interaction, this chapter focuses on a critical enabling capability: the automated understanding and classification of textual content. This capability proves essential for HDT systems that must process and respond to natural language interactions while maintaining computational efficiency.

Our investigation centres on the development of an efficient framework for topic classification that addresses several key challenges in processing human communication. Traditional approaches to text classification often struggle to balance accuracy with computational efficiency, particularly when processing longer documents or handling real-time interaction streams. Additionally, existing systems frequently employ complex architectures that, while capable of achieving high accuracy, prove impractical for deployment in resource-constrained environments.

The framework presented in this chapter introduces innovations designed to address these limitations. At its core lies a novel approach to content selection that enables efficient processing of documents while preserving essential semantic information. This selective processing strategy, combined with neural architectures for feature extraction and classification, creates a system capable of maintaining high accuracy while significantly reducing computational overhead. This framework is a fundamental building block for the broader HDT framework discussed throughout this thesis.

Our work makes several specific contributions to the field of text classification and its application in HDT systems. First, we introduce a Key Sentence Extractor (KSE) algorithm that intelligently identifies and preserves semantically significant content, enabling efficient document processing without sacrificing classification accuracy. Second, we develop an attention-based neural architecture specifically optimised for processing selected content, demonstrating superior performance compared to both traditional classification approaches

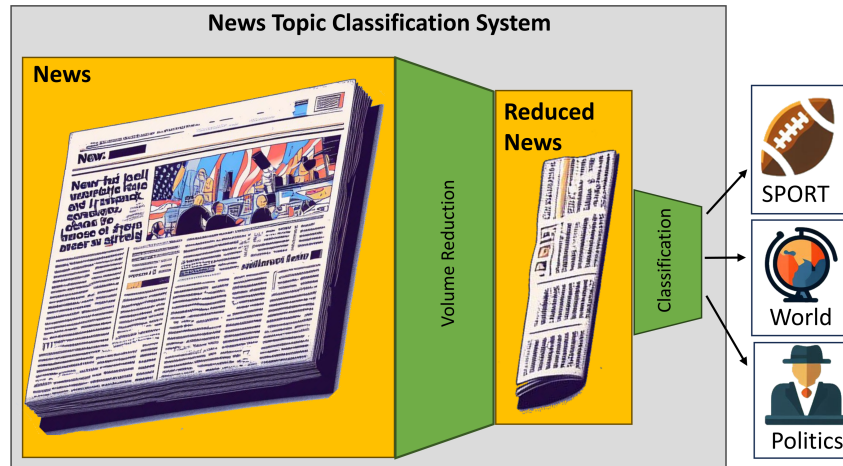


Figure 4.1: News Topic Classification System Overview

and modern LLM. Finally, we provide comprehensive empirical evaluation across multiple datasets, establishing the framework’s effectiveness in real-world classification tasks.

By enabling efficient processing and understanding of textual content, this framework supports the development of digital twin systems capable of maintaining natural and responsive interaction with human users, ultimately contributing to the realisation of the broader HDT vision. The following sections detail our technical approach, empirical evaluation, and the implications of our findings for HDT development.

4.1 Aims and Scope of the Research

This research addresses fundamental challenges in topic classification and text processing, particularly in developing efficient and accurate systems for news article classification. Our work encompasses four primary research objectives, each contributing to advancing text classification capabilities in the context of Human Digital Twin systems. Figure 4.1 depicts the research objective.

The first significant objective focuses on developing a topic classification framework specifically designed for news article processing. This framework integrates established neural network architectures, particularly Bidirectional Long Short Term Memory (BiLSTM) networks, with attention mechanisms to create a more robust classification system. The BiLSTM component captures complex sequential dependencies in both forward and backward directions through the text, while the attention mechanism allows the system to identify and focus on the most relevant portions of input documents. News articles often contain complex temporal and contextual relationships that simple classification architectures might miss. The bidirectional processing capability of our framework ensures a comprehensive capture of these relationships, while the attention mechanism enables adaptive focus on the most informative sections, improving classification accuracy while

maintaining processing efficiency.

Our second objective addresses the crucial challenge of computational efficiency in text processing systems by introducing a novel methodology for reducing data volume through the KSE. This approach represents a significant advancement over traditional document processing methods that typically process entire documents without consideration for the varying importance of different sections. The key sentence extraction algorithm implements content selection mechanisms that identify and preserve the most semantically significant portions of input documents. This selective processing approach serves dual purposes: reducing computational overhead by processing less data and potentially improving classification accuracy by focusing on the most relevant content. The algorithm employs a designed heuristics that consider both local sentence characteristics and global document context when selecting key content.

The third objective encompasses an evaluation of our framework's performance against existing classification systems. This evaluation employed two widely used datasets in the literature, enabling direct comparison with six different classification models. The comparative analysis examined multiple performance metrics including accuracy, precision, recall, and computational efficiency. Our evaluation methodology goes beyond simple accuracy measurements to consider practical deployment considerations such as processing time and resource utilisation. This comprehensive evaluation approach provides valuable insights into the framework's suitability for real-world applications, particularly in resource-constrained environments where efficiency becomes crucial. The final objective addresses a fundamental limitation in current text classification research by introducing the Global News 60K (GN60K) dataset, comprising 60,000 news articles collected from various international sources. This dataset represents a significant contribution to the research community, offering several key advantages over existing collections. The geographic diversity of articles ensures exposure to varied writing styles and cultural perspectives, while the inclusion of ten distinct topic categories provides comprehensive coverage of news domains while maintaining sufficient examples per category for effective training. The international nature of the sources contributes to a diverse vocabulary, reducing the risk of overfitting specific regional or cultural-linguistic patterns. Moreover, careful curation ensures relatively balanced representation across topics, supporting more reliable model training and evaluation.

These research objectives collectively address crucial challenges in text classification while advancing the state of the art in news article processing. The resulting framework and dataset contributions provide essential capabilities for developing HDT systems that can effectively process and understand news content. The following sections detail our technical approach to achieving these objectives and present comprehensive evaluation results demonstrating the effectiveness of our solutions.

4.2 Proposed Solution

In this section, we first mathematically define the problem and then provide an overview of the processes to reach the results presented in Section VI. The process has been split into three distinct phases: Key Sentence Extractor (KSE), word embedding, and finally, the topic classifier.

4.2.1 Problem Definition

The automatic classification branch encompasses various forms of data, including video, audio [60], images [61], and text. Among them, the proposed work focuses on classifying textual data, specifically news articles on the web. In our modelling, each document d_i is decomposed into sentences, defined as a set $\mathcal{S}_i = \{s_{i1}, s_{i2}, \dots, s_{iy}, \dots, s_{iM_i}\}$, where M_i represents the number of sentences in the i -th document. Each sentence s_{iy} consists of a set $\mathcal{W}_{iy} = \{w_{iy1}, \dots, w_{iyz}, \dots, w_{iyZ_{iy}}\}$ of Z_{iy} words, where the words belonging to the document can be represented by the set $\mathcal{W}_i = \bigcup_{y=1}^{M_i} \mathcal{W}_{iy}$. Additionally, we define the set of keywords in the sentence s_{iy} as the set $\mathcal{K}_{iy} = \{k_{iy1}, \dots, k_{iyq}, \dots, k_{iyQ_{iy}}\}$, where for each document d_i , and sentence s_{iy} , the total number of keywords Q_{iy} is at most equal to the number of words Z_{iy} , i.e. the condition $Q_{iy} \leq Z_{iy}$ is always satisfied. It is worth noting that the strict condition is necessary to ensure that the number of selected keywords does not exceed the number of words in the sentence while allowing for the possibility that all words in a sentence could be considered keywords.

The research aims to address two problems: assigning a topic to an unlabelled text corpus and removing irrelevant information from the text. Regarding the first problem, the potential range of topics for news articles is unlimited; our supervised approach considers a limited set of J topics $\mathcal{Y} = \{y_1, y_2, \dots, y_j, \dots, y_J\}$, with each document being assigned to a single topic.

Problem 1 - Topic Assignment Given a generic document d_i , we aim to retrieve the topic $y_j \in \mathcal{Y}$ that maximises the likelihood $P(y_j|d_i)$.

$$\forall d_i \exists! \mathbf{y}_j \in \mathcal{Y} : \max_{y_j \in \mathcal{Y}} P(y_j|d_i)$$

The second problem involves synthesising a document into a shorter version to optimise resource consumption during analysis. This effort becomes particularly significant for medium and long texts, as their complexity makes them more resource-intensive to process. Therefore, distilling the document by removing irrelevant information and retaining only the key concepts becomes crucial to clearly understanding the main topic while minimising resource utilisation.

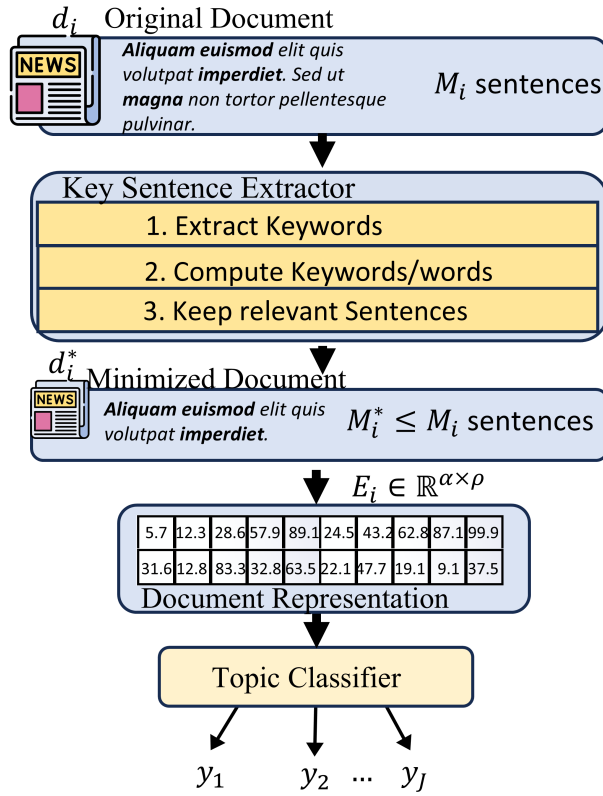


Figure 4.2: System Architecture Flow

Problem 2 - Document Summarisation Given a document d_i belonging to the topic y_j , with a set of sentences $\mathcal{S}_i$, we aim to retrieve a subset $\mathcal{S}_i^* \subseteq \mathcal{S}_i$, such that $M_i^* \leq M_i$, and d_i still belongs to the topic y_j .

Figure 4.2 describes the different phases of the system, which are explained in detail in the following sections. Before delving into each component, Table 4.1 summarizes the main parameters of the proposed framework, from document-processing-related parameters to classification granularity.

4.2.2 Key Sentence Extractor

The Key Sentence Extractor (KSE) represents an approach for distilling text documents into their most semantically meaningful components [62]. This process systematically identifies and extracts the most informative sentences while preserving the document's core meaning and topical relevance. The KSE methodology serves two fundamental purposes in our framework. First, it substantially reduces the computational complexity of the learning process by decreasing the number of features that need to be processed. This reduction directly addresses critical computational constraints such as memory utilization and processing time, making the system more efficient and scalable. Second, it enhances classification accuracy by eliminating potentially misleading sentences - those with minimal

Table 4.1: Main Parameters of the Proposed Framework

Parameter	Description
α	Maximum number of tokens that can be processed by the framework for each document.
ρ	Embedding dimension size that defines the vector representation of each word
λ_{TH}	Threshold parameter in $[0,1]$ used by KSE to determine which sentences to keep based on their keyword density.
J	Number of output topics for classification
M_i	Number of sentences in the i -th document before KSE processing
M_i^*	Number of sentences retained after KSE processing, where $M_i^* \leq M_i$
Z_{iy}	Number of words in the y -th sentence of the i -th document
Q_{iy}	Number of keywords in the y -th sentence of the i -th document, where $Q_{iy} \leq Z_{iy}$

keyword content that contribute little to topic determination and might introduce noise into the classification process. To achieve these objectives, the KSE algorithm processes an input document d_i and produces a refined subset of sentences $\mathcal{S}_i$, where $M_i^* \leq M_i$, with M_i^* and M_i representing the number of sentences in the processed and original documents respectively. The algorithm determines sentence importance through keyword analysis, beginning with the extraction of a keyword set $\mathcal{K}_i$ from document d_i . For quantitative assessment of sentence importance, we introduce the concept of keyword density (λ_{iy}). This metric is calculated for each sentence as the ratio between the number of keywords (Q_{iy}) and the total word count (Z_{iy}) in that sentence:

$$\lambda_{iy} = \frac{Q_{iy}}{Z_{iy}} \quad (4.1)$$

The keyword density provides a normalized measure of how concentrated the important terms are within each sentence, allowing for objective comparison across sentences of varying lengths. The KSE algorithm, expressed in Algorithm 1, employs a threshold-based selection mechanism, where sentences are retained if their keyword density exceeds a predefined threshold: $\lambda_{iy} \geq \lambda_{TH}$, with $\lambda_{TH} \in [0,1]$. This threshold acts as a tunable parameter that controls the stringency of the selection process. The resulting document d_i^*

contains the selected subset of sentences $\mathcal{S}_i^*$ and their corresponding words $\mathcal{W}_i^*$. To ensure robustness in cases where the threshold might be too stringent, the algorithm implements a fallback mechanism. If the initial filtering results in an empty set ($\mathcal{S}_i^* = \emptyset$), indicating that no sentences met the threshold criterion, the algorithm automatically selects the sentence with the highest keyword density:

$$\mathcal{S}_i^* = \arg \max_y (\lambda_{iy}) \text{ if } \mathcal{S}_i^* = \emptyset \quad (4.2)$$

This fail-safe mechanism guarantees that the algorithm always produces a non-empty output, preserving at minimum the most informative sentence from the original document. Such an approach ensures the system's reliability while maintaining its primary goal of feature reduction and noise elimination.

Algorithm 1 Key Sentence Extractor Algorithm

Require: d_i ▷ i-th document

- 1: Define $\lambda_{TH} \in [0, 1]$, $d_i^* \leftarrow \emptyset$
- 2: Define $\mathcal{K}_i \leftarrow \text{ExtractKeywords}(d_i)$
- 3: Define $\mathcal{S}_i \leftarrow \text{ExtractSentences}(d_i)$
- 4: **for each** $s_{iy} \in \mathcal{S}_i$ **do** Compute λ_{iy}
- 5: **if** $\lambda_{iy} \geq \lambda_{TH}$ **then** $d_i^* \leftarrow [d_i^*, s_{iy}]$
- 6: **end if**
- 7: **end for**
- 8: **if** $d_i^* == \emptyset$ **then** $d_i^* \leftarrow \text{getMaxLambdaSentence}(d_i)$
- 9: **end if**
- 10: **return** d_i^*

4.2.3 Word Embedding

Word embedding represents a sophisticated technique for transforming textual data into a dense numerical representation that captures the semantic relationships between words. This transformation is fundamental to our framework as it enables the processing of natural language by machine learning algorithms while preserving crucial contextual information. The embedding process creates a high-dimensional vector space where words with similar semantic meanings are positioned proximally to each other. This spatial arrangement ensures that semantic relationships between words are quantifiable represented - synonyms and contextually related words cluster together, while words with distinct meanings are separated in the vector space [63]. This property is particularly valuable for topic classification as it allows the system to recognize semantic similarities even when different vocabulary is used to express similar concepts. Given a synthesized document d_i^* containing a set of words $\mathcal{W}_i^*$ (referred to as word tokens), the embedding module transforms these tokens into a matrix representation $E_i \in \mathbb{R}^{|\mathcal{W}_i^*| \times \rho}$. Here, ρ represents the dimensionality of

the embedding space, typically determined through empirical optimization according to the rule of thumb proposed by [64]. This dimensionality plays a crucial role in balancing the model’s expressive power with computational efficiency.

Length Normalization and Padding A significant challenge in processing document collections is handling varying document lengths while maintaining consistent input dimensions for the neural network. To address this, we implement a fixed-length representation strategy using a parameter α that defines the maximum number of tokens our framework processes for any document. This standardization is crucial for efficient batch processing and memory management. The length normalization process depends on the relationship between the document length and α :

- When $|\mathcal{W}_i^*| > \alpha$: The document is truncated to the maximum length. In this case, the prior application of the KSE algorithm helps mitigate information loss by ensuring that only the most relevant sentences are retained, maintaining the relationship $|\mathcal{W}_i^*| \leq |\mathcal{W}_i|$.
- When $|\mathcal{W}_i^*| \leq \alpha$: Zero padding is applied with dimensions $[(\alpha - |\mathcal{W}_i^*|) \times \rho]$ to achieve consistent dimensionality across all documents.

The resulting embedding matrix E_i is structured as a transposed row vector where each component represents an embedded token. As shown in Equation 4.3, each token $\mathbf{e}_{i\sigma}$ is represented as a vector of length ρ , capturing its semantic properties in the embedding space:

$$E_i = \begin{bmatrix} \mathbf{e}_{i1} & \mathbf{e}_{i2} & \dots & \mathbf{e}_{i\sigma} & \dots & \mathbf{e}_{i\alpha} \end{bmatrix}^T \quad (4.3)$$

This matrix representation serves as the foundation for subsequent neural network processing, providing a dense, fixed-dimensional representation of the document’s semantic content while maintaining computational efficiency through careful dimensionality management and padding strategies.

4.2.4 Topic Classifier

Our framework implements a supervised learning approach to assign one of J predefined topics to input documents. The classification architecture leverages deep learning techniques to learn the complex relationships between document representations and their corresponding topics. As illustrated in Figure 4.3, the classifier consists of three primary components arranged sequentially: a BiLSTM network, a Self-Attention mechanism, and a fully connected layer.

Bidirectional LSTM Layer The BiLSTM layer processes the input sequence in both forward and backward directions, enabling the model to capture contextual dependencies

from both past and future tokens in the sequence. This bidirectional processing is crucial for understanding complex semantic relationships within the text. Given a document d_i^* with its embedding matrix E_i as input, the BiLSTM computes two hidden state components $\vec{h}_{i\sigma}, \overleftarrow{h}_{i\sigma}$ for each token position σ . These hidden states are computed as follows:

$$\vec{h}_{i\sigma} = f(\vec{\omega}_{xh} \mathbf{e}_{i\sigma} + \vec{\omega}_{hh} \vec{h}_{i\sigma-1} + \vec{b}_h) \quad (4.4)$$

$$\overleftarrow{h}_{i\sigma} = f(\overleftarrow{\omega}_{xh} \mathbf{e}_{i\sigma} + \overleftarrow{\omega}_{hh} \overleftarrow{h}_{i\sigma-1} + \overleftarrow{b}_h) \quad (4.5)$$

where:

- ω_{xh} represents the weight matrix connecting input and hidden layers
- ω_{hh} denotes the weight matrix between consecutive hidden states
- b_h is the bias vector of the hidden layer
- f is a non-linear activation function

The final output of the BiLSTM layer for each position is obtained by concatenating the forward and backward hidden states: $h_{i\sigma} = [\vec{h}_{i\sigma}, \overleftarrow{h}_{i\sigma}]$.

Self-Attention Mechanism The self-attention layer enhances the model's ability to capture long-range dependencies by allowing it to weigh the relative importance of different tokens in the sequence. This mechanism computes attention weights $\delta_{\sigma\tau}$ that quantify the semantic relationship between pairs of tokens:

$$\delta_{\sigma\tau} = \frac{\exp(h_{i\sigma}^\top h_{i\tau})}{\sum_{\theta=1}^{\alpha} \exp(h_{i\sigma}^\top h_{i\theta})} \quad (4.6)$$

Here, $h_{i\sigma}^\top h_{i\tau}$ represents the dot product between the BiLSTM outputs for tokens σ and τ , measuring their semantic similarity. The denominator normalizes these attention scores using a softmax function, ensuring that the attention weights for each token sum to one. Using these attention weights, the layer computes a context-aware representation for each token:

$$o_{i\sigma} = \sum_{\tau=1}^{\alpha} \delta_{\sigma\tau} h_{i\tau} \quad (4.7)$$

where $o_{i\sigma}$ represents the new token representation that incorporates contextual information from the entire sequence.

4.2.4.1 Fully Connected Layer

The final component of our architecture comprises a dense, fully connected layer that transforms the attention-weighted representations into topic predictions. This layer implements a learnable mapping between the document's semantic space and the target topic

space through trainable weights and biases. The architecture of this layer is specifically designed to handle the high-dimensional input features while producing discriminative topic classifications.

Given the output from the attention layer $O_i \in \mathbb{R}^{d \times h}$, where d represents the sequence length and h the hidden dimension, the fully connected layer first applies a linear transformation:

$$z_i = WO_i + b$$

where $W \in \mathbb{R}^{J \times h}$ represents the weight matrix, $b \in \mathbb{R}^J$ the bias vector, and J denotes the number of possible topics. This transformation projects the high-dimensional document representation into a lower-dimensional space corresponding to the number of target topics.

The resulting logits z_i are then processed through a softmax activation function to generate a probability distribution over the possible topics. The softmax function normalizes the network outputs, ensuring they sum to one and can be interpreted as classification probabilities:

$$P(y_j|d_i^*) = \frac{\exp(z_{ij})}{\sum_{k=1}^J \exp(z_{ik})}$$

where $P(y_j|d_i^*)$ represents the probability that document d_i^* belongs to topic j . This normalization is crucial for maintaining numerical stability during training and providing interpretable classification scores during inference.

The final topic assignment for document d_i^* is determined by selecting the class with the highest probability:

$$\text{topic}(d_i^*) = \arg \max_{j \in \{1, \dots, J\}} P(y_j|d_i^*)$$

This architecture enables the network to learn complex, non-linear decision boundaries in the semantic space while maintaining computational efficiency. The combination of the fully connected layer with softmax activation proves particularly effective for topic classification tasks, as it can capture subtle distinctions between related topics while providing clear, probabilistic predictions. During training, the weights and biases of this layer are optimised through backpropagation, allowing the network to discover the most discriminative features for topic classification automatically.

This architecture's combination of bidirectional processing, attention-based context modelling, and non-linear classification enables effective topic identification across diverse document types and domains.

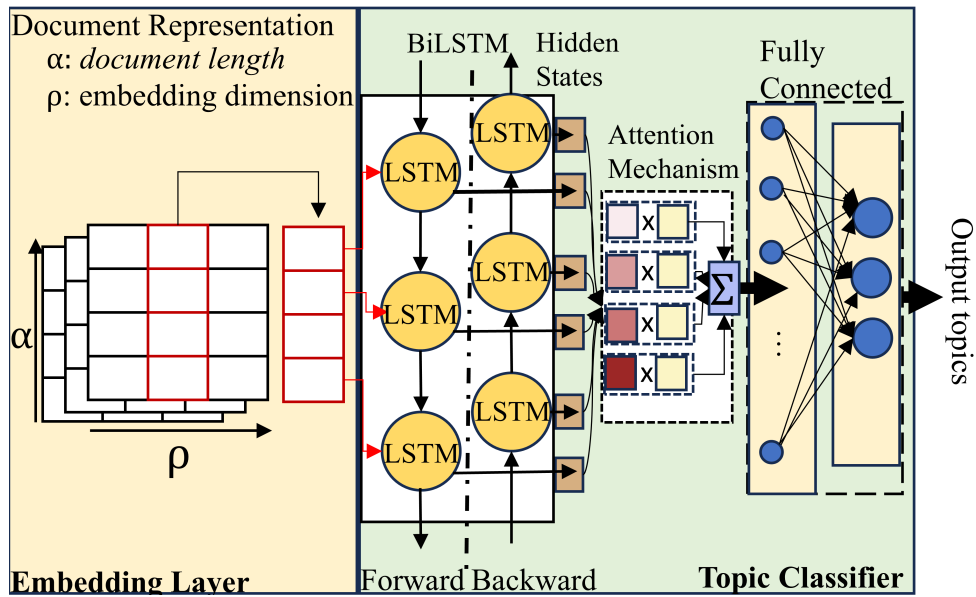


Figure 4.3: Proposed architecture with BiLSTM, Attention Mechanism, and a fully connected layer.

4.3 Experimental Setup

Our experimental evaluation framework was designed to provide a comprehensive assessment of the proposed topic classification system. This evaluation strategy encompassed both the use of established benchmark datasets and the creation of a novel, more diverse dataset to address existing limitations in the field. We begin by outlining our general experimental methodology and then detail the creation and characteristics of our new dataset. The core of our experimental evaluation relied on datasets that have become standard benchmarks in topic classification research. These established datasets, including BBC News and AG News, provide valuable baseline comparisons and enable direct performance evaluation against existing approaches. Each dataset brings unique characteristics and challenges, allowing us to assess our system's performance across different writing styles, article lengths, and topic distributions.

4.3.1 Global News 60K: A Novel Dataset for Topic Classification

The effectiveness of machine learning models for topic classification heavily depends on the quality and diversity of training data. Despite the growing number of text classification datasets, current research faces significant challenges due to limited dataset availability and diversity. Most existing datasets suffer from several limitations: they often cover only a narrow range of topics, draw from a limited number of news sources, and frequently exhibit geographical or temporal biases. These constraints can lead to models that perform well on specific datasets but fail to generalize effectively to real-world applications. The development

of the dataset, GN60K, was guided by the need to address these limitations. Rather than focusing on a single region or cultural context, we deliberately sampled news articles from multiple continents, including North America, Europe, and Australia. This geographical diversity ensures exposure to different writing styles, cultural perspectives, and journalistic approaches. Furthermore, while many existing datasets contain outdated content, we focused on collecting contemporary articles from 2022-2023, ensuring the inclusion of modern terminology and current events, including coverage of global phenomena like the COOronaVirus Disease 2019 (COVID-19) pandemic.

4.3.1.1 Dataset Creation and Methodology

Our dataset creation process began with topic selection informed by the comprehensive UK media consumption study presented in [65], which identified 12 primary topics of reader interest. We refined this list by eliminating geographically-bound topics to ensure global relevance and separating science and technology into distinct categories, recognizing their unique characteristics. This refinement process resulted in 10 final categories optimized for classification tasks.

Using web scraping tools, we gathered 60,000 articles, with approximately 6,000 articles per topic category to ensure balanced representation. The collection process included rigorous duplicate detection and removal, content quality verification, and metadata validation for accurate source attribution and dating. As shown in Table 4.2, the dataset incorporates articles from a diverse set of news sources across multiple countries. For instance, Breitbart.com provides extensive coverage of politics, sports, business, and technology topics from a US perspective, while Bristolpost.co.uk contributes articles focused on entertainment, health, education, and security from a UK viewpoint. CNET offers technology and political coverage across multiple regions including the USA, UK, and Australia. This geographical and topical diversity in the source distribution helps ensure comprehensive coverage and varied perspectives across all categories. A key strength of GN60K is its balanced nature, with a variation coefficient of 0.17 between majority and minority classes. This low disparity indicates a near-uniform distribution of articles across topics, minimizing class imbalance issues for model training. The dataset has been made publicly available through IEEE DataPort¹ to facilitate reproducible research in topic classification and the development of robust, generalizable models. Through its carefully considered design and comprehensive coverage, GN60K represents a significant contribution to the field of topic classification, providing researchers with a robust tool for developing and evaluating classification models that can generalize effectively to real-world applications.

¹<https://ieee-dataport.org/documents/global-news-60k>

Table 4.2: List of sources employed to construct the dataset

Name	Country	Topic
Breitbart.com	USA	Politics, Sports, Business, Economy, Tech, Entertainment & celebrity
Bristolpost.co.uk	UK	Entertainment & Celebrity, Health & Education, Crime & security
Cnet.com	USA, UK, AUS	Politics, Tech
Csmonitor.com	USA	Science, Arts & Culture
Dailycoller.com	USA	Business & Economy, Entertainment & Celebrity, Health & Education, Sports, Politics
Mirror.co.uk	UK	Crime & Security, Weird

4.3.2 Dataset Analysis

The evaluation of our proposed solution required careful selection and analysis of benchmark datasets commonly used in topic classification research. Our analysis encompassed both traditional metrics of dataset size and structure, as well as deeper analytical measures of dataset complexity and challenge level. This comprehensive evaluation approach enables us to understand not only the basic characteristics of each dataset but also their inherent difficulties and potential biases. The AG News dataset [66] represents one of the largest and most widely used benchmarks in the field. Comprising 127,568 news articles (120,000 for training and 7,568 for testing), this dataset offers a substantial volume of data for robust model evaluation. The dataset is structured in a CSV format with three key attributes: title, description (containing the main news content), and topic classification. The four main categories, i.e., world, sports, business, and science-technology, provide a test of classification capabilities across broad news domains. Despite its relatively small number of categories, our analysis revealed that AG News presents substantial complexity, with an interclass similarity score of 0.539 and a Flesch-Kincaid readability score of 15.06, indicating sophisticated language use and significant semantic overlap between categories. The BBC News dataset [67], while smaller in size with 2,225 articles, offers a different kind of challenge through its five distinct categories: business, entertainment, politics, sport, and tech. This dataset exhibits an interclass similarity of 0.466 and a Flesch-Kincaid score of 11.30, suggesting moderately complex language and reasonable category separation. The smaller size but higher quality of articles in this dataset makes it particularly valuable for evaluating model generalization capabilities. To provide context for the complexity of these datasets, we conducted a comparative analysis including other significant datasets in the field, specifically Reuters and 20NewsGroups. This analysis revealed interesting patterns in dataset complexity that challenge common assumptions about dataset difficulty. Table 4.3 summarizes the key characteristics of all analyzed datasets, including our newly

Table 4.3: Comparative Analysis of Dataset Characteristics

Dataset	Categories	Interclass Similarity	Flesch-Kincaid Score
GN60K	10	0.564	12.21
AG News	4	0.539	15.06
BBC News	5	0.466	11.30
20NewsGroups	20	0.259	9.00
Reuters	81	0.143	11.27

introduced GN60K dataset. Our analysis revealed several important insights about dataset complexity. Interclass similarity, measured through TF-IDF vectors and cosine similarity between aggregated category documents, proved to be a crucial metric. Higher values in this metric indicate more challenging classification tasks due to less distinct boundaries between categories. The Flesch-Kincaid readability scores provided another dimension of complexity, measuring text sophistication based on sentence length and word difficulty. Notably, the analysis demonstrated that having more categories does not necessarily correlate with higher classification difficulty. For instance, while Reuters boasts 81 categories, it shows the lowest interclass similarity (0.143), suggesting more distinct category boundaries. Conversely, AG News, with only four categories, presents a higher interclass similarity (0.539), making it a more challenging classification task despite its simpler category structure. Our newly introduced GN60K dataset adds another dimension to this landscape with the highest interclass similarity (0.564) among all datasets analyzed. This high similarity, combined with its intermediate number of categories (10) and substantial Flesch-Kincaid score (12.21), positions it as a particularly challenging benchmark that balances category granularity with classification difficulty. This comprehensive analysis of dataset characteristics has important implications for evaluating topic classification systems. It suggests that traditional metrics like the number of categories or dataset size may not be sufficient indicators of classification difficulty. Instead, measures of semantic overlap and text complexity provide better insights into the real challenges these datasets present to classification systems.

4.3.3 Simulation Setup

The implementation of our framework required careful preparation of the datasets and configuration of various system components. We conducted extensive preprocessing, parameter tuning, and architectural decisions to ensure optimal performance across different types of news articles. Our preprocessing pipeline began with a comprehensive cleaning phase applied to all three datasets. This involved removing special characters, HTML tags, and other artifacts that commonly remain after web scraping procedures. We paid particular attention to maintaining the semantic integrity of the text while eliminating

noise that could interfere with classification accuracy.

A critical aspect of preprocessing involved handling stop words, common words like “the”, “is”, and “at” that carry minimal topical information. These were systematically removed using a curated stop word list, as their presence could dilute the distinctive features needed for accurate topic classification. This step was essential for improving the signal-to-noise ratio in our input data. The morphological processing of words represented another crucial preprocessing step. We implemented both stemming and lemmatization techniques to standardize word forms. Stemming reduces words to their root form through rule-based truncation, while lemmatization employs linguistic knowledge to convert words to their dictionary base form. For example, the term “scientists” was transformed to “scientist”, and various verb forms like “researching”, “researched”, and “researches” were unified to “research”. This normalization process helped reduce vocabulary size and improve the model’s ability to recognize semantically equivalent terms.

For keyword extraction, we employed the YAKE algorithm [68], which has demonstrated robust performance in news article analysis. YAKE identifies keywords using statistical features such as word position, frequency, and co-occurrence patterns, making it particularly effective for news content where important terms often follow predictable patterns of usage [69]. The KSE algorithm’s performance was controlled through the threshold parameter λ_{TH} , which we systematically varied to study its impact on classification accuracy. This parameter essentially determines how aggressive the sentence filtering should be, with higher values leading to more selective retention of sentences. The optimal value of λ_{TH} was determined through extensive experimentation, with results detailed in Section 4.4. For text embedding, we utilized the BERT model [70] with a dimension of $\rho = 768$, following the original BERT paper’s specifications. To manage memory constraints effectively, we implemented a document truncation strategy where the maximum document length α was set to the mean length of articles in the training set. This approach, suggested by Xiong et al. [71], provided a good balance between computational efficiency and information preservation. The architecture of our neural network comprised several carefully designed layers. The embedding matrix E_i for each processed document d_i^* was first passed through a BiLSTM layer, which captured bidirectional contextual dependencies in the text. The BiLSTM’s output was then processed by a dense layer containing 512 neurons, utilizing the Rectified Linear Unit (ReLU) activation function. The final classification layer employed a softmax activation function to generate probability distributions over the possible topics:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^J e^{x_j}} \quad (4.8)$$

where J represents the number of output classes specific to each dataset ($J = 4$ for AGNews, $J = 5$ for BBC News, and $J = 10$ for GN60K). This architecture ensured that the output probabilities summed to 1, providing a well-formed probability distribution

Table 4.4: Hardware and Software Configuration for Experimental Evaluation

Component	Specification	Relevance
CPU	Intel® Core™ i9-7980XE @ 2.60GHz	High core count facilitates parallel data processing and efficient model training
RAM	64GB DDR4	Sufficient capacity for handling large datasets and enabling larger batch sizes
GPU	NVIDIA® GeForce GTX 1080Ti (11GB VRAM)	Provides hardware acceleration for deep learning tasks, including efficient computation of BERT embeddings
Operating System	Ubuntu 20.04 LTS	Offers a reliable and stable environment for running deep learning frameworks
Deep Learning Framework	TensorFlow 2.4	Provides robust tools and libraries for implementing and optimizing neural network architectures
NLP Libraries	Hugging Face Transformers	Simplifies the implementation of state-of-the-art NLP models such as BERT
Programming Language	Python 3.9	Ensures compatibility with modern NLP and deep learning libraries

over possible topics.

The entire system was implemented using modern deep learning frameworks, with careful attention to computational efficiency and numerical stability. We utilized batch processing and GPU acceleration where available to optimize training and inference times while maintaining high accuracy across all datasets.

4.4 Experimental Evaluation

Our experimental evaluation was conducted using the computing environment, detailed in Table 4.4. The hardware configuration was chosen to handle the computational demands of training deep learning models on large text datasets, with the GPU providing essential acceleration for neural network operations. The software stack was carefully selected to ensure compatibility and optimal performance across all components.

4.4.1 Evaluation Metrics

The performance evaluation of our topic classification system requires careful consideration of multiple aspects, as different metrics capture various facets of classification performance. We employ a comprehensive set of evaluation metrics, each providing unique insights into the model's capabilities. The fundamental basis for our evaluation metrics is the confusion matrix, which records the relationships between predicted and actual classifications. For a classification task with J topics, the confusion matrix C is a $J \times J$ matrix where element c_{ij} represents the number of instances from true class i that were predicted as class j . From this matrix, we derive our primary evaluation metrics: **Accuracy** serves as our primary evaluation metric, measuring the overall correctness of classifications:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (4.9)$$

where TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives) are derived from the confusion matrix. In the context of topic classification, accuracy provides an intuitive measure of how often our model correctly identifies the topic of a news article. **Precision** quantifies the trustworthiness of positive predictions for each topic:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4.10)$$

This metric is particularly important in scenarios where false positives are costly, such as when automated topic classification drives content recommendations or feeds into downstream analysis systems. **Recall**, also known as sensitivity, measures the model's ability to identify all relevant instances of each topic:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4.11)$$

High recall is crucial when missing articles of a particular topic would be problematic, such as in comprehensive news aggregation or research applications. **F1-Score** provides a balanced measure combining precision and recall:

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.12)$$

This harmonic mean is particularly valuable when dealing with imbalanced datasets or when we need to consider both precision and recall simultaneously. In our case, it helps evaluate the model's performance across topics with varying frequencies in the dataset. These metrics are calculated both per-class and as weighted averages across all classes, providing a comprehensive view of the model's performance. The per-class metrics help identify any biases or weaknesses in the classification of specific topics, while the averages give an overall performance assessment.

True Class	1	1806	23	40	31
	2	18	1865	8	9
	3	52	13	1694	141
	4	56	18	112	1714
		1	2	3	4
		Predicted Class			

Figure 4.4: Confusion Matrix for AGNews dataset classification. The classes are mapped as follows: 1 - World, 2 - Sports, 3 - Business, 4 - Sci/Tech. The diagonal elements represent correct classifications, while off-diagonal elements indicate misclassifications between categories.

4.4.2 Classification Results

We employed standard metrics, including precision, recall, F1-score, and accuracy, to evaluate the system’s effectiveness in classifying news articles into single topics. Our evaluation methodology was designed to provide comprehensive insights into both overall system performance and category-specific strengths and weaknesses, with particular attention to the impact of the KSE algorithm’s parameter settings on classification outcomes.

Our work primarily focuses on the accuracy metric for several key reasons. First, the prevalence of accuracy measures in related works enables direct benchmarking of our model against existing approaches. Second, accuracy provides an intuitive and clear indicator of the model’s ability to identify topics, particularly valuable in single-label topic classification tasks where each article must be assigned to exactly one category. However, we also analyse precision, recall, and F1-score to provide a more nuanced understanding of classification performance.

For each dataset, we compute these metrics at both the system and category level using the KSE algorithm with optimized λ_{TH} values. The comprehensive classification results for all three datasets are presented in Table 4.5, revealing several interesting patterns and insights about our system’s performance across different contexts. The BBC News dataset evaluation, conducted on a test set of 735 news articles, demonstrates exceptional performance with $\lambda_{TH} = 0.15$, achieving an accuracy of 0.997. This outstanding result is accompanied by remarkably well-balanced precision, recall, and F1-score values across all categories. The Business, Politics, and Entertainment categories achieved perfect

scores across all metrics, suggesting flawless topic separation. The Tech and Sports categories showed marginally lower but still impressive performance (0.984 and 0.990 recall respectively), with the few misclassifications primarily occurring between these technologically and competition-oriented categories. While these results are remarkable, they should be contextualized within the dataset’s characteristics - all articles come from a single source with consistent editorial style and vocabulary usage, potentially simplifying the classification task compared to more diverse sources.

The GN60K dataset evaluation employed an 80/20 train-test split while maintaining the original topic distribution to ensure representative testing. The results demonstrate the robustness of our approach when handling a more complex ten-category classification task, achieving an overall accuracy of 0.945. Science-related articles showed particularly strong performance across all metrics (Precision: 0.986, Recall: 0.981, F1-Score: 0.984), likely benefiting from their distinctive technical vocabulary and well-defined scope. The Health & Education category similarly performed well (Precision: 0.952, Recall: 0.971), suggesting that specialized terminology aids classification accuracy. Interestingly, the Politics category showed lower performance metrics (Recall: 0.838), revealing challenges in political content classification. This lower recall indicates that many political articles were misclassified into other categories, likely due to the inherent overlap between political news and other domains such as Business & Economy (Precision: 0.909) and Crime & Security (Precision: 0.977). This pattern suggests that political discourse often intersects with multiple domains, making clean classification more challenging. The Arts & Culture category achieved surprisingly high metrics (Precision: 0.977, Recall: 0.970), indicating that cultural content maintains distinctive linguistic patterns despite potential thematic overlap with entertainment and news categories. The AGNews dataset evaluation, performed on the standard test set of 7,568 articles, achieved a robust accuracy of 0.9455. Figure 4.4 presents the confusion matrix for this dataset, revealing detailed patterns of classification success and error. The Sports category emerged as the best-performing class (Accuracy: 0.9904, Precision: 0.981, Recall: 0.990), likely due to its distinctive vocabulary and well-defined scope. This exceptional performance suggests that sports journalism maintains consistent linguistic patterns that facilitate accurate classification. The Business category showed comparatively lower performance (Accuracy: 0.9004), with analysis of the confusion matrix revealing frequent misclassifications with World news. This pattern likely stems from the global nature of business reporting, where international business news often overlaps with general world news coverage. The Sci/Tech category’s performance (Accuracy: 0.9109) reveals an interesting phenomenon - the combination of science and technology topics into a single category appears to reduce classification precision. This observation is supported by our GN60K results, where separate Science and Tech categories achieved higher individual performance metrics. These results provide valuable insights into the strengths and limitations of our classification approach. The high accuracy across all datasets (0.997 for BBC News, 0.945

Table 4.5: Classification Results

Dataset	Topic	Metric			
		Precision	Recall	F1	Accuracy
BBC News	Business	1.0	1.0	1.0	1.0
	Tech	1.0	0.984	0.992	0.984
	Politics	1.0	1.0	1.0	1.0
	Sports	1.0	0.99	1.0	1.0
	Entertainment	1.0	1.0	1.0	1.0
GN60K	Arts & Culture	0.9766	0.9702	0.9734	0.9702
	Business & Economy	0.9086	0.9317	0.92	0.9317
	Crime & Security	0.9772	0.9748	0.976	0.9748
	Entertainment	0.9242	0.9361	0.9301	0.9361
	Health & Education	0.9518	0.9705	0.9611	0.9705
	Politics	0.9063	0.8382	0.8709	0.8382
	Science	0.9856	0.9814	0.9835	0.9814
	Sports	0.9328	0.9779	0.9548	0.9779
	Tech	0.9171	0.9345	0.9257	0.9345
	Weird	0.9824	0.9392	0.9603	0.9392
AGNews	World	0.9436	0.9593	0.9514	0.9593
	Sports	0.9807	0.9904	0.9855	0.9904
	Business	0.9225	0.9004	0.9113	0.9004
	Sci/Tech	0.9133	0.9109	0.9121	0.9109

for GN60K, and 0.9455 for AGNews) demonstrates the robustness of our framework across different dataset sizes and complexities. The category-specific analysis reveals how topic distinctiveness, vocabulary overlap, and content intersection influence classification success. Furthermore, the confusion matrix visualisation supports these findings, showing strong diagonal dominance indicating excellent overall classification while highlighting specific areas where topic disambiguation could be improved. The variation in performance across categories and datasets offers important insights for practical applications. Topics with specialized vocabulary (like Sports and Science) consistently achieve higher classification accuracy, while categories with significant content overlap (like Politics and Business) present greater challenges. These patterns suggest that future improvements might focus on better handling of articles that span multiple thematic domains.

4.4.3 Performance Analysis of the Key Sentence Extractor

A critical aspect of our framework’s evaluation focused on analyzing the Key Sentence Extractor (KSE) algorithm’s performance, particularly its dual impact on classification accuracy and computational efficiency.

The investigation revealed complex relationships between sentence selection criteria, system performance, and processing times, providing valuable insights into the algorithm’s behaviour across different types of news articles. To establish a comprehensive under-

standing of the KSE algorithm’s behaviour, we first conducted a detailed analysis of sentence distributions across our datasets. This analysis centred on the λ parameter, which quantifies the density of keywords within each sentence and serves as a proxy for information content. For each news article across our three datasets, we meticulously extracted keywords and computed λ values for every sentence. These distributions were visualised in Figure 4.5 using a histogram with a carefully chosen quantisation step of 0.05. To ensure meaningful comparisons across datasets of varying sizes, we normalised the histogram heights according to the total number of sentences in each dataset. The distribution analysis revealed distinctive patterns across our datasets. The BBC News and GN60K datasets exhibited similar characteristics, with a higher concentration of sentences showing relatively low λ values. This pattern indicates that in these datasets, most sentences contain a relatively small proportion of keywords, suggesting a more elaborate writing style where key information is distributed throughout the text and supplemented with contextual details and supporting information. This writing pattern is typical of traditional news outlets that follow established journalistic practices of providing comprehensive coverage with background information and context. In contrast, the AGNews dataset demonstrated a markedly different distribution pattern. Its λ values peaked at approximately 0.8, indicating that up to 80% of words in these articles were identified as keywords. This high keyword density reflects the more concise, focused nature of these articles, which often prioritise direct information delivery over elaborate context. This characteristic aligns with modern digital news formats that favour brevity and immediate information access.

To evaluate the KSE algorithm’s impact on system performance, we conducted extensive experiments varying the threshold parameter λ_{TH} . The results of these experiments are presented in Figure 4.6, where blue solid lines represent classification accuracy, and dashed lines show training and testing times on a semi-logarithmic scale. For both BBC News and GN60K datasets, we identified an optimal operating range for λ_{TH} between 0.1 and 0.15. Within this range, the system not only maintained but actually exceeded the baseline accuracy achieved without KSE ($\lambda_{TH} = 0$). This improvement demonstrates that the KSE algorithm successfully identifies and removes non-essential sentences while preserving and even highlighting the most relevant content for topic classification. The efficiency gains achieved through KSE implementation were substantial and varied across datasets. With λ_{TH} set within the optimal range, the GN60K dataset demonstrated a 34% reduction in training time and an even more impressive 66% decrease in testing time. The improvements were even more dramatic for the BBC News dataset, achieving an 82% reduction in training time and a 63% decrease in testing time. However, our analysis also revealed important limitations: setting λ_{TH} too high (e.g., 0.25) resulted in excessive information loss, retaining only 36% of content for GN60K and 28% for BBC News, with corresponding decreases in classification accuracy. To ensure robust and fair evaluation

across datasets with varying article lengths, we standardized our testing procedure using a consistent sample size of 2,000 articles from each dataset. The AGNews dataset presented unique challenges due to its inherently shorter articles and high information density. The concentrated nature of its content made it more difficult for the KSE algorithm to effectively discriminate between essential and non-essential sentences, resulting in relatively consistent performance across different λ_{TH} values. This finding suggests that for very short news articles, the benefits of feature space reduction may be limited, and full-text analysis might be more appropriate. A crucial aspect of our evaluation focused on real-world applicability, specifically examining inference time in relation to human reading speed. Recent research indicates that adults read silently at an average rate of 230 words per minute [72]. Our system achieved remarkable performance with an average inference time of just 0.011 ms, substantially faster than human reading times across all document categories: 5.22 seconds for short documents (S, 20 words), 15.65 seconds for medium documents (M, 60 words), and 52.17 seconds for long documents (L, 200 words). This exceptional speed ensures that topic classification can be completed well before a human reader finishes the article, enabling real-time applications and responsive user experiences. The performance analysis of our KSE algorithm demonstrated its effectiveness in balancing accuracy and computational efficiency, particularly for medium and long-form news articles. The algorithm’s ability to significantly reduce processing times while maintaining or even improving classification accuracy represents a valuable contribution to practical news classification systems. Moreover, the detailed characterization of its behaviour across different article lengths and writing styles provides valuable guidance for implementing similar systems in various real-world applications.

4.4.4 Comparative Analysis with State-of-the-Art Methods

To establish the effectiveness of our proposed framework, we conducted a comprehensive comparison with leading approaches in the field of news topic classification. This comparative analysis focused on systems evaluated using the same portions of the BBC News and AGNews datasets, ensuring a fair and meaningful comparison. We emphasize accuracy as our primary comparative metric, as it provides the most direct measure of classification performance and is consistently reported across the literature. Furthermore, as demonstrated in our performance analysis, the training and inference times of our system remain well within practical limits, making accuracy the decisive factor for comparison. Our analysis of systems using the BBC News dataset revealed several strong competing approaches. Kumar et al. [73] implemented a transfer learning approach using RoBERTa, achieving an accuracy of 99.1%. This implementation demonstrated the potential of transfer learning in topic classification, leveraging pre-trained knowledge to achieve high accuracy. Shah et al. [53] explored a different direction, utilizing BERT transformers for English

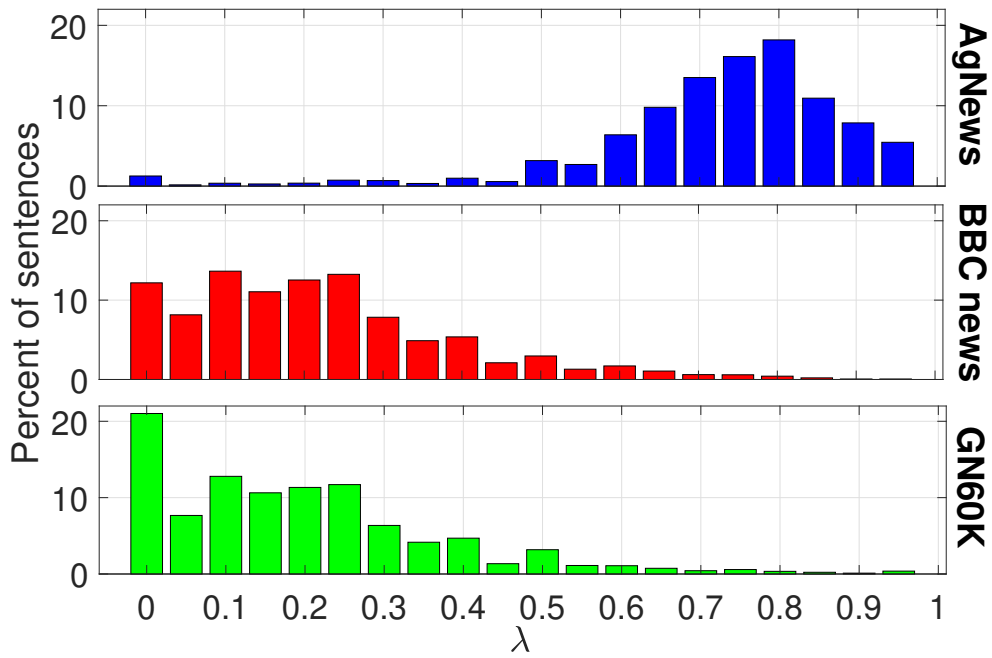


Figure 4.5: Sentence percentage distribution

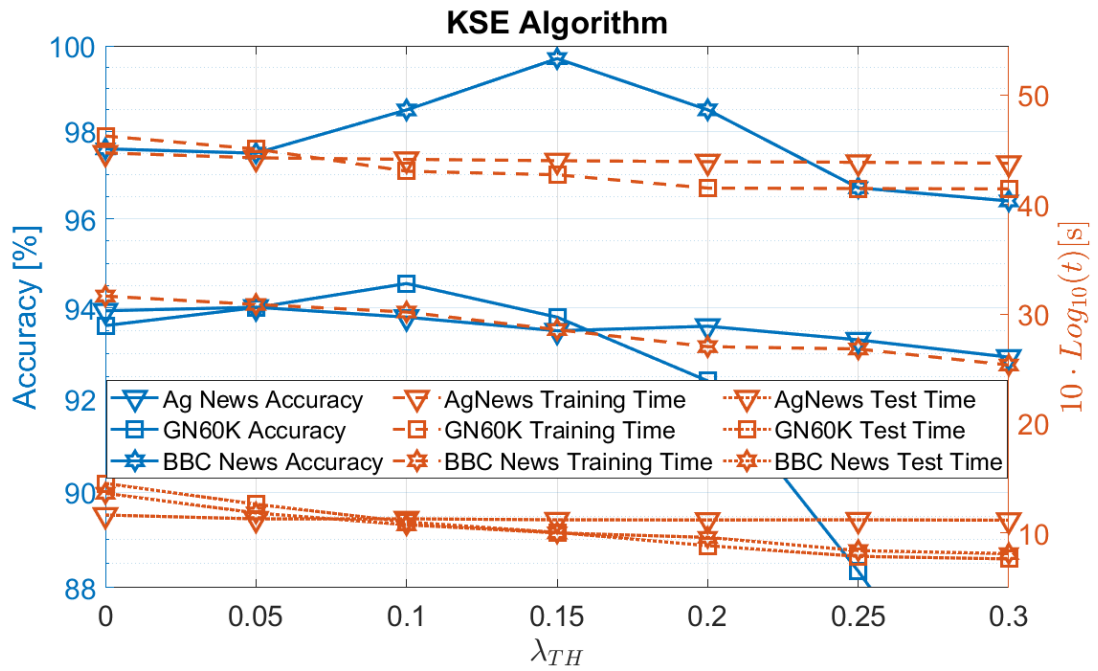


Figure 4.6: Accuracy and Efficiency Trade-off

text classification and achieving an accuracy of 89.1%. While this result highlighted the capabilities of transformer-based architectures, it also suggested room for improvement in their application to topic classification. A particularly notable result came from Kavitha et al. [54], who developed a hybrid CNN-LSTM architecture that achieved an impressive 99.97% accuracy on BBC News. Interestingly, the same architecture showed significantly

lower performance (87.76%) on AGNews, highlighting the challenges of maintaining consistent performance across different dataset characteristics. Our framework achieved an overall accuracy of 99.7% on the BBC News dataset, surpassing most existing approaches. This performance is particularly significant given that our system maintains this high accuracy while incorporating the KSE algorithm for computational efficiency. The ability to maintain such high accuracy while reducing computational overhead demonstrates the effectiveness of our architectural choices and the robustness of our sentence selection mechanism. The evaluation on the AGNews dataset presented different challenges due to its distinct characteristics, particularly the shorter average article length and larger scale. Despite these challenges, our framework demonstrated superior performance compared to existing approaches. Kumar Velu et al. [55] achieved 92.67% accuracy using a supervised approach, while Waly et al. [74] reached 89.91% through various embedding models and classifiers. Our framework surpassed these results with an accuracy of 94.55%, with particularly strong performance in the sports category, reaching 97.2% accuracy. These comparative results, summarized in Table 4.6, demonstrate the robust performance of our framework across different dataset characteristics. The table provides a clear visualisation of our framework’s competitive edge, showing consistent superior performance across both datasets. This consistency is particularly noteworthy given the distinct challenges presented by each dataset: BBC News with its longer, more formally structured articles, and AGNews with its shorter, more concise format. The strong performance across both datasets validates our architectural decisions, particularly the integration of the KSE algorithm with the BiLSTM-attention mechanism. The ability to maintain high accuracy while processing different article lengths and writing styles suggests that our framework has successfully captured fundamental aspects of topic classification that generalize well across different news formats. This generalization capability is crucial for real-world applications, where news articles may come from diverse sources and follow varying stylistic conventions. Furthermore, our framework’s superior performance can be attributed to several key innovations: the adaptive sentence selection mechanism that preserves essential topical information, the effective combination of BiLSTM and attention mechanisms for feature extraction, and the robust classification architecture. These elements work together to create a system that not only achieves high accuracy but also maintains computational efficiency through intelligent content selection. These results suggest that our approach represents a significant advancement in news topic classification, offering both superior accuracy and practical efficiency. The consistency of our performance across different datasets indicates that our framework could be effectively deployed in real-world applications, where handling diverse news formats and maintaining high classification accuracy are crucial requirements.

Table 4.6: Works comparison on BBC News and AGNews

Dataset	Work	Accuracy [%]
BBC News	Shah et al.	89.1
	Kavitha et al.	97.97
	Abhishek et al.	99.1
	Proposed Framework	99.7
AG News	Kavitha et al.	87.76
	Waly et al.	89.91
	Kumar Velu et al.	92.67
	Proposed Framework	94.55

4.4.5 Comparative Analysis with Large Language Models

The rapid advancement of LLMs in NLP has necessitated a thorough comparison between our specialized framework and these powerful general-purpose systems. To provide a comprehensive evaluation, we conducted extensive experiments comparing our framework with both lightweight and state-of-the-art LLMs, focusing specifically on the task of news topic classification. Our comparative analysis began with TinyLlama², a compact yet capable language model that represents an interesting middle ground between computational efficiency and performance. We chose TinyLlama as our initial comparison point because it offers a balanced perspective on what smaller-scale LLMs can achieve in specialized tasks like topic classification. This comparison is particularly relevant for real-world applications where computational resources might be constrained. To ensure a rigorous and fair evaluation, we designed our experiments around a practical use case involving fake news detection and topic-based information retrieval. This scenario requires not only accurate topic classification but also efficient processing to enable real-time response. Both our framework and TinyLlama were evaluated using identical computational resources and testing conditions, eliminating any hardware-related performance variations. The results revealed significant differences in both accuracy and processing efficiency: while our framework maintained its 99.7% accuracy with minimal inference time, TinyLlama achieved an accuracy of 82% with a considerably longer average inference time of 45 seconds per news article when tested on the BBC dataset. This performance gap highlights the advantages of our specialized architecture in handling focused classification tasks. To broaden our understanding of how our framework compares against more powerful language models, we extended our evaluation to include GPT-3.5, one of the most advanced LLMs currently available. This comparison was particularly interesting as it allowed us to assess our framework against a model that represents the current state-of-the-art in general-purpose natural language processing. For this evaluation, we utilized the OpenAI API to access GPT-3.5’s capabilities, focusing specifically on the task of classifying news articles from the BBC News test set into their

²<https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0>

respective topics. The results of this comparison were particularly illuminating. While GPT-3.5 demonstrated impressive capabilities, achieving a 92% accuracy rate in topic classification, our specialized framework still maintained superior performance with its 99.7% accuracy. This performance differential of 7.7 percentage points is especially significant given the scale and sophistication of GPT-3.5. It demonstrates that while LLMs excel in their versatility, specialized architectures can still outperform them in focused tasks. These comparative results offer several important insights into the relative strengths of specialized versus general-purpose architectures. First, they demonstrate that while LLMs possess remarkable general capabilities, there remains substantial value in developing targeted solutions for specific tasks. Our framework's superior performance can be attributed to its specialized design, which incorporates domain-specific knowledge about news article structure and topic classification requirements. This specialization allows for more efficient and accurate processing compared to the broader, more general approach taken by LLMs. Furthermore, the significant difference in inference times between our framework and TinyLlama (0.011 ms versus 45 seconds) highlights another crucial advantage of specialized architectures. In real-world applications, where rapid classification might be necessary for content filtering, recommendation systems, or real-time news aggregation, this performance difference becomes particularly significant. The ability to maintain high accuracy while processing articles orders of magnitude faster than LLMs makes our framework particularly suitable for practical applications. The implications of these results extend beyond mere performance metrics. They suggest that while LLMs represent a remarkable achievement in general-purpose natural language processing, there remains a crucial role for specialized architectures in specific domains. Our framework's superior performance in topic classification demonstrates that targeted solutions can achieve better results while maintaining computational efficiency, especially in scenarios where specific task requirements must be met under real-world constraints. These findings also contribute to the broader discussion about the trade-offs between general-purpose and specialized AI systems. While LLMs offer impressive versatility and can perform reasonably well across a wide range of tasks, our results suggest that for specific applications, carefully designed specialized systems can still offer substantial advantages in both accuracy and efficiency. This insight has important implications for the future development of AI systems, suggesting that a hybrid approach combining both general-purpose and specialized systems might offer the best path forward for many real-world applications. The comparison with LLMs, summarised in table 4.7, underscores the effectiveness of our framework's design choices, particularly the integration of the KSE algorithm with domain-specific architectural elements. These components work together to create a system that not only achieves higher accuracy than state-of-the-art LLMs in topic classification but does so with significantly better computational efficiency. This balance of performance and efficiency makes our framework particularly valuable for practical applications in news classification and content organization systems.

Table 4.7: Performance Comparison between Our Framework and Large Language Models

Model	Accuracy [%]	Inference Time	Model Type
Our Framework	99.7	0.011 ms	Specialized Architecture
GPT-3.5	92.0	N/A*	General Purpose LLM
TinyLlama	82.0	45 s	Lightweight LLM

Inference time for GPT-3.5 is not directly comparable as it runs on OpenAI's servers

All results are obtained on the BBC News test dataset

Inference times measured on identical hardware configuration for Our Framework and TinyLlama

4.4.6 Conclusions and Future Directions

The comprehensive evaluation of our topic classification framework demonstrates several significant advances in processing and understanding textual information. Through careful integration of KSE, word embeddings, and attention-based neural architectures, we have developed a system that not only achieves superior classification accuracy but also maintains computational efficiency across diverse document types. This balance of performance and efficiency establishes a foundation for applications in human-digital interaction systems.

Our experimental results revealed several key insights that inform future developments in text processing and classification. First, the effectiveness of the KSE algorithm in maintaining or improving classification accuracy while substantially reducing computational overhead demonstrates the value of intelligent feature selection in natural language processing. The algorithm's ability to identify and preserve semantically significant content while eliminating non-essential information suggests promising applications in real-time processing of human-digital interactions.

The comparative analysis with LLMs provided particularly valuable insights for future development directions. While our specialized framework demonstrated superior performance in focused classification tasks, the capabilities of LLMs in handling broader contextual understanding suggest opportunities for hybrid approaches. Such systems might combine the efficiency and accuracy of specialized architectures with the versatility and contextual awareness of LLMs.

Several key findings from our evaluation warrant further investigation:

1. The relationship between sentence-level semantic density and classification performance suggests opportunities for content selection mechanisms. Future work might explore adaptive thresholds that respond to document characteristics and classification requirements.
2. The superior performance of our framework compared to general-purpose LLMs in specific classification tasks indicates the continued value of specialised architectures. This finding suggests a development path where targeted solutions complement rather than replace general-purpose language models.

3. The effectiveness of our attention mechanisms in capturing document-level semantic relationships points toward enhanced approaches for understanding conversational context. This capability becomes particularly relevant as we move toward human-digital interaction systems.

These insights establish an important foundation for the development of more advanced HDT systems. The ability to efficiently process and classify textual information while maintaining high accuracy addresses a fundamental requirement for systems that must understand and respond to human communication in real-time. The framework’s success in balancing computational efficiency with classification accuracy suggests approaches for handling the complex, multi-modal data streams characteristic of HDT applications.

As we transition to discussing Human Digital Twins in Chapter 3, the capabilities developed in our topic classification framework become essential building blocks for interaction systems. The ability to efficiently process and understand textual information provides a crucial foundation for HDTs that must interpret and respond to various forms of human communication. This foundation enables the development of systems that can maintain a comprehensive understanding of user context while operating within practical computational constraints.

The transition from focused topic classification to broader human-digital interaction presents several important challenges that we will address in the following chapter:

- Integration of topic classification capabilities with other modalities of human-computer interaction, creating a comprehensive understanding of user context and intent.
- Extension of our efficient processing techniques to handle real-time streams of multi-modal interaction data while maintaining responsive system performance.
- Development of proactive interaction capabilities that leverage our enhanced understanding of textual content to anticipate and respond to user needs.

These challenges frame our discussion of HDTs, where the technical capabilities demonstrated in our topic classification framework enable approaches to human-digital interaction. The following chapter explores how these foundational capabilities contribute to the development of HDT systems that can maintain a comprehensive understanding of user context while delivering personalised and proactive assistance.

Chapter 5

From LLM Reliability to Fake News Detection: A Case Study

5.1 Reliability in Large Language Model

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding (NLU) and generation, but they also exhibit a concerning tendency to “hallucinate”, generating content that appears plausible but is factually incorrect or entirely fabricated. Recent research has systematically categorised these hallucinations into distinct types, including “overfitting”, “logic errors”, “reasoning errors”, and “unfounded fabrication”, highlighting how these models can fail in predictable patterns despite their sophistication. Understanding these patterns of hallucination is crucial not only for improving LLM but also for developing better approaches to detecting misinformation in general. The mechanisms through which LLMs generate false information, maintaining linguistic coherence and surface-level plausibility while diverging from factual truth, mirror patterns found in human-generated disinformation. This parallel suggests valuable insights for developing more robust approaches to detecting unreliable information, regardless of its source.

Drawing from our previous work in topic classification and HDTs, we explore how the study of fallacious reasoning in news articles can inform our understanding of information reliability more broadly. While we do not directly address LLM hallucinations in this work, we present a case study examining how logical fallacies manifest in news content. This analysis provides a framework for understanding how false or misleading information maintains apparent credibility while containing fundamental flaws in reasoning, a pattern common to both machine and human-generated content.

5.2 From Hallucinations to Disinformation: A Framework for Analysis

The parallel between LLM hallucinations and human-generated disinformation extends beyond surface similarities. Both phenomena represent instances where information systems, whether artificial or human, produce content that maintains coherence and plausibility while deviating from truth. This systematic organisation of distorted information suggests common underlying patterns that can be studied and detected. The classification system presented by [75] offers valuable insights into these patterns. Their comprehensive analysis of ChatGPT’s distorted outputs identified eight primary categories of errors, including overfitting, logic errors, and unfounded fabrication. While their work focused specifically on machine-generated content, these categories show remarkable overlap with patterns found in human disinformation:

Logical Failures: Just as LLMs can produce responses that appear reasonable but contain fundamental logical flaws, human-generated disinformation often relies on fallacious reasoning that maintains surface plausibility while failing basic logical tests.

Unfounded Claims: Both systems can generate content that appears authoritative while lacking actual evidential support, relying on the appearance of expertise rather than genuine knowledge.

Context Manipulation: In both cases, information can be technically accurate but presented in misleading contexts that alter its meaning or implications.

These parallels suggest that studying fallacies in news articles can provide insights relevant to both human and machine-generated misinformation. By examining how logical fallacies manifest in news content, we can better understand the mechanisms through which false information maintains apparent credibility despite fundamental flaws in reasoning. This case study focuses specifically on developing a framework for detecting logical fallacies in news articles. While we do not directly apply this framework to LLM hallucinations, the patterns we identify may inform future work on detecting unreliable information across multiple domains. The following sections present our methodology for identifying and categorizing fallacious reasoning, followed by empirical analysis of its effectiveness in detecting problematic content in news media.

The remainder of this chapter is organised as follows. Section 5.3 reviews the current state of the art in disinformation detection and fallacy identification, examining both traditional approaches and emerging work in this field. Section 5.4 introduces our methodology for detecting and analysing fallacies in news articles, building on approaches from both computational linguistics and traditional media analysis. In Section 5.5, we present our

framework for fallacy detection in news content, detailing its components and operational logic. This framework integrates NLP techniques with established logical analysis principles to identify various types of fallacious reasoning. We validate our approach using a real-world news dataset, providing both quantitative performance metrics and qualitative analysis of specific cases.

5.3 State of the Art on Disinformation and Trustworthiness

In recent years, the literature has seen a substantial emphasis on analysing fake news, with numerous studies proposing methods to detect it [76]. The expansion of social media and the vast amount of online information have significantly complicated this task [77]. When disseminating news, individuals often overlook the possibility of encountering fake news and tend to believe only the information that aligns with their pre-existing beliefs. This lack of critical evaluation results in a failure to assess the reliability and truthfulness of the information they encounter on social media platforms [78]. Additionally, the rapid dissemination of fake news, which can spread much faster, deeper, and broader than accurate news, means that a significant portion of the information people come across daily is false [79, 80]. Moreover, while fake news is not a new phenomenon, it is rapidly increasing and attracting more public attention [81]; a primary reason for this is that fake news can be produced more cheaply and quickly than traditional news media [82]. Consequently, fake news detection is becoming an essential mechanism aimed at identifying fake content as swiftly as possible and aiding journalists and fact-checkers [83]. The most important detectors are analysed and categorised below based on their techniques and approaches.

In this context, two prominent fake news detectors, which analyse news features using various machine-learning techniques, are discussed in [84] and [85]. In the first study, the authors employ different machine-learning methods as classifiers for fake news, focusing on linguistic and count-based features such as length and word count. They observe that fake news articles tend to be shorter, more negative, and adopt a more personal tone. In the second study, the authors present a similar approach and demonstrate how utilizing various sources of sentiment, such as images and visual media, can enhance accuracy. This method is evaluated using multiple datasets and similarity techniques. Both studies achieve the best results with the SVM algorithm, which is then applied in the fake news detection processes. Additionally, an approach primarily based on SVM is presented in [86], where the authors introduce a fake news detection model that combines n-gram analysis, a technique used in language modelling and natural language processing, with an Linear SVM (LSVM) classifier. Various sequences of characters or words, known as n-grams, are

generated from a training set and compared to classify fake news from genuine news. All the n-grams are then used as input for a machine-learning technique responsible for the final classification.

Two other approaches that primarily focus on machine learning are detailed in [87] and [88], where the authors employ neural network architectures for detection. In the first paper, the authors concentrate on feature extraction, examining the most relevant attributes of text news. They identify various types of features: content features, such as word count and character frequency, user features based on news readers and specifically on users who have interacted with the news, and social features, which pertain to the social connections of the users. All these features, encompassing both text and news context, are then evaluated and compared using a Convolutional Neural Network (CNN) and an LSTM to classify the news. In the second work, the authors present a classification model for fake news detection based on linguistic features and automatic fact-checking. The model assesses the news by considering linguistic features, such as the number of words and sentences, and then compares them with verified mainstream articles. A deep learning algorithm is trained to recognize common patterns and produce the classification.

Furthermore, two distinct machine-learning techniques are presented in [89] and [90]. In the first approach, the authors introduce a fake news detector based on the analysis of term frequency and unique words. Following this feature extraction process, a Naive Bayes Classifier (NBC) model is trained to classify the news. The system's performance is evaluated based on the precision and accuracy of the classification. In the second approach, the authors investigate the relationship between fake news and clickbait, highlighting how fake news producers aim to profit through clickbait. Clickbait attracts users and piques their curiosity with flashy ads or designed click links to boost revenue. In this context, the authors propose a fake news detection model based on context analysis, such as collecting URLs commonly used for clickbait and linguistic features like the number of capitalized characters or exclamation marks. Additionally, a work based on well-known machine-learning techniques is described in [91]. The authors present an automatic fake news detection system using BERT and ALBERT models, which retrieve the most relevant facts related to news claims and verify their truthfulness through textual comparison. A series of transformer models are employed to compare news with facts retrieved from a manually curated dataset.

The final group of articles explores different approaches that do not rely on machine learning techniques for fake news detection. In this context, two studies are highlighted in [92] and [93], where the authors primarily focus on sentiment analysis. In the first study, the authors propose a framework to promote fact-checked content. They investigate active Twitter users, referred to as guardians, who share validated information to correct fake content in online discussions. The framework provides these guardians with a URL-based fact-checking recommendation model to enhance their engagement and mitigate the

Table 5.1: Analysis of existing fake news detectors.

Ref	Quant.	Informal.	Complex.	Divers.	Fact-Checking	Sentim.	Ads
[84]	✓	✓	-	-	-	✓	-
[85]	-	-	✓	-	-	✓	-
[86]	✓	✓	✓	✓	-	-	-
[87]	✓	-	✓	✓	-	-	-
[88]	✓	-	✓	✓	✓	-	-
[89]	✓	-	✓	✓	-	-	-
[90]	✓	-	-	-	-	-	✓
[91]	✓	-	✓	-	✓	-	-
[92]	-	-	✓	-	✓	-	-
[93]	✓	-	✓	✓	-	✓	-
FAKE	✓	✓	✓	✓	✓	✓	✓

negative impacts of fake news. Initially, the model identifies guardian users, then analyses textual claims, and recommends fact-checking URLs from guardians to other users. In the second study, the authors present a model for detecting fake news using sentiment analysis as the primary feature. This model integrates sentiment related to the text with style features, such as word count and statement frequency, and performs classification using various machine learning algorithms.

In summary, Table 5.1 shows a classification of the fake news detection models based on the analysed approaches. The classification is based on three metrics, i.e., text quality, fact-checking, and context analysis. In general, the features needed to measure the news quality can be divided into four main categories [94]:

- *Quantity*, which considers features such as the number of characters, the number of words and the number of sentences;
- *Informality*, which takes into account the fact that fake news often contains more mistakes than trustworthy ones, and therefore characteristics such as misspellings and typos are used as indicators of the authenticity of the news;
- *Complexity*, which is represented by parameters such as the average word length, the words per sentence and the average ratio of punctuation per sentence. In general, the higher the linguistic complexity of a text, the less likely it is a fake.
- *Diversity*, which considers the percentage of different terms in the text, the occurrence of the words, and their spatial distance; deceptive texts are perceived to be limited in terms of vocabulary usage and usually make use of several redundant terms.

Another classification metric is represented by *Fact-Checking*. Even if the text analysis is largely used for classification, other approaches are used to support it. To verify the

correctness of the information, researchers propose systems to check the credibility of the news. The last metric concerns the analysis of the news context. It classifies news through the analysis of the website in which the news is published with features such as the *Sentiment* related to the news content or the presence of advertisements (*Ads*) as well. The solution proposed in this research takes into consideration all kinds of news topics, and no user information is collected; for this reason, properties related to the news readers are out of the scope of our solution.

5.3.1 Trustworthiness on News Providers

The issue of trust in news providers has been gaining huge popularity in recent years, and the community is trying to find new approaches to study the news providers' trustworthiness [95]. This is due to the fact that non-verified news media have exploited technological development to spread misleading or fraudulent news [96, 97]. In general, trust in news providers is at historically low levels and new mechanisms to measure their trust are essential to overcome this age of uncertainty [98].

Below, we provide a brief background on the most accepted properties used in literature to evaluate the trust of a news provider. One of these is presented in [99], where authors introduce the concept of trustworthiness in news media as a relationship between a trustee, represented by the user or the actor who trusts, and a trustor, i.e., the news source that provides the news and receives the trust. The news media is then evaluated, and the trust is obtained through 4 dimensions: trust in the selectivity of topics, selectivity of facts, the accuracy of depictions, and journalistic assessments. Another two works that analyse the impact of trust in news media are illustrated in [100] and [101]. The first paper proposes various measures of trust at different levels of analysis. The source is evaluated both generally and in each specific topic, taking into account its fairness and accuracy in distinguishing between facts and opinions. Then, in the same way, the paper judges the author of the news, the journalist, and the media outlet, according to their objectivity and subjectivity. In the second paper, the authors examine the news media trust in terms of credibility in a political topic for five different countries. The authors present a regression model that assesses news providers based on various features, including media attributes such as overall credibility and reading frequency, as well as demographic metrics of the news reader.

In these terms, we propose an automatic algorithm to evaluate the trust of news sources based on the most accredited measures of credibility. It makes use of novel parameters, which consider the expertise, the popularity of the news source, and its past behaviour.

5.3.2 State of the Art in Fallacy Classification

Recent research in fallacy classification has seen significant advances through both traditional machine learning approaches and neural architectures. Early work by Habernal et al. [102] established a framework for computational detection of logical fallacies, focusing primarily on argument structure analysis in online discussions. Their work introduced essential distinctions between surface-level language features and deeper argumentative patterns that characterize different types of fallacies. Building on this foundation, [103] developed more sophisticated approaches incorporating transformer-based models for fallacy detection, achieving notable improvements in classification accuracy across multiple fallacy types. Their research highlighted the importance of contextual understanding, demonstrating that fallacy detection requires analysis beyond simple pattern matching. Several key challenges have emerged in the field of automated fallacy detection:

- **Context Dependency:** [104] demonstrated that accurate fallacy classification often requires understanding broader contextual elements beyond the immediate text, particularly for complex fallacies like false equivalence and hasty generalization.
- **Implicit Premises:** Work by Sourati et al. [105], highlighted the challenge of identifying fallacies that rely on unstated assumptions, proposing novel architectures that incorporate background knowledge representation.
- **Cross-Domain Applicability:** [106] addressed the challenge of developing fallacy detection systems that maintain accuracy across different domains and discourse styles.

More recent approaches have focused on incorporating multi-modal analysis. Zanartu et al. [107], introduced a framework combining psychological insights with computational techniques for detecting fallacies in climate misinformation. Their work demonstrated remarkable improvements in fallacy detection, achieving F1 scores 2.5-3.5 times higher than previous approaches. Notably, their analysis revealed that certain fallacy types like fake experts and anecdotal arguments were more readily detectable, while fallacies requiring deeper background knowledge - such as oversimplification and misrepresentation - presented greater detection challenges. The integration of LLMs has opened new avenues for fallacy detection. Lei et al. [108], showed that fine-tuned transformer models can achieve superior performance in identifying subtle logical fallacies, particularly in cases where traditional rule-based systems struggle. However, their work primarily focused on isolated statements rather than analyzing fallacies within the broader context of news articles. While these works have advanced our understanding of fallacy detection, significant gaps remain in applying these techniques to real-world news content. Current approaches lack the ability to both identify and extract fallacious segments from longer news articles while maintaining awareness of the broader narrative context. Additionally, no existing work

has successfully integrated fallacy detection with comprehensive news verification systems. The development of such an integrated approach, capable of both detecting fallacies and evaluating their impact on news credibility, represents a crucial next step in combating misinformation.

5.4 A Framework for Fake News Detection and Trust Management

The increasing sophistication of fake news creation and distribution demands a comprehensive approach that goes beyond simple content analysis. While existing solutions often focus on either content verification or source credibility, there is a clear need for integrated systems that can handle both aspects while maintaining user privacy and system security. In this section, we present a framework that combines fake news detection capabilities with trust management features. The framework is designed to provide users with real-time evaluation of news content while building and maintaining trust profiles of news sources. This dual approach allows for both immediate content assessment and long-term source credibility evaluation. The framework consists of three main components working together through a browser plugin interface, as illustrated in Figure 5.1. At its core lies the fake news detector, responsible for evaluating individual news items through a multi-faceted analysis approach. This component works alongside a trust management model, which maintains and updates reliability scores for news sources. These two components are connected through an encrypted database that serves as both a storage system and a communication channel. The fake news detector represents the system's primary evaluation engine. When a user encounters news content, this component initiates an analysis process that considers multiple aspects of the content, extracting relevant features for classification. Rather than relying on a single evaluation method, the detector implements several complementary approaches to achieve more robust results. The trust management model focuses on the broader context of news sources rather than individual articles. This component maintains historical data about news providers and implements novel parameters for assessing their reliability. By tracking source behaviour over time, the system can identify patterns and changes in content quality, providing users with valuable context before they engage with specific content. The database plays a crucial role in the framework's operation, serving as more than just a storage system. It implements selective encryption through the Blowfish algorithm, allowing for efficient handling of both sensitive and non-sensitive data. This design choice reflects a careful balance between security requirements and system performance. Sensitive feedback data is encrypted to protect against potential breaches, while non-sensitive data remains unencrypted to minimize computational overhead. The interaction between these components is orchestrated through a browser plugin that serves

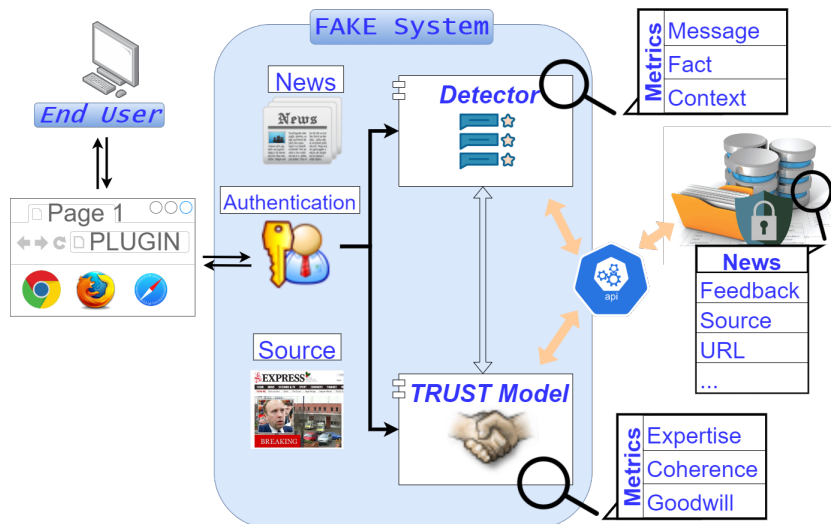
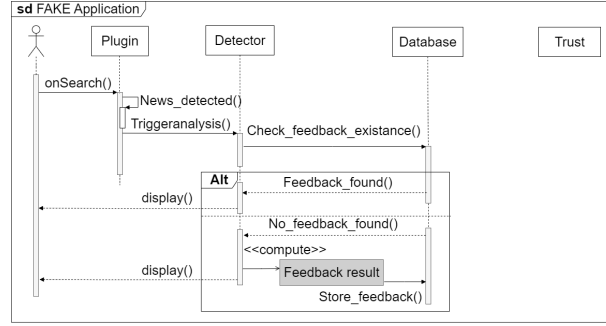


Figure 5.1: Overall system architecture.

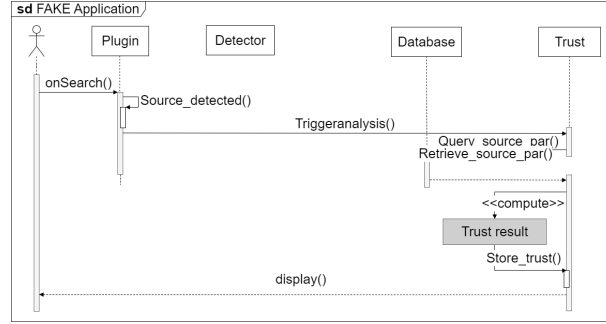
as the user interface. Figure 5.2 illustrates the sequence of operations that occur when users interact with news content. The plugin continuously monitors web pages, activating either the fake news detector or trust management model depending on the user’s browsing behaviour. This design allows for seamless integration into users’ normal browsing habits while maintaining privacy by limiting personal data collection to the minimum necessary for system operation. The framework’s architecture reflects careful consideration of privacy and security requirements. Personal data collection occurs only during the initial plugin installation, and browsing data is collected anonymously. This approach ensures that while the system can effectively evaluate news content and sources, it does so without compromising user privacy or creating potential security vulnerabilities. Through this integrated approach, the framework provides a comprehensive solution to the challenges of fake news detection while maintaining user privacy and system security. The following sections will detail each component’s operation, beginning with the fake news detector’s evaluation methodology.

5.4.1 Feedback Evaluation Model

The assessment of news content requires an evaluation approach that considers multiple aspects of the content and its context. In this section, the feedback evaluation model is presented, which implements a comprehensive assessment strategy for news content selected by users. This model extends beyond simple binary classification by incorporating multiple evaluation factors and producing a continuous score that reflects not just the reliability of the news but also its potential impact and intent. The model evaluates news through three distinct yet complementary factors: message-based, fact-based, and context-based parameters. The message-based factor analyses the stylistic and linguistic characteristics



(a) Fake news detector.



(b) Trust management model.

Figure 5.2: Sequence diagrams for the fake news detector and the trust management model.

of the news text, examining aspects such as quantity (length and word count), informality (presence of errors and informal language), complexity (sentence structure and vocabulary sophistication), and diversity (vocabulary richness and word distribution). This multi-dimensional linguistic analysis helps identify patterns commonly associated with deceptive content. The fact-based factor implements a two-pronged approach to verify the accuracy of news claims. First, it compares news claims against a comprehensive pre-trained model that contains verified information. Second, it analyses similarities with previously evaluated news articles, leveraging the system's growing knowledge base to improve accuracy over time. This dual verification process helps identify both blatant falsehoods and more subtle forms of misinformation. The context-based parameters examine elements beyond the news text itself, considering factors such as the presence and quantity of advertisements on the hosting webpage - a feature often associated with low-credibility sources. Additionally, this factor includes sentiment analysis to evaluate both the emotional tone and objectivity of the content, as extreme or highly subjective language often signals potential bias or manipulation. These three factors are combined into a single feedback score using a weighted sum approach. For a news article n_i , the feedback score f_i is computed as:

$$f_i = \alpha M_i + \beta F_i + (1 - \alpha - \beta) C_i \quad (5.1)$$

where:

- $M_i \in [-1, 1]$ represents the message-based factor contribution
- $F_i \in [-1, 1]$ represents the fact-based factor contribution
- $C_i \in [-1, 1]$ represents the context-based factor contribution
- $\alpha, \beta \in [0, 1]$ are weighting parameters with $\alpha + \beta \leq 1$

The weighting parameters α and β allow the system to adjust the relative importance of each factor based on empirical performance data. The constraint that their sum must be less than or equal to 1 ensures that the resulting feedback score remains within the range $[-1, 1]$, providing a normalized measure of content reliability. The interpretation of the feedback score follows a continuous spectrum from disinformation to reliable information. A score of -1 indicates content deliberately created to cause harm or spread disinformation [109], while a score of 1 represents highly reliable information supported by strong evidence and proper context. The middle range around zero represents misinformation - false or misleading content shared without malicious intent [110]. The model acknowledges two critical transition zones in the feedback spectrum: the boundary between disinformation and misinformation, and between misinformation and reliable information. These zones represent areas where classification becomes more challenging due to the subtle nature of content manipulation. To address this uncertainty, we introduce a threshold parameter TH such that all content with feedback scores satisfying $f_i \leq |TH|$ is classified as misinformation. This threshold helps create clear boundaries between different categories of content while maintaining the nuanced nature of the continuous scoring system. The following subsections detail the specific components and implementation of each factor, describing how they contribute to the overall assessment of news content reliability. This detailed examination will demonstrate how the model captures both obvious and subtle indicators of potentially unreliable content, providing a robust framework for news evaluation.

5.4.1.1 Message-based Analysis of News Content

The message-based component of the evaluation model focuses on the linguistic and stylistic characteristics of news content. This analysis is founded on the premise that deceptive content often exhibits distinct patterns in its textual presentation, from basic structural elements to more complex linguistic features. Our approach implements a comprehensive analysis framework that examines multiple dimensions of text quality to effectively distinguish between authentic and false statements. The analysis process begins with a crucial pre-processing phase. During this stage, the news text undergoes several cleaning operations to standardize the input for our model. These operations include the removal of special characters that might interfere with analysis and the elimination of stopwords, common words that carry little semantic value. This preprocessing ensures that

Table 5.2: Message-based factor features.

Category	Feature	Description
Quantity	Characters	Number of characters in the news text.
	Words	Number of words in the news text.
	Average words	Average number of words per sentence.
	Punctuation	Average punctuation marks with respect to the number of characters.
Informality	Bad words and toxic content	Presence of bad words or abusive language in the message.
	Typos and misspellings	Number of unknown words and their similarity to other words.
Diversity	Redundancy	Occurrence of words and spatial distance between them.
Complexity	Term frequency	Frequency of words in the entire text.

subsequent analysis focuses on the meaningful content of the text. Following preprocessing, the model implements feature extraction across four primary categories of textual analysis, as detailed in Table 5.2. Each category captures different aspects of writing quality and style:

The first category, Quantity, provides foundational metrics about the text’s structure. These features include raw character and word counts, which can reveal patterns in content length typical of certain types of news articles. The average words per sentence metric offers insight into syntactic complexity, while punctuation density can indicate writing style and formality level. Together, these quantitative measures provide a baseline assessment of the text’s structural characteristics. Informality analysis focuses on indicators that might suggest lower editorial standards or deliberately inflammatory content. This category examines two key aspects: the presence of inappropriate or toxic language, and the frequency of typographical errors or misspellings. These features are particularly valuable as legitimate news sources typically maintain high editorial standards, making the presence of such elements a potential indicator of unreliable content. The Diversity category examines the text’s lexical richness through redundancy analysis. This involves studying both the frequency of word repetition and the spatial distribution of repeated terms throughout the text. Low diversity often characterizes content created with less attention to quality or deliberately designed to emphasize certain points through repetition. The final category, Complexity, analyses term frequency patterns across the entire text. This analysis helps identify sophisticated writing patterns typically associated with professional journalism, as opposed to more simplistic or repetitive patterns often found in lower-quality content. After extracting these features, our model employs the Extreme Gradient Boosting (XGBoost) algorithm [111] to compute the message-based value $M_i \in [-1, 1]$. The selection of XGBoost was motivated by several key advantages. Unlike other classification algorithms, XGBoost demonstrates robust performance even with imperfectly curated

datasets, a common scenario in real-world news analysis. Additionally, XGBoost provides valuable feature importance metrics, allowing us to identify which textual characteristics are most indicative of reliable versus unreliable content. The computed message-based value serves as a crucial component in our overall feedback model, providing a quantitative assessment of content quality based purely on textual characteristics. The continuous range of values from -1 to 1 allows for nuanced classification, with negative values indicating characteristics typically associated with unreliable content and positive values suggesting more professional, reliable content.

5.4.1.2 Fact-based Analysis and Verification

The fact-based component of our evaluation model focuses on verifying the accuracy and truthfulness of news content through a dual verification approach. This component combines automated fact-checking against a verified knowledge base with comparative analysis against previously evaluated news articles. The integration of these two methods provides a more robust assessment of content truthfulness than either method could achieve independently. The fact-based analysis produces two distinct but complementary parameters. The first, denoted as $F_{i,f}$, evaluates news content by matching its claims against FEVER, a pre-trained network containing verified claims. The second, represented by $F_{i,c}$, implements a comparative analysis approach, evaluating news content by examining its similarity to previously analysed and verified articles. These parameters work in concert to provide a comprehensive assessment of content accuracy.

Automated Fact Verification Through FEVER The FEVER (Fact Extraction and VERification) system [112] serves as our primary tool for automated fact verification. This pre-trained network, built on a foundation of Wikipedia-sourced claims, provides a sophisticated mechanism for assessing the veracity of news content. FEVER analyses individual claims within news articles and categorizes them into three distinct classifications: supports (indicating verified truthful claims), refutes (signifying demonstrably false claims), and insufficient information (where verification cannot be definitively determined). The implementation of FEVER-based verification begins with content preprocessing, where news articles are decomposed into their constituent claims. For a given news article n_i , we define a set of claims $\mathcal{C}_i = \{c_{p,i}\}$. Each claim $c_{p,i}$ is evaluated against the FEVER knowledge base and assigned a binary state: 1 for claims that align with verified information in the FEVER dataset, and 0 for claims that are either false or cannot be verified. The total number of supported claims V_i is then calculated as:

$$V_i = \sum_{p=1}^{|\mathcal{C}_i|} c_{p,i} \quad (5.2)$$

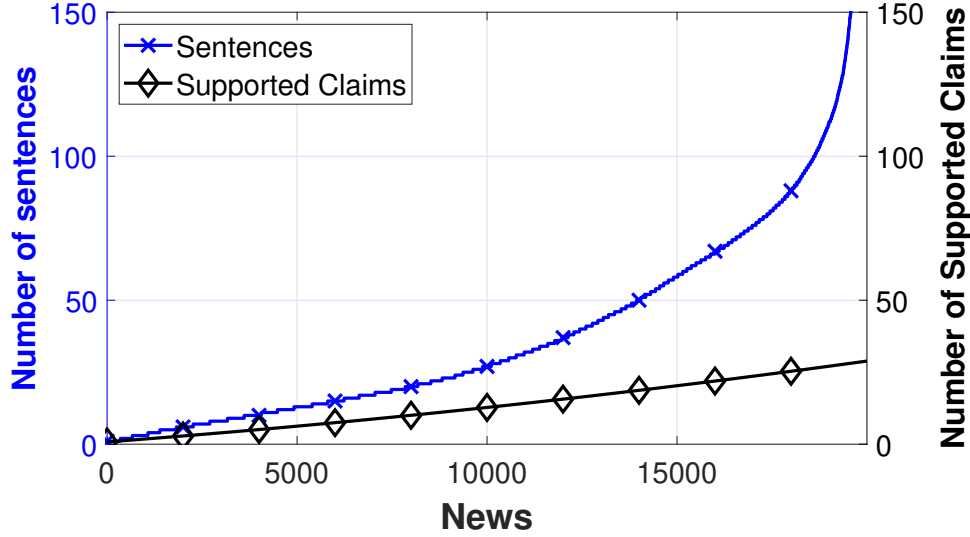


Figure 5.3: Analysis of the number of supported claims and news length for the Kaggle dataset.

An important consideration in this analysis is the relationship between article length and the number of verifiable claims. Naturally, longer articles tend to contain more claims that can be verified, regardless of their overall truthfulness. To account for this length dependency, we normalize the verification score by the article length L_i , computing the ratio V_i/L_i . This normalized score, falling within the interval $[0, 1]$, provides a more balanced indicator of content reliability. As illustrated in Figure 5.3, our analysis of the Kaggle dataset shows that the number of verified claims increases at a significantly slower rate than article length, suggesting that articles with disproportionately high verification ratios may actually indicate potential reliability issues. Based on these observations, we compute the FEVER parameter $F_{i,f}$ as:

$$F_{i,f} = \begin{cases} 0.5 - \frac{V_i}{L_i} & \text{for } V_i > 0 \\ -0.5 & \text{for } V_i = 0 \end{cases} \quad (5.3)$$

The parameter is constrained to the range $[-0.5, 0.5]$ to account for the inherent limitations of the FEVER dataset’s coverage.

Comparative Fact Analysis The comparative fact analysis component complements FEVER’s automated verification by implementing a similarity-based evaluation approach. This method assesses content accuracy by comparing news articles against a database of previously evaluated content, providing an additional verification mechanism particularly valuable when direct fact verification is challenging or incomplete. Our implementation adopts the similarity analysis framework proposed in [113], enhanced with modern text embedding techniques. The process begins with content preprocessing, where articles are

converted into bag-of-words representations using a combination of embedding algorithms including Word2vec and GloVe. These representations enable sophisticated similarity comparisons against previously evaluated content. For a given article n_i , we identify a set of similar articles $\mathcal{D}_i$ defined as: $\mathcal{D}_i = n_j : f_j \geq f_{th} : \& : A_{i,j} > A_{min}$ where f_{th} represents a reliability threshold, $A_{i,j} \in [0, 1]$ quantifies the similarity between articles n_i and n_j , and A_{min} establishes a minimum similarity threshold. This formulation ensures that comparisons are made only against articles that have been previously verified as reliable and demonstrate substantial content similarity. The fact comparison parameter $F_{i,c} \in [-1, 1]$ is computed for articles with identified similar content ($|\mathcal{D}_i| > 0$) as:

$$F_{i,c} = \frac{\sum_{j=1}^{|\mathcal{D}_i|} f_j \cdot A_{i,j}}{\sum_{j=1}^{|\mathcal{D}_i|} A_{i,j}} \quad (5.4)$$

The final fact-based score F_i integrates both verification approaches:

$$F_i = \begin{cases} F_{i,c} & \text{if } |\mathcal{D}_i| > 0 \\ F_{i,f} & \text{if } |\mathcal{D}_i| = 0 \end{cases} \quad (5.5)$$

This integrated approach ensures that fact verification leverages both direct fact-checking and comparative analysis, providing a more comprehensive assessment of content truthfulness. When similar articles are available, the system prioritizes comparative analysis; when no similar content exists, it relies on FEVER’s automated verification capabilities.

5.4.2 Context-based Analysis

The context-based analysis represents a crucial component in our framework for evaluating news credibility. This factor extends beyond the textual content itself, incorporating elements that characterize the broader context in which the news appears. Our approach considers two primary contextual indicators: the presence and distribution of advertisements on the web page ($C_{i,a}$) and the sentiment patterns within the content ($C_{i,s}$). These parameters provide valuable insights into the credibility and professional standards of news presentation.

Advertisement Analysis The presence and distribution of advertisements serve as a significant indicator of news credibility. Research has consistently shown that websites propagating low-credibility news often employ aggressive advertising strategies to maximize revenue through user engagement [114]. This correlation between excessive advertising and questionable content quality provides a valuable metric for assessing news reliability. To quantify this relationship, we have developed a mathematical model that evaluates the advertising patterns on news websites. The advertisement score $C_{i,a}$ is computed using a

piecewise function that accounts for both the number of advertisements and their density relative to content:

$$C_{i,a} = \begin{cases} 1 - 2\frac{N_{i,a}}{N'_{i,a}} & \text{for } N_{i,a} \leq N'_{i,a} \\ -1 & \text{for } N_{i,a} > N'_{i,a} \end{cases} \quad (5.6)$$

In this formulation, $C_{i,a}$ is normalized to the range $[-1, 1]$, where $N_{i,a}$ represents the count of advertisements detected on a given news page. The threshold $N'_{i,a}$ serves as a critical parameter, determining the maximum acceptable number of advertisements before the content is flagged as potentially unreliable. When the number of advertisements exceeds this threshold ($N_{i,a} > N'_{i,a}$), the model assigns the lowest credibility score of -1, reflecting the strong correlation between excessive advertising and questionable content quality. The linear relationship in the first case ($N_{i,a} \leq N'_{i,a}$) provides a gradual decrease in credibility score as the number of advertisements increases, allowing for nuanced evaluation of pages with moderate advertising presence. This approach acknowledges that while some level of advertising is standard in digital news media, excessive advertising often indicates prioritization of revenue generation over content quality. This mathematical framework provides a systematic way to evaluate the advertising context of news articles, contributing to the overall assessment of content credibility. The model's sensitivity can be adjusted through the threshold parameter $N'_{i,a}$, allowing for calibration based on evolving digital media standards and empirical observations of advertising patterns in reliable news sources.

Sentiment Analysis and Objectivity Assessment The second major component of our context-based analysis focuses on sentiment patterns and objectivity measurement. This parameter serves a dual purpose: detecting potential discrepancies between the sentiment expressed in the news title versus its main text, and evaluating the overall objectivity of the content. This analysis is particularly crucial as research has shown that misleading content often exhibits specific patterns in sentiment expression and objectivity levels.

Sentiment analysis in our framework operates on multiple levels. First, it determines whether the information in both the title and main text is conveyed in a positive, neutral, or negative manner [115]. Beyond this basic categorization, we also measure sentiment polarity - the intensity of positive or negative sentiments [116]. This dual approach provides a more nuanced understanding of the content's emotional undertones.

Our research has identified several patterns characteristic of deceptive content. Fake news articles often deliberately mix information with contrasting emotional undertones to create a specific emotional response in readers. Additionally, providers of misleading content frequently employ highly subjective language, prioritizing personal interpretation over objective factual reporting [117]. This tendency toward subjectivity and emotional manipulation serves as a valuable indicator in our detection framework.

To implement this analysis, we employ a sophisticated combination of sentiment analysis tools. The core sentiment evaluation utilizes the VADER algorithm [118], which analyzes both the news title sentiment ($S_{i,title}$) and the main text sentiment ($S_{i,text}$). This is complemented by TextBlob’s sentiment analyzer [119], which provides additional capabilities for assessing text objectivity.

Our sentiment analysis framework incorporates two primary components:

1. Title-Text Sentiment Coherence: This measures the alignment between the sentiment expressed in the title and the main text, helping identify potential clickbait or misleading headlines.

2. Global Objectivity Assessment: This evaluates the overall objectivity of the content, distinguishing between fact-based reporting and opinion-heavy content.

These components are mathematically combined into a single sentiment analysis factor $C_{i,s}$, normalized to the range $[-1, 1]$:

$$C_{i,s} = \sigma(2O_i - 1) - (1 - \sigma)(2\Delta S_i - 1)$$

where: - $\sigma \in [0, 1]$ is a weighting parameter that balances the importance of objectivity versus sentiment consistency - $O_i \in [0, 1]$ represents the objectivity measure, with 0 indicating highly subjective content and 1 representing purely objective content - $\Delta S_i = |S_{i,title} - S_{i,text}|$ measures the sentiment disparity between title and text - $S_{i,title}, S_{i,text} \in [0, 1]$ represent the normalized sentiment scores for title and text respectively

The combined context-based evaluation factor C_i for a news article n_i integrates both the advertisement analysis and sentiment analysis:

$$C_i = \rho C_{i,a} + (1 - \rho) C_{i,s}$$

where $\rho \in [0, 1]$ serves as a balancing parameter between the two contextual components. The resulting score $C_i \in [-1, 1]$ provides a comprehensive assessment of the content’s contextual credibility, with values approaching -1 indicating likely fake content and values near 1 suggesting credible information.

This approach to sentiment and context analysis enables our system to capture subtle indicators of content reliability that might not be apparent from purely textual or structural analysis.

5.4.3 Trust Model Framework

The development of a reliable trust model for news sources represents a critical component in combating misinformation. Our framework implements a sophisticated approach to evaluating news producers and quantifying their credibility across different topics. This section presents the mathematical foundation and theoretical underpinnings of our trust

evaluation system. The model operates by accessing and analyzing comprehensive historical data about news sources stored in our database. This includes detailed feedback on previously published articles and topic-specific performance metrics. When evaluating a particular source s within a specific topic domain t , the system employs a weighted combination of three fundamental trust factors, expressed through the following equation:

$$T_s^t = \gamma E_s^t + \delta H_s^t + (1 - \gamma - \delta) G_s^t \quad (5.7)$$

This trust formulation incorporates three novel components:

- *Expertise* (E_s^t): Measures the source’s demonstrated knowledge and accuracy within the specific topic domain
- *Coherence* (H_s^t): Evaluates the consistency of the source’s reporting over time
- *Goodwill* (G_s^t): Quantifies the source’s overall contribution to reliable information dissemination

Each component is carefully normalized to ensure values fall within the range $[0, 1]$. The weighting parameters γ and δ allow for dynamic adjustment of the relative importance of each factor, while maintaining the constraint that their sum, along with the weight of the Goodwill factor $(1 - \gamma - \delta)$, equals 1. This normalization ensures that the final trust value T_s^t also falls within the interval $[0, 1]$, providing an intuitive scale where higher values indicate greater trustworthiness. A crucial aspect of our model’s implementation is its handling of multidimensional feedback. For any given piece of news, the associated feedback $f_{s,i}^t$ incorporates multiple parameters:

- The source identifier s
- The topic classification t
- The specific news item identifier i

For clarity in subsequent discussions, we adopt a simplified notation where feedback is denoted as f_i , omitting explicit reference to topic and source parameters. This simplification extends to our treatment of the trust factors, where we use E , H , and G instead of their fully parameterized forms. This notational convention allows us to focus on the core relationships and mechanisms while maintaining mathematical precision. The model’s design reflects several key insights from our research into trust dynamics in news ecosystems:

1. **Topic Specificity:** Trust is not a monolithic quantity but varies across different subject domains. A source might be highly reliable in one area while showing less expertise in others.

2. **Temporal Dynamics:** The model captures how trust evolves over time, reflecting both recent performance and historical patterns.
3. **Multifaceted Evaluation:** By combining multiple factors, the model provides a more nuanced and robust assessment than single-metric approaches.

This framework provides a foundation for quantitative trust assessment while maintaining the flexibility to adapt to different news environments and evolving patterns of misinformation. The following sections will examine each component in detail, exploring how they contribute to the overall trust evaluation process.

5.4.3.1 Source Expertise Evaluation

The assessment of source expertise represents a fundamental component in our trust evaluation framework. This factor aims to quantify how well-informed and competent a source is within specific topic domains. Our approach recognizes that expertise is multifaceted, encompassing both the depth of engagement with particular topics and the demonstrated quality of content production. To capture these dimensions comprehensively, we evaluate expertise through two distinct but complementary parameters: Topic Importance (E_m) and Writing Competence (E_c).

The Topic Importance metric examines the source’s level of specialization and commitment to particular subject areas, while Writing Competence assesses the technical and professional quality of the source’s content production. This dual approach allows us to distinguish between sources that frequently cover a topic but with varying levels of proficiency, and those that may publish less frequently but maintain high professional standards.

Topic Specialization Assessment The evaluation of Topic Importance provides a quantitative measure of a source’s specialization level in specific subject areas. This parameter is particularly crucial in distinguishing between specialized, expert sources and general news outlets that cover a broader range of topics with less depth. Our approach recognizes that expertise often correlates with focused, sustained engagement in particular subject domains.

To quantify this specialization, we develop a Topic Importance metric E_m , normalized to the interval $[0, 1]$. This metric is computed using the following formula:

$$E_m = \frac{N_s^t}{\bigcup_t N_s^t} \quad (5.8)$$

In this formulation, N_s^t represents the number of articles published by source s on topic t , while $\bigcup_t N_s^t$ denotes the total volume of news articles published by the source across all topics. This ratio provides several key insights:

The numerator N_s^t directly measures the source's engagement with the specific topic under evaluation. A higher number of articles in a particular domain suggests greater investment in understanding and covering that area. The denominator $\bigcup_t N_s^t$ provides context by considering the source's overall publication volume, allowing us to distinguish between truly specialized sources and those that simply publish large volumes of content across many topics.

This metric effectively identifies sources that demonstrate sustained, focused attention to specific topics. For instance, a specialized economic news outlet might show a high E_m value for financial topics, while a general news source would likely show lower values across multiple topics. This distinction is crucial for assessing the likelihood that a source possesses deep, reliable knowledge in particular areas.

The normalization of E_m to the $[0, 1]$ interval provides an intuitive scale where values closer to 1 indicate higher levels of topic specialization. This facilitates both direct comparisons between different sources and integration with other trust metrics in our broader evaluation framework.

Writing Competence Assessment Beyond the quantitative measure of topic engagement, our framework incorporates a sophisticated evaluation of content quality through the Writing Competence parameter. This metric builds upon the message-based analysis detailed in Section 5.4.1.1, providing a comprehensive assessment of a source's ability to produce well-crafted, professional content within their chosen domains.

The Writing Competence score E_c is computed through a normalized aggregation of message-based evaluations across all articles published by a source within a specific topic:

$$E_c = \frac{1}{2N} \left(1 + \sum_{i=1}^N M_i \right) \quad (5.9)$$

where $M_i \in [-1, 1]$ represents the message-based evaluation factor for each news article n_i published by source s within topic t . This formulation ensures that the Writing Competence score remains within the normalized interval $[0, 1]$, facilitating integration with other trust metrics. The addition of 1 in the summation, combined with the factor of $\frac{1}{2N}$, provides this normalization while preserving the relative differences in writing quality between sources.

Integrated Expertise Evaluation The comprehensive assessment of source expertise combines both Topic Importance and Writing Competence through a weighted aggregation:

$$E = \tau E_m + (1 - \tau) E_c \quad (5.10)$$

This formulation yields an overall expertise score $E \in [0, 1]$, where values approaching zero indicate sources with minimal expertise in the topic, while scores near unity identify

highly proficient and specialized sources. The weighting parameter τ plays a crucial role in balancing these two components, adapting to the volume and consistency of a source's publications.

The dynamic nature of source expertise is captured through an adaptive weighting mechanism:

$$\tau = \frac{N}{\frac{1}{\omega}N + \psi} \quad (5.11)$$

The weighting function incorporates two key parameters:

- $\omega \in (0, 1]$ controls the asymptotic behavior of the weight as the number of publications increases
- $\psi \in [1, \infty)$ determines the rate at which the weighting factor responds to changes in publication volume

This formulation ensures that the relative importance of Topic Importance increases with the volume of published content, reflecting the principle that sustained engagement with a topic provides stronger evidence of expertise than occasional coverage. Conversely, for sources with limited publication history in a topic, the Writing Competence factor carries greater weight, as individual article quality becomes a more reliable indicator of expertise.

The parameters ω and ψ allow fine-tuning of this adaptive behaviour, with their specific calibration detailed in Section 5.4.6. This calibration ensures that the expertise evaluation remains robust across varying patterns of publication frequency and consistency, while maintaining sensitivity to changes in source behaviour and quality.

The resulting expertise metric provides a nuanced evaluation that captures both the breadth and depth of a source's topic-specific expertise. By combining quantitative measures of topic engagement with qualitative assessment of content production, this approach offers a more complete picture of source reliability than either measure alone could provide.

5.4.3.2 Social Impact Analysis: Relevance and Goodwill

The evaluation of news source credibility must consider not only the inherent quality of content but also its social impact and circulation patterns. Our framework incorporates two interconnected factors—Relevance and Goodwill—to analyse how news propagates through user networks and how this propagation affects perceived information quality. This approach acknowledges the complex relationship between content dissemination and credibility, particularly in the context of misinformation spread.

Contemporary research has highlighted the significant role of social networks in information dissemination and the complex dynamics of fake news propagation. Studies

have demonstrated that the perceived quality of information is strongly influenced by sharing patterns and user engagement [120]. This relationship creates a feedback loop where content visibility and perceived credibility can reinforce each other, potentially amplifying the impact of both accurate and misleading information.

To capture these dynamics, we introduce a sophisticated Relevance metric that quantifies the reach and impact of individual news articles. For any given article n_i , we define a Relevance factor $R_i \in (0, 1]$ that measures the relative frequency with which users access and engage with the content:

$$R_i = \frac{N_{r,i}}{N_r} \quad (5.12)$$

This formulation incorporates several key components:

- $N_{r,i}$ represents the total number of user requests for news article n_i
- N_r denotes the aggregate number of requests for all articles from source s within topic t
- The relationship $\sum_{i=1}^N N_{r,i} = N_r$ ensures proper normalization of the metric

The Relevance factor serves as a weighting mechanism that adjusts the impact of individual articles based on their reach and engagement levels. This adjustment is crucial because highly circulated content, whether accurate or misleading, tends to have a disproportionate impact on public perception and discourse.

Building upon this foundation, we develop a comprehensive Goodwill factor that integrates both content quality and social impact. This metric $G \in [0, 1]$ is computed as:

$$G = \frac{1}{2} \left(1 + \sum_{i=1}^N R_i \cdot f_i \right) \quad (5.13)$$

where:

- f_i represents the quality feedback score for article n_i
- R_i weights each article's contribution based on its relevance
- The normalization factor $\frac{1}{2}$ and the addition of 1 ensure the result falls within $[0, 1]$

This formulation creates a sophisticated measure that balances raw content quality with social impact. Articles that receive significant attention (high R_i) contribute more strongly to the overall Goodwill score, reflecting their greater influence on public perception. However, this increased weight applies to both positive and negative feedback, ensuring that widely circulated misinformation can appropriately damage a source's credibility.

The Goodwill metric provides several key insights into source reliability:

1. It captures the cumulative impact of a source’s content, weighted by actual user engagement
2. It penalizes sources that achieve high visibility through misleading content
3. It rewards sources that maintain high quality standards across their most impactful articles

This approach recognizes that trustworthiness emerges not just from producing accurate content, but from consistently delivering reliable information that reaches and influences audiences. The combination of feedback scores and relevance weights creates a more nuanced picture of source credibility than either measure alone could provide.

The integration of social impact metrics into the trust evaluation framework represents an enrichment of the system’s context over systems that consider only content characteristics. By incorporating user engagement patterns and their relationship with information quality, we create a more complete and realistic model of source credibility in modern digital media environments.

5.4.3.3 Temporal Coherence Analysis

While the Goodwill factor provides valuable insights into a source’s overall reliability, it may not adequately capture rapid changes in source behaviour, particularly in the context of sophisticated deception strategies. Our research has identified two primary attack patterns that can exploit this limitation: On-Off Attacks (OOAs) [121] and WhiteWashing Attacks (WWAs) [122]. These attack patterns represent significant challenges to traditional trust evaluation systems.

In On-Off Attacks, malicious sources alternate between periods of reliable reporting (ON state) and deliberate misinformation (OFF state). This oscillating behaviour can accumulate positive reputation during benevolent phases while maximizing damage during malicious periods. Similarly, Whitewashing Attacks involve sources periodically abandoning compromised identities and re-registering as new entities, effectively resetting negative reputation scores.

To address these sophisticated attack vectors, we introduce a Coherence factor that emphasizes recent behaviour patterns within a carefully calibrated temporal window. This approach considers the last N_{last} news items evaluated for a source s within topic t , providing a more responsive measure of source reliability. The Coherence factor $H \in [0, 1]$ is computed using a weighted summation:

$$H = \frac{1}{2} \left(1 + \sum_{z=1}^{N_{last}} w_z \cdot f_z \right) \quad (5.14)$$

where f_z represents the feedback score for news item z , and N_{last} defines the temporal window size. This formulation incorporates a sophisticated weighting scheme that prioritizes recent behavior while maintaining awareness of historical patterns.

The weighting mechanism employs a modified geometric distribution to assign greater significance to recent activities while retaining some influence from older interactions. For each feedback entry z within the temporal window, the weight w_z is computed as:

$$w_z = p(1-p)^{z-1} + \frac{\xi^{res}}{N_{last}} \quad (5.15)$$

This weighting scheme incorporates two key components: a geometric decay factor controlled by parameter p , and a residual weight distribution term ξ^{res} . The geometric component ensures that more recent feedback has stronger influence, while the residual term maintains proper normalization and historical context.

The residual weight term ξ^{res} accounts for the truncated tail of the geometric distribution beyond our temporal window, computed as:

$$\xi^{res} = \sum_{r=N_{last}+1}^N p(1-p)^{r-1} \quad (5.16)$$

This sophisticated approach to temporal weighting provides several key advantages:

First, it ensures rapid response to behavioral changes, enabling quick detection of transitions between benevolent and malicious states in OOA scenarios. The emphasis on recent behavior means that sudden shifts in content quality quickly impact the overall coherence score, making it harder for malicious sources to exploit accumulated positive reputation.

Second, the retention of some historical influence through the residual term helps identify patterns in oscillating behavior. This feature is particularly valuable for detecting systematic manipulation attempts, as it maintains awareness of past behavioral cycles while still prioritizing recent actions.

Third, the normalization of weights within $[0, 1]$ ensures that the coherence metric remains comparable across sources with different publication frequencies and histories. This standardization is crucial for fair comparison and integration with other trust metrics in our framework.

The effectiveness of this approach depends significantly on proper calibration of the temporal window size N_{last} and the geometric distribution parameter p . A smaller window size provides faster response to behavioral changes but may be more susceptible to short-term fluctuations, while a larger window offers greater stability but reduced sensitivity to recent changes. The geometric parameter p similarly balances the trade-off between recency bias and historical awareness.

Through careful tuning of these parameters, our Coherence metric provides robust

protection against sophisticated manipulation attempts while maintaining fair evaluation of legitimate sources that may experience natural variations in content quality. This temporal awareness complements the broader perspective provided by the Goodwill factor, creating a more comprehensive and resilient trust evaluation framework.

5.4.4 Experimental Evaluation

This section presents a comprehensive evaluation of our proposed framework, encompassing both the fake news detection system and the trust management algorithm. We describe our experimental methodology, dataset characteristics, and parameter configurations, followed by detailed analysis of the results.

5.4.4.1 Experimental Framework and Dataset Selection

The evaluation of our proposed framework requires a robust dataset that captures the complexity and diversity of real-world news content. For this purpose, we selected the comprehensive news dataset available through Kaggle [123], comprising 20,387 news articles. This dataset provides an excellent testbed for our analysis due to its size, diversity, and detailed labelling. The dataset structure reflects the key components necessary for news content analysis. Each entry contains four primary attributes:

- Title: The headline of the news article
- Author: The original content creator
- Text: The main body of the article
- Label: A binary classification (0 or 1) indicating fake or real news respectively

To adapt this dataset for our specific research objectives, we implemented several pre-processing steps. First, we modified the binary classification scheme to align with our evaluation framework, remapping the labels to the interval $[-1, 1]$, where -1 indicates fake news and 1 represents genuine content. This transformation provides a more nuanced scale for evaluation while maintaining the fundamental distinction between reliable and unreliable content. Given the substantial size of our dataset, we employed a strategic split for training and evaluation. We allocated 90% of the articles (18,348 items) to train the message-based factor, reserving the remaining 10% (2,039 items) for testing and validation. This division ensures robust model training while maintaining a significant test set for performance evaluation. A critical enhancement to the original dataset involved the integration of topic classification capabilities. For this purpose, we leveraged the topic classification framework developed in Chapter 3, which has demonstrated robust performance in categorizing news content. This system provides automatic classification

Table 5.3: Simulation Parameters and Their Optimal Values

Parameter	Description	Value
f_{th}	Minimum feedback threshold for fact-comparison evaluation	0.8
A_{min}	Similarity threshold for content comparison	0.86
$N'i, a$	Maximum acceptable advertisement count	2
σ	Sentiment analysis weighting factor	0.6
ω	Expertise evaluation asymptotic parameter	0.3
ψ	Expertise evaluation rate parameter	70
p	Coherence geometric distribution parameter	0.03
N_{last}	Temporal window size for trust assessment	50
γ	Expertise contribution weight	0.25
δ	Coherence contribution weight	0.5

into ten distinct domains: Arts & Culture, Business & Economy, Crime & Security, Entertainment & Celebrity, Health & Education, Politics, Science, Sports, Tech, and Weird. The integration of this pre-existing classification system allows us to maintain consistency across our research while benefiting from its proven effectiveness in topic categorization. The preprocessing pipeline for our experimental framework built upon this classification system, implementing additional text cleaning procedures and standardization steps to ensure optimal performance in the fake news detection context. This approach allowed us to leverage classification capabilities previously developed while adapting them specifically for fake news detection and trust evaluation. Table 5.3 presents the optimal configuration parameters determined through extensive experimentation. These parameters were carefully tuned to balance system sensitivity with robustness across different types of content and attack scenarios. The selection process for these parameters involved systematic evaluation of system performance across multiple metrics, including detection accuracy, false positive rates, and resilience to various deception strategies. For example, the feedback threshold $f_{th} = 0.8$ was chosen to ensure high confidence in fact-comparison evaluations, while the similarity threshold $A_{min} = 0.86$ provides a stringent criterion for content matching. The temporal window size $N_{last} = 50$ represents an optimal balance between historical awareness and responsiveness to behavioural changes. The subsequent sections provide detailed analysis of system performance using these parameter values, demonstrating their effectiveness in real-world scenarios. The integration of the original dataset with our enhanced topic classification and parameter optimization creates a robust framework for evaluating both the detection accuracy of our system and its effectiveness in establishing

and maintaining trust metrics for news sources. This comprehensive approach ensures that our experimental results reflect realistic challenges in news content evaluation while maintaining rigorous scientific standards.

5.4.5 Performance Analysis of the Feedback Model

The effectiveness of our feedback model hinges on its ability to evaluate news content from multiple perspectives and synthesize these evaluations into a single, meaningful reliability score. This section presents a detailed analysis of the model’s performance, examining how different components contribute to the overall assessment of news reliability.

The feedback score f_i associated with news article n_i represents a sophisticated integration of message-based, fact-based, and context-based parameters. This composite score is normalized to the interval $[-1, 1]$, where negative values approaching -1 indicate strong likelihood of disinformation, while positive values near 1 suggest reliable informational content. The intermediate values provide nuanced gradations of content reliability, allowing for fine-grained distinction between different levels of content quality.

5.4.5.1 Message-Based Classification Performance

Our initial experimental focus centred on validating the effectiveness of various machine-learning approaches for message-based classification. We conducted a comprehensive comparison of five state-of-the-art classification algorithms:

The XGBoost algorithm served as our primary candidate, evaluated against four established benchmarks:

- NBC [124], selected for its proven effectiveness in text classification tasks
- SVM [125], known for its robust performance in high-dimensional spaces
- K-Nearest Neighbours (KNN) [126], chosen for its intuitive approach to classification
- Random Forest (RF) [127], included for its ensemble learning capabilities

To ensure rigorous evaluation of classifier performance, we employed two complementary metrics: the Receiver Operating Characteristic (ROC) curve and the Area Under Curve (AUC). The ROC curve provides a graphical representation of classification performance across various decision thresholds by plotting the True Positive Rate against the False Positive Rate. This visualization allows for detailed analysis of the trade-offs between sensitivity and specificity at different operating points. The AUC metric complements this by providing a single scalar value representing overall classification performance, where an AUC of 1.0 indicates perfect classification and 0.5 represents random chance.

Figure 5.4 presents the comparative ROC curves for all evaluated classifiers. The results demonstrate clear superiority of both XGBoost and RF algorithms over their

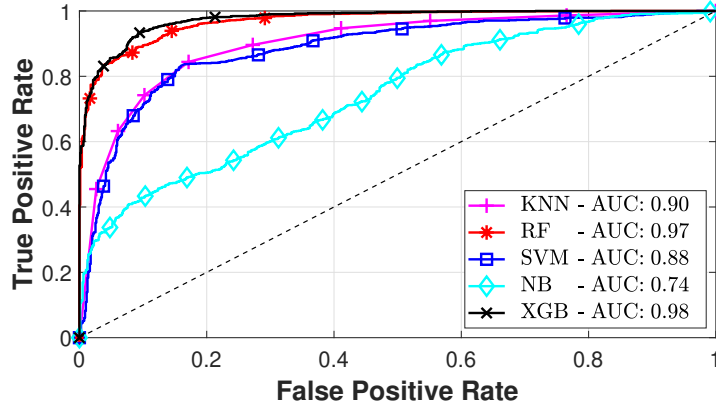


Figure 5.4: ROC curves for the proposed machine learning classifiers.

competitors, with both methods achieving notably higher AUC values. This superior performance can be attributed to their ensemble learning approaches, which effectively capture complex patterns in the textual features of news articles. In addition, Figure 5.5 shows the message-based factor reached a 92% accuracy, with 91% concerning fake and 93% for real news.

While both XGBoost and RF demonstrated excellent classification capabilities, our final implementation employs XGBoost due to its significant advantages in computational efficiency. In our experimental setup, XGBoost achieved complete training convergence in less than one minute, substantially outperforming RF in terms of processing speed while maintaining comparable accuracy.

The practical performance of our XGBoost implementation proved exceptional, achieving an overall accuracy of % across the test dataset. More specifically, the system demonstrated balanced performance across both classes, achieving 91% accuracy in identifying fake news and 93% accuracy in recognizing legitimate content. This balanced performance is particularly noteworthy as it indicates the model's robust capability to handle both types of content without significant bias toward either category.

These results validate our choice of XGBoost as the primary classification algorithm for the message-based component of our feedback model. The high accuracy rates, combined with excellent computational efficiency, provide a strong foundation for reliable news content classification. Moreover, the balanced performance across fake and legitimate news categories suggests that the model has successfully learned meaningful patterns that distinguish between these content types, rather than relying on superficial or biased features.

5.4.5.2 Analysis of Fact-Based Detection Performance

The evaluation of our fact-based detection component represents a critical aspect of the overall system assessment. This analysis encompasses two complementary approaches:

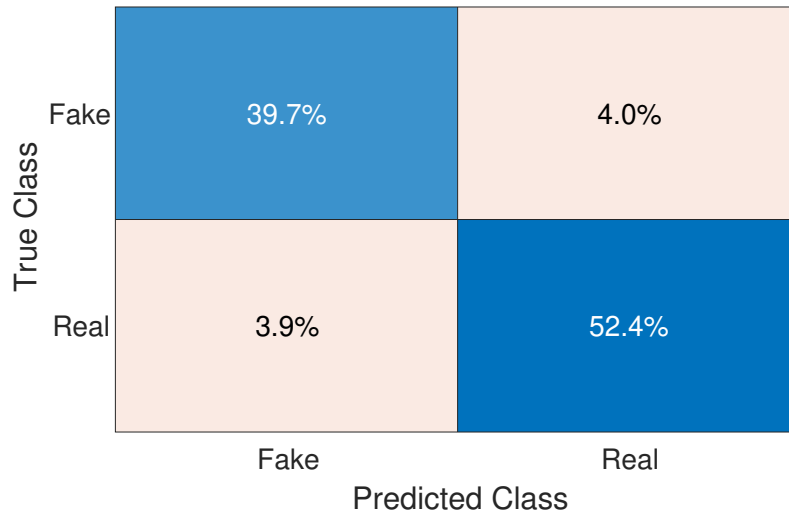


Figure 5.5: XGBoost model - Test Confusion Matrix

the FEVER-based claim verification system and the similarity-based fact comparison mechanism. Each component contributes distinct capabilities to the detection framework, and their combined performance provides insights into the effectiveness of our multi-faceted approach to fact verification.

FEVER-Based Claim Verification Our implementation of the FEVER system for claim verification builds upon significant research demonstrating its effectiveness in identifying supported claims [128]. The integration of FEVER into our framework required careful consideration of input preparation and result interpretation to maximize detection accuracy.

To optimize the system’s performance, we implemented a sophisticated preprocessing pipeline utilizing the Stanford Parser tool [129]. This preprocessing stage serves several crucial functions:

First, it decomposes news articles into fundamental semantic units consisting of subject-verb-object triplets. This decomposition provides a structured representation that facilitates more accurate claim verification. Second, it performs selective removal of non-essential speech components, producing simplified input that enhances FEVER’s processing efficiency while maintaining semantic integrity.

Following this preprocessing, we compute the supported claims metric V_i for each news article n_i . The distribution of these results, visualized in Figure 5.6, reveals several significant patterns in the relationship between supported claims and article veracity. The visualization employs varying color intensities to represent the proportion of genuine news within each category, with each bar aggregating data from 400 articles to provide meaningful statistical representation.

The analysis reveals a clear inverse relationship between the ratio V_i/L_i and article

authenticity. Higher values of this ratio consistently correspond to increased concentrations of fake news, while lower values indicate a predominance of genuine content. This pattern becomes particularly pronounced near the distribution boundaries, where classification accuracy shows marked improvement. The overall performance achieves a 70% accuracy rate using a straightforward classification rule where articles are labelled as fake when $F_{i,f} < 0$ and genuine when $F_{i,f} \geq 0$.

Similarity-Based Fact Comparison Complementing the FEVER-based analysis, our framework incorporates a similarity-based fact comparison component that leverages previously evaluated content to enhance detection accuracy. As detailed in Section 5.4.1.2, this approach requires careful calibration of two critical parameters: the similarity threshold A_{min} and the minimum feedback threshold f_{th} . These parameters allow us to selectively consider only the most relevant and reliable previously analyzed articles.

Figure 5.7 presents a comprehensive analysis of how detection accuracy varies across different threshold configurations. The visualization contrasts these results with FEVER’s consistent performance baseline, revealing several important insights about parameter optimization:

Maximum accuracy can be achieved with highly stringent thresholds—specifically, requiring 90% similarity and a minimum feedback score of 0.8. However, this configuration proves impractical in real-world applications, as it allows evaluation of only 5% of incoming articles. This limitation led us to investigate more balanced parameter settings that could maintain high accuracy while achieving broader coverage.

Through systematic evaluation, we identified an optimal configuration using $A_{min} = 0.87$. This adjustment expands the system’s coverage to 23% of articles while maintaining superior accuracy compared to the FEVER baseline. For articles falling outside this coverage, the system defaults to FEVER-based evaluation, ensuring comprehensive assessment capabilities across all content types.

This dual-component approach to fact-based verification demonstrates the benefits of combining complementary detection strategies. While FEVER provides broad coverage with consistent performance, the similarity-based component offers enhanced accuracy for a significant subset of content. The integration of these approaches, coupled with carefully tuned parameters, creates a robust framework for fact verification that balances accuracy with practical applicability.

5.4.5.3 Analysis of Context-Based Detection Performance

The evaluation of our context-based detection framework encompasses two primary analytical dimensions: advertisement pattern analysis and sentiment-based content evaluation. These complementary approaches provide distinct insights into content credibility, leveraging both structural and linguistic indicators of potential misinformation.

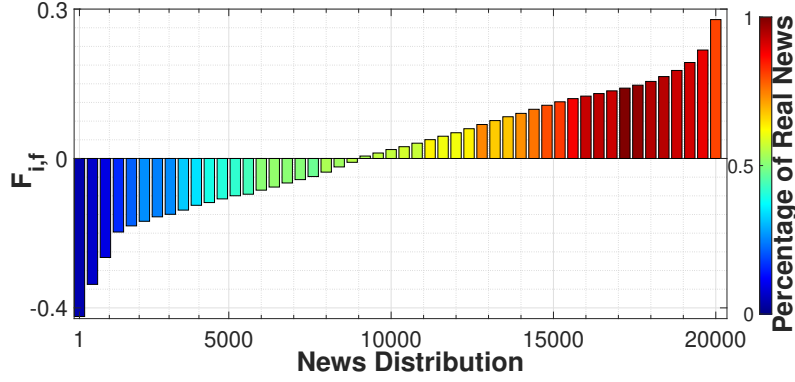


Figure 5.6: FEVER Score distribution

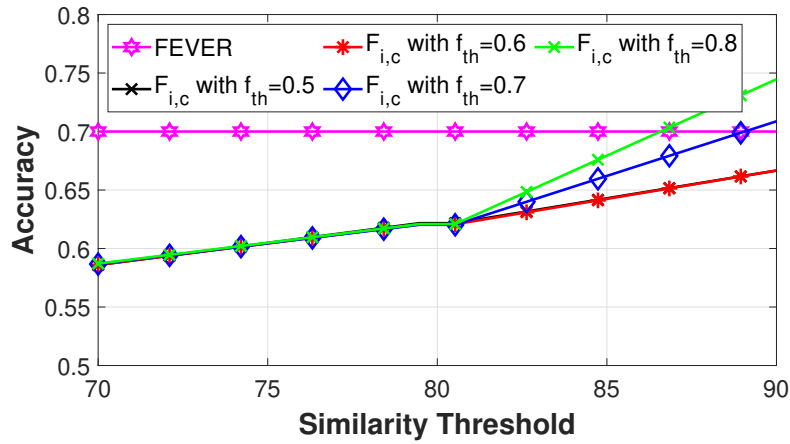


Figure 5.7: Fact-based overall accuracy over different thresholds of feedback and similarity

Our analysis of advertisement patterns reveals significant correlations between advertising density and content reliability. Figure 5.8 illustrates a clear relationship between advertisement frequency and content authenticity, with reliable news sources demonstrating notably restrained advertising practices. The data shows a pronounced threshold effect at $N'_{i,a} = 2$ advertisements, beyond which the probability of genuine content decreases precipitously. This finding allows for the implementation of a robust classification criterion where content with excessive advertising ($N_{i,a} > 2$) receives a strongly negative credibility score ($C_{i,a} = -1$).

The advertisement analysis provides particularly strong discriminative power, as genuine news outlets typically maintain strict standards regarding advertising density to preserve readability and professional appearance. The sharp decline in genuine content representation beyond our identified threshold suggests that excessive advertising often serves as a reliable indicator of questionable content sources.

The sentiment analysis component of our evaluation reveals additional subtle patterns distinguishing genuine from fraudulent content. Table 5.4 presents a comprehensive statistical analysis of sentiment indicators across both title and main text components. The

analysis reveals several key patterns in sentiment distribution:

The sentiment scores for article titles (S_{title}) show notably different characteristics between genuine and fake content. Genuine news demonstrates more consistent sentiment expression ($\sigma^2 = 0.011$) compared to fraudulent content ($\sigma^2 = 0.022$), with genuine content maintaining higher average sentiment scores ($\bar{x} = 0.78$ versus 0.71). This pattern suggests that fake news often employs more extreme or variable emotional tones in headlines to attract attention.

Main text sentiment analysis (S_{text}) reveals even more pronounced differences, with genuine content showing higher average sentiment (0.65 versus 0.55) but lower variance (0.075 versus 0.085). This finding indicates that legitimate news maintains more consistent emotional tone throughout the article, while fake news often exhibits greater emotional volatility.

The sentiment differential (ΔS) between title and text provides particularly valuable insights. Fake news shows both higher average discrepancy (0.25 versus 0.21) and greater variance in this discrepancy (0.041 versus 0.036), suggesting that fraudulent content more frequently employs clickbait tactics where headline sentiment diverges significantly from article content.

Objectivity analysis (O) reveals subtle but significant differences, with fake news showing slightly higher average objectivity scores (0.42 versus 0.41) but notably higher variance (0.017 versus 0.007). This counter-intuitive finding suggests that some fake news attempts to mimic the objective tone of legitimate journalism, but does so inconsistently.

The integration of these sentiment metrics yields a robust sentiment-based classification score ($C_{i,s} = 0.72$), which contributes significantly to the overall context-based detection accuracy of 76% . This performance level demonstrates the effectiveness of combining advertisement pattern analysis with sophisticated sentiment evaluation in identifying potentially fraudulent content.

The strong performance of our context-based approach, achieving 76% accuracy through the combination of advertisement and sentiment analysis, validates the importance of considering contextual factors in fake news detection. The complementary nature of these indicators, structural (advertisements) and linguistic (sentiment), provides robust detection capabilities that can identify suspicious content through multiple independent signals.

5.4.6 Trust Model Performance Analysis

The evaluation of our trust model's performance required careful calibration of various parameters and comprehensive testing across multiple behavioural scenarios. This section presents our analysis of the model's effectiveness in detecting and responding to different patterns of source behaviour, including both steady-state reliable reporting and various forms of deceptive behaviour.

Table 5.4: Investigation for the sentiment analysis.

	Real		Fake	
	$\bar{x}$	σ^2	$\bar{x}$	σ^2
S'_{title}	0.78	0.011	0.71	0.022
S_{text}	0.65	0.075	0.55	0.085
ΔS	0.21	0.036	0.25	0.041
O	0.41	0.007	0.42	0.017
$C_{i,s} = 0.72$				

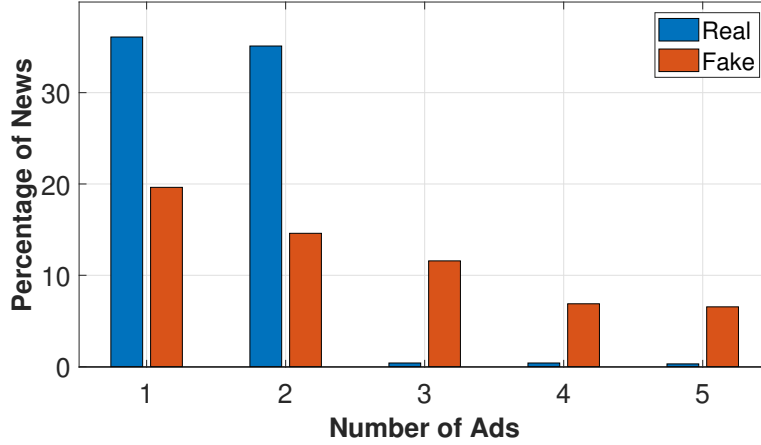


Figure 5.8: Impact of advertisements (ads)

Parameter Optimization Initial experimentation focused on optimizing the weighting parameters for our three primary metrics: Expertise, Goodwill, and Coherence. Through systematic evaluation, we determined that optimal performance was achieved with balanced weighting between Expertise ($\gamma = 0.25$) and Coherence ($\delta = 0.25$), with the remaining weight allocated to Goodwill. This configuration provides robust detection capabilities while maintaining responsiveness to behavioural changes.

Behavioural Response Analysis To evaluate our model’s effectiveness under realistic conditions, we designed a series of experiments simulating three distinct behavioral patterns commonly observed in news media. These patterns reflect the competitive pressures in modern journalism that often lead to rushed publication without proper verification [130]. Figure 5.9 illustrates the trust score evolution under these three scenarios:

The first scenario represents consistently benevolent behaviour, where the source maintains high standards of verification and accuracy. This establishes our baseline for trustworthy source behaviour and demonstrates the model’s ability to recognize and reward consistent reliability.

The second scenario examines a more complex pattern where a source initially builds credibility through 50 reliable articles before transitioning to publishing 100 pieces of

fake news. This pattern tests our model’s ability to detect and respond to deteriorating source quality, simulating cases where previously reliable sources may be compromised or deliberately shift their practices.

The third scenario inverts this pattern, beginning with 50 fake news articles before attempting to establish credibility through 100 genuine publications. This tests our model’s capacity to recognize genuine improvement while maintaining appropriate scepticism of sources with problematic histories.

The results demonstrate our model’s robust adaptability, particularly through the Coherence factor’s ability to rapidly detect and respond to behavioural changes. The trust scores show appropriate sensitivity to both positive and negative behavioural transitions, maintaining a balance between responsiveness and stability.

Attack Resistance Analysis To further validate our model’s robustness, we conducted a series of detailed experiments specifically designed to evaluate the system’s resistance to targeted manipulation attempts. Our experimental protocol implemented a sophisticated attack pattern that tested both the system’s ability to detect coordinated disinformation campaigns and its capacity to recognize genuine rehabilitation efforts. Figure 5.10 presents a comparative analysis of system performance under two configurations: one incorporating our proposed relevance algorithm and another operating without it. The experiment was structured in four distinct phases, each designed to test different aspects of the system’s response to manipulation: The initial convergence phase established baseline behavior through the publication of 20 genuine articles. During this period, both system configurations demonstrated identical performance trajectories, confirming the consistency of our baseline measurements. This phase was crucial for establishing the natural behavior patterns of legitimate news sources. The second phase implemented a concentrated attack strategy, where a single piece of false content was repeatedly requested 20 times. This phase revealed significant divergence between the two system configurations. The relevance-enabled system demonstrated rapid detection of the repeated false content, triggering an accelerated decrease in trust scores. This behavior reflects the system’s ability to recognize and respond to coordinated amplification attempts. In contrast, the baseline system without relevance consideration showed only modest response to the repeated content, treating each instance as an independent occurrence. The third phase introduced a distributed attack pattern, introducing 20 distinct pieces of false content. This phase revealed another important distinction between the two configurations. The relevance-enabled system maintained its aggressive response to false content, while the baseline system showed delayed recognition of the systematic nature of the attack. The trust value in the relevance-enabled system exhibited a steeper decline during this phase, demonstrating enhanced sensitivity to patterns of deceptive behavior. The final recovery phase, consisting of 70 genuine articles, provided insights into the system’s ability to recognize and respond

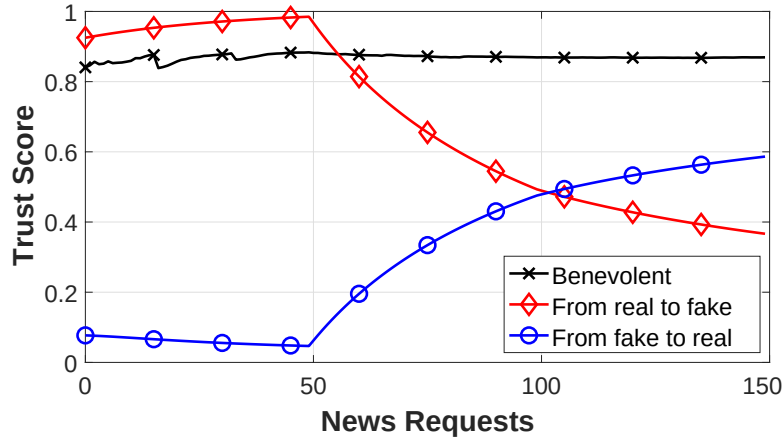


Figure 5.9: Final trust score: Three different behaviours.

to genuine rehabilitation efforts. Both configurations showed capacity for trust recovery, but with notable differences in their response patterns. The relevance-enabled system demonstrated more nuanced recovery behavior, with trust scores increasing more gradually and showing distinct inflection points corresponding to the temporal window boundaries (N_{last}) of the Coherence factor. Particularly noteworthy were the abrupt changes observed during the recovery phase’s rising edge. These discontinuities correspond to moments when the temporal analysis window (N_{last}) becomes predominantly populated with positive feedback scores. This behavior demonstrates the effectiveness of our temporal weighting scheme in capturing recent behavioral trends while maintaining appropriate skepticism of sources with problematic histories. The comparative analysis conclusively demonstrates the superior attack resistance provided by the relevance-enhanced configuration. While both systems eventually recognize and respond to systematic manipulation attempts, the relevance-enabled system provides substantially faster detection and more appropriate penalization of coordinated disinformation efforts. This enhanced protection against manipulation, combined with the system’s ability to recognize genuine rehabilitation efforts, represents a significant advancement in trust modeling for news sources. These results validate our architectural decisions and parameter choices, particularly the integration of relevance metrics with temporal analysis windows. The system successfully balances the competing requirements of rapid response to behavioral changes and resistance to manipulation attempts, while maintaining the flexibility to accommodate genuine changes in source behavior.

Building upon our analysis of attack resistance, we conducted additional experiments using real-world data to evaluate the system’s performance on established news sources. This analysis provided crucial insights into how our trust model handles actual news organizations with different editorial standards and reporting practices.

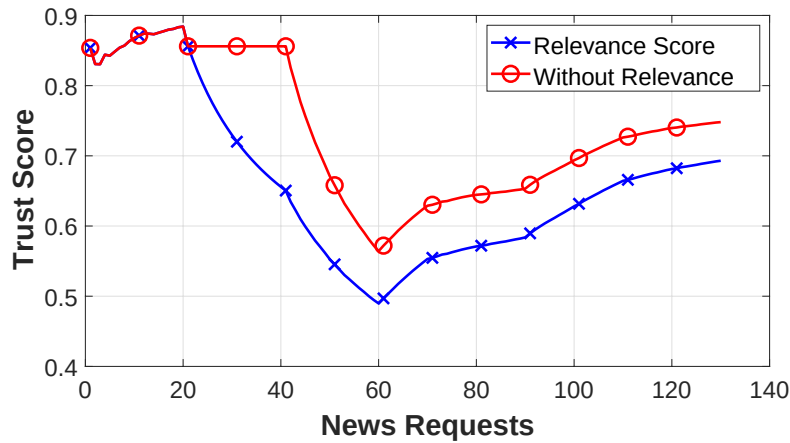


Figure 5.10: Trust score with and without the relevance algorithm.

Analysis of Real-World News Sources Table 5.5 presents a comparative analysis of two prominent news sources, the New York Times and Breitbart, across two major topic categories: Business & Economy (B&E) and Politics. This analysis reveals several significant insights about our trust model’s behavior with real-world content: First, the results demonstrate the model’s ability to capture nuanced differences in content quality that go beyond simple binary classification of fake versus real news. A particularly noteworthy finding emerges from the Business & Economy coverage, where Breitbart shows a lower percentage of explicitly fake news (0%) compared to the New York Times (1.25%). However, the average feedback and trust scores tell a more complex story, with the New York Times achieving significantly higher scores (0.94 for feedback, 0.92 for trust) compared to Breitbart (0.76 for feedback, 0.73 for trust). This apparent paradox highlights a crucial feature of our evaluation system: its ability to assess the qualitative aspects of news content rather than merely identifying explicitly false information. The continuous nature of our feedback scoring system, particularly influenced by the message-based component, allows for fine-grained differentiation in content quality. Articles that may not contain outright false information can still receive lower scores based on factors such as writing quality, source attribution, and argumentative coherence. The political coverage analysis provides additional validation of this pattern. While both sources show relatively low rates of explicitly fake news (0% for NYTimes, 0.54% for Breitbart), the substantial difference in average feedback scores (0.94 vs 0.72) and trust scores (0.91 vs 0.69) reflects deeper qualitative differences in their reporting approaches. These differences manifest in various aspects captured by our model:

Writing Quality: The higher scores achieved by the New York Times reflect more sophisticated writing, better source attribution, and more comprehensive coverage of topics. **Consistency:** The minimal variation between topics (0.94/0.92 for B&E vs 0.94/0.91 for Politics) in the New York Times scores suggests consistent editorial standards across

Table 5.5: Trust and average feedback scores related to the most evident sources

Source	Topic	% Fake news (number of fake, number of real)	Average Feedback, Trust score
NYTimes	B&E	1.25 (1, 80)	0.94, 0.92
Breitbart	B&E	0 (0, 86)	0.76, 0.73
NYTimes	Politics	0 (0, 170)	0.94, 0.91
Breitbart	Politics	0,54 (1, 185)	0.72, 0.69

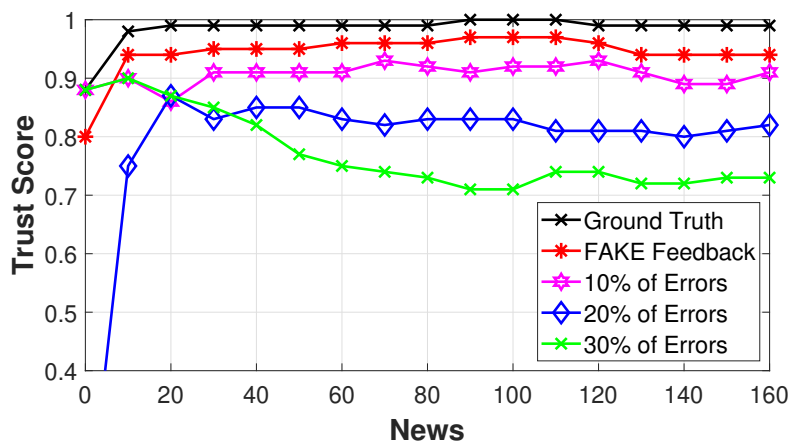


Figure 5.11: Trust model performance through different fake news detector accuracy systems.

different subject areas. Topic-Specific Performance: Breitbart shows slightly better performance in Business & Economy coverage (0.76/0.73) compared to Politics (0.72/0.69), indicating potential variations in editorial rigor across different topics.

This real-world validation demonstrates the sophistication of our trust model in capturing the multifaceted nature of news quality. Rather than relying solely on binary classifications of true versus false, the system provides a more nuanced evaluation that better reflects the complex reality of modern news media. This capability is particularly valuable in an environment where the distinction between reliable and unreliable news sources often lies not in obvious falsehoods but in more subtle aspects of journalistic quality and professional standards. The results also validate our decision to implement a continuous scoring system rather than a binary classification approach. This granular evaluation enables the system to reflect meaningful differences in content quality even among sources that rarely publish explicitly false information, providing users with more informative assessments of source reliability.

Impact of Detection Accuracy on Trust Evaluation To comprehensively understand our trust model’s robustness, we conducted a detailed analysis of how variations in fake news detection accuracy influence the overall trust evaluation system. This investigation is crucial because real-world detection systems rarely achieve perfect accuracy, and understanding

how detection errors propagate through the trust evaluation system is essential for assessing its practical reliability. Figure 5.11 presents a comparative analysis across multiple accuracy scenarios, using a reliable news source as our test case. The experiment establishes several key reference points for evaluation: The black line represents the idealized ground truth case, where we assume perfect detection accuracy. In this scenario, the feedback system operates in a purely binary mode, assigning discrete values of $f_i = -1$ to fake news and $f_i = 1$ to genuine content. This idealized case serves as our baseline for comparing the performance of more realistic scenarios. The red line demonstrates the actual performance of our proposed detection system, revealing how closely our implementation approximates the ideal case. The comparison between these two curves provides valuable insights into the practical effectiveness of our approach under real-world conditions. To systematically evaluate the system's resilience to detection errors, we introduced controlled levels of inaccuracy into the feedback evaluation process. This analysis included scenarios with 10% and 20% detection error rates. The results reveal several important characteristics of our trust model's behavior under varying detection accuracy:

1. **Robust Error Tolerance:** The trust evaluation system demonstrates remarkable resilience to moderate levels of detection error. Even with a 10% error rate in the detection system, the trust scores maintain a strong correlation with the ground truth trajectory.
2. **Performance Degradation:** As error rates increase, the system's performance degrades gradually rather than catastrophically. This behavior is particularly evident in the 20% error scenario, where the trust scores continue to track the general trend of the ground truth, albeit with reduced precision.
3. **Threshold Identification:** The analysis reveals that the system maintains acceptable performance up to approximately 20% detection error. Beyond this threshold, particularly at the 30% error level, the trust evaluations begin to deviate significantly from the ground truth pattern.

These findings have important implications for the practical deployment of our trust evaluation system: First, they demonstrate that the trust model can effectively compensate for moderate inaccuracies in the underlying detection system. This robustness is crucial for real-world applications where perfect detection accuracy is unattainable. Second, the identification of a 20% error threshold provides valuable guidance for system deployment and maintenance. This threshold can serve as a key performance indicator, helping to determine when detection system recalibration or retraining might be necessary. Third, the gradual degradation in performance suggests that the system can continue to provide useful, if less precise, trust evaluations even under suboptimal detection conditions. This graceful degradation characteristic is particularly valuable in practical applications where

Table 5.6: Machine Specifications

	Client (Plugin)	Server (FAKE System)
Processor	Intel(R) Core(TM) i7-5500U CPU @2.40GHz	Intel(R) Xeon(R) E5-1620 CPU @3.60GHz
Memory	8GB Samsung 1600MHz	32GB Samsung 1600 MHz
Storage	Samsung SSD 870 QVO 1TB	SanDisk SSD PLUS 240GB
Operating System	Windows 10 Home 64 bit	Ubuntu 22.04 LTS

temporary reductions in detection accuracy might occur. The ability of our trust model to maintain reliable performance despite moderate detection errors validates its practical utility in real-world scenarios. This resilience, combined with the system’s ability to provide meaningful trust evaluations even under imperfect detection conditions, represents a significant advancement in the development of robust trust evaluation systems for news sources.

5.4.7 System Performance Analysis in Real-World Conditions

The research required a comprehensive evaluation of the complete system’s performance under real-world operating conditions. This section presents our findings from extensive testing that simulated actual usage patterns across multiple users and diverse news sources. Our evaluation focused particularly on system responsiveness, resource utilization, and scalability under concurrent user loads.

Experimental Setup and Methodology To ensure realistic testing conditions, we implemented a simulation environment that reflected typical user behavior patterns when accessing news content. The test environment consisted of a two-tier architecture: a client-side component implemented as a browser plugin, and a server-side system handling the core FAKE system functionality. Table 5.6 details the hardware and software specifications for both client and server configurations, demonstrating our commitment to testing under realistic hardware constraints. For our primary test scenarios, we selected three prominent news sources with distinct editorial approaches and content strategies: BBC, representing traditional public broadcasting; Breitbart, offering alternative perspectives; and USA Today, providing mainstream commercial news coverage. This diverse selection allowed us to evaluate system performance across varying content styles and editorial standards.

Performance Under Concurrent Load The system’s ability to handle multiple simultaneous users represents a critical performance metric for real-world deployment. Our testing protocol involved simulating concurrent access from 100 simultaneous users, each actively engaging with news content through the browser plugin. This testing revealed

several key performance characteristics:

1. Initial Content Evaluation:

- Average processing time: < 2 seconds for new content
- This includes complete analysis through all evaluation components (message-based, fact-based, and context-based)
- Represents full pipeline processing including feature extraction and model inference

2. Cached Content Retrieval:

- Average response time: < 0.2 seconds for previously evaluated content
- Demonstrates effective caching and database optimization
- Highlights the system's efficiency in handling repeated requests

Resource Utilization Analysis Table 5.7 provides detailed measurements of the system's resource impact on client machines. The comparison between baseline browser performance and plugin-enabled operation reveals several important characteristics: The average execution time shows a modest increase from 1.23 seconds to 1.54 seconds with the plugin active, representing approximately a 25% overhead. This increase remains within acceptable bounds for interactive web browsing, ensuring that the user experience isn't significantly impacted. Memory utilisation demonstrates a more substantial impact, with average usage increasing from 25.74 MB to 79.12 MB. While this represents a significant percentage increase, the absolute values remain well within the capabilities of modern client systems. Furthermore, this memory footprint includes all necessary caching and state management components that enable the plugin's rapid response capabilities. These performance metrics validate our architectural decisions and implementation strategies, demonstrating that the system can provide sophisticated news evaluation capabilities while maintaining acceptable resource utilization levels. The measured response times, particularly for cached content, suggest that the system can scale effectively to support larger user bases while maintaining responsive performance. The combination of reasonable processing latency for new content and rapid retrieval of cached evaluations indicates that the system achieves an effective balance between computational thoroughness and user experience requirements. This balance is crucial for maintaining user engagement while ensuring the reliability of content evaluation.

Table 5.7: Average impact of the plugin over 100 users.

	No-Plugin	Plugin
Average Time Execution [s]	1.23	1.54
Average Memory Usage [MB]	25.74	79.12

Table 5.8: Database and Blockchain comparison.

Database	Write			Read		
# News	1	500	1000	1	500	1000
Latency [s]	0.0019	0.0036	0.0054	0.00046	0.00042	0.00048
THR [MBps]	31.45	24.20	18.03	647.90	714.15	693.53

Blockchain	Write			Read		
# News	1	500	1000	1	500	1000
Latency [s]	1.20	10.04	20.44	0.080	0.077	0.078
THR [MBps]	0.30	0.06	0.032	4.46	4.37	4.40

5.4.8 Alternative Approaches to Data Management: A Blockchain Perspective

The secure management of feedback data and source evaluations represents a critical component of our fake news detection system. While our primary implementation utilizes traditional database technologies, the emergence of blockchain as a potential alternative merits careful consideration. This section presents a comparative analysis of blockchain technology and conventional database approaches, examining their respective strengths and limitations in the context of fake news detection systems.

Blockchain Technology in Content Verification Recent years have witnessed growing interest in blockchain technology as a tool for combating misinformation [131]. Several researchers have demonstrated the potential of blockchain for storing and validating content evaluations [132, 133], highlighting its capacity to create immutable records of news assessments and source credibility scores. The fundamental architecture of blockchain presents several compelling advantages for content verification systems. At its core, blockchain functions as a distributed digital ledger, replicating transaction records across a decentralized network of nodes. This distributed structure eliminates the need for central authentication authorities, enabling trustworthy peer-to-peer interactions through consensus mechanisms rather than institutional oversight. The security benefits of blockchain stem from its fundamental design principles. Unlike centralized systems that present single points of failure, blockchain’s distributed architecture provides inherent resistance to tampering and manipulation. The consensus algorithms that maintain the distributed ledger ensure that any attempted modifications to historical records would require simultaneous alterations across multiple network nodes, making unauthorized changes practically impossible [134].

Implementation Architecture Our comparative analysis involved implementing two parallel systems: our primary solution using an encrypted database and an alternative implementation utilizing the Ethereum blockchain platform through Alastria [135]. The blockchain implementation stores each news evaluation as a distinct transaction, creating new blocks that become permanently linked to the existing chain. This structure ensures the immutability of evaluation records, as modifying any single evaluation would require alterations to all subsequent blocks in the chain. Smart contracts, implemented in the Solidity programming language, provide the operational framework for our blockchain solution. These contracts define the data structures for storing evaluations and establish the protocols for creating and retrieving transaction records. Through careful contract design, we ensured that all operations remain verifiable while maintaining necessary access controls.

Performance Analysis Table 5.8 presents a detailed performance comparison between the two implementations, focusing on read and write operations across varying workloads. Our analysis considered two critical metrics: latency (measuring operation completion time) and throughput (quantifying data transfer rates in megabits per second). The results reveal significant performance differences between the two approaches. The traditional database consistently demonstrates superior performance metrics, particularly in write operations. For single-record operations, database write latency averages 0.0019 seconds, compared to 1.20 seconds for blockchain. This performance gap widens substantially under higher loads, with the database maintaining sub-second response times even for 1000-record operations, while blockchain write times extend to over 20 seconds. Read operations show similar patterns, though with less dramatic differences. Database read latency remains consistently below 0.5 milliseconds, while blockchain reads require approximately 80 milliseconds, representing a significant but less extreme performance differential than write operations. Throughput measurements further emphasize these performance disparities. Database write throughput ranges from 31.45 MBps for single records to 18.03 MBps under maximum load, while blockchain write throughput decreases from 0.30 MBps to 0.032 MBps. Read throughput shows even more dramatic differences, with database performance exceeding 600 MBps compared to blockchain's approximately 4.40 MBps. A significant limitation of the Ethereum blockchain implementation relates to scalability. Current platform limitations restrict processing to approximately 500,000 transactions daily, with a maximum throughput of 15 transactions per second [136]. These constraints present serious challenges for high-volume applications requiring real-time responsiveness.

Technology Selection Considerations Despite its performance limitations, blockchain technology offers compelling advantages for specific use cases. In scenarios where data provenance and tamper resistance take precedence over raw performance, blockchain's

inherent security characteristics [137] may justify its adoption despite the associated performance overhead. The choice between blockchain and traditional database implementations ultimately depends on specific application requirements and priorities. Our analysis suggests that applications demanding real-time responsiveness should favour traditional database implementations, while those prioritizing absolute data integrity and audit trails might benefit from blockchain’s security guarantees despite its performance implications.

5.4.9 Conclusions and Future Directions

This research has presented a comprehensive framework for detecting and mitigating the spread of misinformation in online news content. Our work addresses two fundamental challenges in modern digital media: the detection of unreliable content and the assessment of source credibility. Through the development and implementation of sophisticated analytical tools and trust evaluation mechanisms, we have demonstrated the feasibility of creating automated systems that can assist users in navigating the complex landscape of online news.

The experimental validation of the system using the Kaggle dataset, comprising diverse news articles across multiple domains, confirms the framework’s effectiveness in real-world scenarios. The system’s ability to maintain high accuracy while processing content from various sources and topics suggests its potential for broader application in combating online misinformation. However, several areas merit further investigation as this field continues to evolve. Future research might explore: The integration of more sophisticated natural language processing techniques to enhance detection accuracy across different writing styles and content formats. The adaptation of trust evaluation metrics to account for emerging forms of misinformation and changing patterns of online news consumption. The development of more efficient blockchain implementations that might better balance security requirements with performance needs. These opportunities for future work notwithstanding, our current implementation represents a significant step forward in automated news verification and source credibility assessment. By combining sophisticated content analysis with dynamic trust evaluation, we have created a framework that can effectively assist users in identifying reliable information sources while adapting to the evolving challenges of online misinformation.

5.5 A Framework for Fallacy Detection

The challenge of identifying fallacies in news articles presents two fundamental problems: determining the exact boundaries of fallacious statements within the text and classifying the type of logical fallacy present. To formalise this problem, we define a news article as a sequence of words.

Let a_i represent a news article consisting words sequence $\mathcal{W}_i = \{w_1, \dots, w_j, \dots, w_{m_i}\}$, where m_i is the number of words in a_i . Within this sequence, a fallacious statement may span any continuous subsequence of words. We define a fallacious segment as $\mathcal{F}_{ijk} = \{w_j, w_{j+1}, \dots, w_k\}$, where j and k represent the starting and ending positions of the fallacy within article a_i .

The task of fallacy detection then becomes determining for each possible subsequence whether it contains a complete fallacious argument. This is particularly challenging because fallacious reasoning often relies on context and may not be confined to neat grammatical boundaries. A fallacious argument might begin mid-sentence or span multiple sentences, requiring analysis of both local and broader textual context.

For each identified fallacious segment $\mathcal{F}_{ijk}$, we must then determine its type from a set of possible fallacy categories $\mathcal{C} = \{c_1, c_2, \dots, c_n\}$. This classification task assigns to each fallacious segment a label $y_{ijk} \in \mathcal{C}$ indicating the specific type of logical fallacy present. The classification must consider not only the text within the fallacious segment but also its relationship to the broader context of the article.

The complete task can be formalized as finding a set of tuples $\mathcal{T}_i = \{(\mathcal{F}_{ijk}, y_{ijk})\}$ for each article a_i , where each tuple identifies both the boundaries and type of a fallacious segment. This framing captures the interdependent nature of the detection and classification tasks, accurate boundary detection often requires understanding the type of fallacy present, while proper classification relies on correctly identifying the complete fallacious argument.

This formulation acknowledges several key challenges inherent in fallacy detection. First, fallacious segments may overlap or be nested within one another, requiring careful consideration of how to represent and identify such relationships. Second, the boundaries of fallacious segments may be ambiguous, particularly when dealing with implicit premises or complex rhetorical structures. Finally, the classification task must account for the fact that a single piece of text might exhibit characteristics of multiple fallacy types, necessitating a nuanced approach to classification.

Evaluating the effectiveness of fallacy detection within news articles presents a unique challenge. It's not enough to simply identify the presence of fallacies; a robust solution must also accurately pinpoint their boundaries within the text and correctly classify the specific type of fallacy employed. This dual requirement significantly increases the complexity of the task, distinguishing it from simpler problems like general text classification or basic segment detection.

To illustrate this challenge, 5.12 visually represents the problem we aim to solve. The figure shows the process of extracting a fallacious statement from a news article and then classifying it according to its specific type.

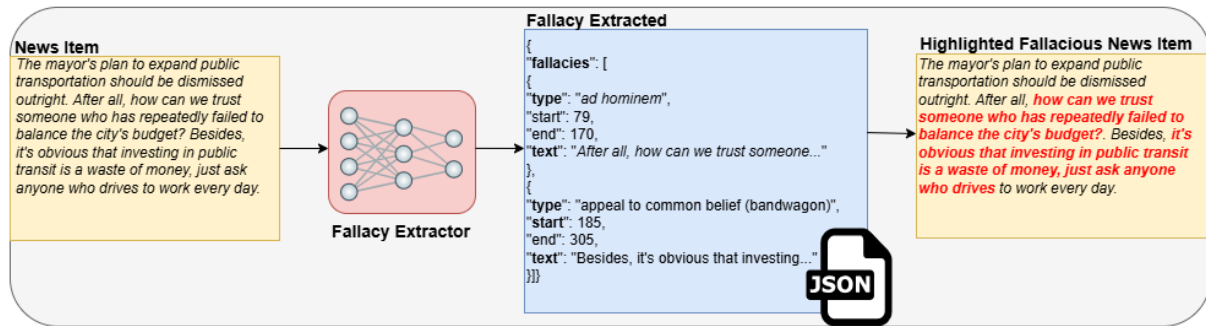


Figure 5.12: Fallacy Extraction and Detection

5.5.1 Challenges in Dataset Availability

The development of automated fallacy detection systems has historically been hampered by the scarcity of high-quality annotated datasets. This limitation stems from several factors inherent to the nature of logical fallacies: First, the identification of logical fallacies requires expert knowledge and careful analysis, making the creation of large-scale annotated datasets both time-consuming and expensive. Unlike more straightforward classification tasks where annotations can be crowdsourced, fallacy identification often requires trained analysts who understand the nuances of argumentative structure and logical reasoning. Second, there exists significant subjectivity in identifying the exact boundaries of fallacious arguments within the text. As discussed by Fleisig et al. [138], even expert annotators often have differing views on creating annotations. This inherent ambiguity makes it difficult to create reliable ground truth datasets. Third, logical fallacies can manifest in complex, context-dependent ways that make them difficult to categorise definitively. A single piece of text might exhibit characteristics of multiple fallacy types, or the classification might depend on broader contextual information not contained within the immediate text segment. The context-dependent nature of logical fallacies presents a third major obstacle. Fallacious reasoning often relies on implicit assumptions or requires understanding broader contextual information not contained within the immediate text. A single piece of text might exhibit characteristics of multiple fallacy types, or its classification might depend on understanding the broader discourse context in which it appears. This complexity makes it difficult to create datasets that capture all relevant contextual information needed for accurate fallacy identification.

5.5.2 Implementation Framework and Methodology

The development of our fallacy detection system follows a systematic approach comprising three main phases: data augmentation, fine-tuning process, and inference, as illustrated in Figure 5.13. This framework implements sophisticated processing techniques while addressing the fundamental challenges of fallacy detection in news articles.

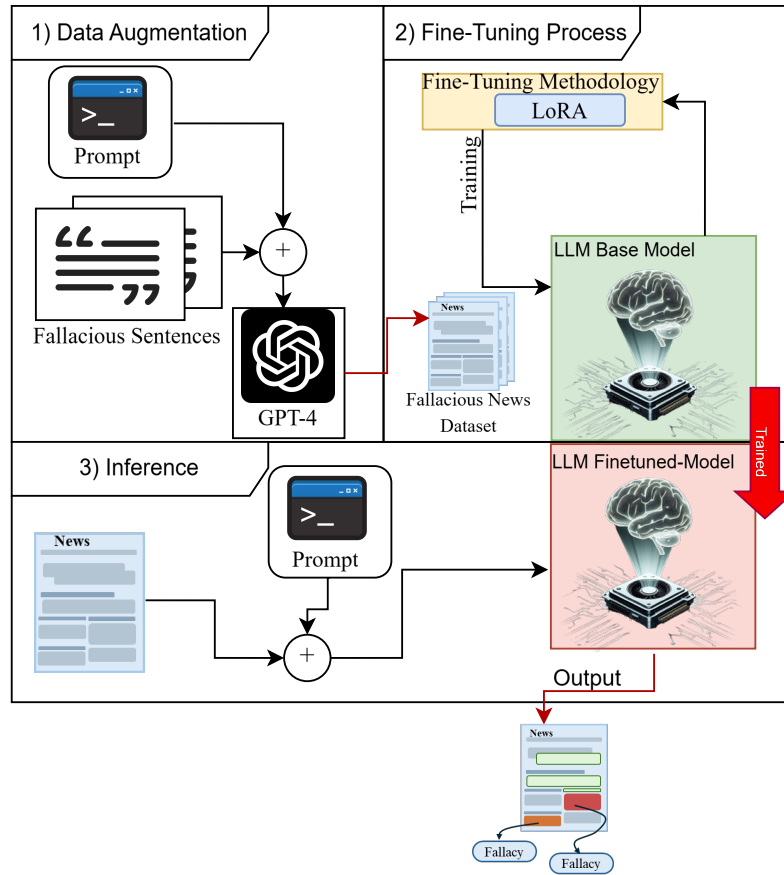


Figure 5.13: System Overview

5.5.3 Dataset Construction and Model Adaptation

To address these challenges while leveraging the capabilities of LLMs, we developed an innovative approach combining existing annotated data with synthetic augmentation. Our foundation dataset builds upon the work of Sourati et al. [105], which provides 3,425 annotated fallacious statements. Table 5.9 presents the comprehensive categorisation of fallacy types used in our research, along with their descriptions and typical manifestations in news content.

The composition of this foundational dataset reveals significant insights into the distribution of logical fallacies within news content. As shown in Figure 5.14, the dataset exhibits a natural imbalance across different fallacy types, reflecting both the inherent characteristics of news media discourse and the varying complexity of different fallacious reasoning patterns. Faulty Generalization emerges as the predominant category, accounting for 17.67% of the dataset. This prevalence aligns with common journalistic practices where limited evidence or isolated incidents are often used to support broader conclusions. Such generalizations frequently appear in coverage of social trends, political movements, or economic analyses, where the pressure to draw comprehensive conclusions from limited data points can lead to oversimplified narratives. The second most frequent category is Ad Hominem attacks at 12.83%, followed by False Causality at 10.40%. This clustering of

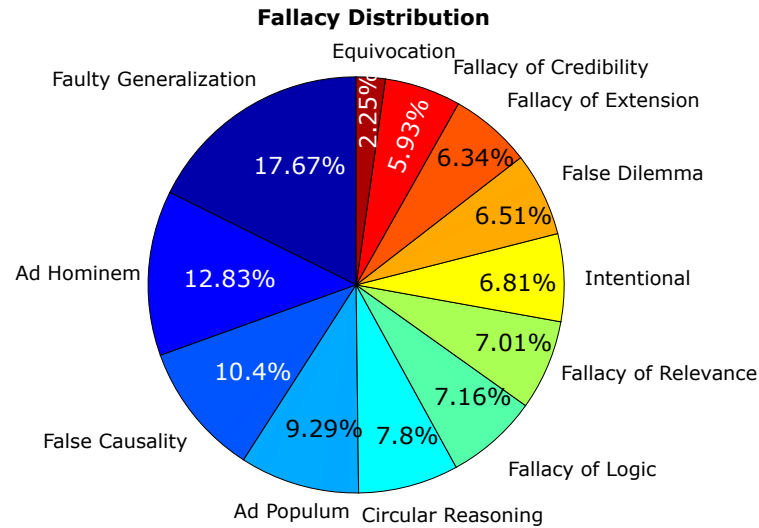


Figure 5.14: Fallacy Extraction and Detection

personal attack-based fallacies and causation errors suggests a systematic bias in news content toward attributional reasoning. The substantial presence of these fallacy types likely reflects media tendencies to personalize complex issues and establish direct causal relationships where more nuanced analysis might be warranted. The middle range of the distribution encompasses several categories including Ad Populum (9.29%), Circular Reasoning (7.80%), and various logic-based fallacies (7.16%). These moderately represented categories demonstrate the diverse ways in which logical errors manifest in news discourse, from appeals to popular opinion to more structured failures in argumentative reasoning. Notably, Equivocation appears as the least common fallacy type at 2.25%. This lower representation likely stems from the subtle nature of equivocation fallacies, which often require sophisticated linguistic analysis to identify. Their relative scarcity in the dataset presented a particular challenge for our augmentation strategy, requiring special attention to ensure adequate representation in the training data. This natural distribution informed our approach to dataset augmentation and model development. The varying representation of different fallacy types necessitated careful consideration in our synthetic data generation process to ensure balanced coverage while maintaining realistic proportions. Moreover, understanding these inherent patterns in fallacious reasoning helped guide our model architecture choices and training strategies, particularly in addressing the challenges posed by less frequent but equally important fallacy types. The foundation dataset required significant augmentation to create training examples that better reflect the complexity of real-world news articles. This augmentation process aimed to address two critical challenges: providing sufficient context for fallacy identification and ensuring the model could distinguish between fallacious reasoning and valid arguments presented in similar linguistic patterns.

Table 5.9: Categories and Descriptions of Logical Fallacies in News Content

Fallacy Type	Description	Common Manifestation in News
Ad Hominem	Attacking the character or personal traits of an opponent rather than addressing their argument	Discrediting a policy proposal by focusing on a politician’s personal life rather than the policy’s merits
Ad Populum	Appealing to widespread belief or popularity as evidence of truth	Citing public opinion polls as proof of a policy’s correctness without examining its actual effects
Circular Reasoning	Using a conclusion as a premise to support itself	Claiming a source is reliable because it has previously published accurate information, which was deemed accurate because the source reported it
Deductive	Errors in the logical structure of an argument	Drawing incorrect conclusions from statistical data through flawed reasoning chains
Equivocation	Using ambiguous terms that shift meaning within the argument	Conflating different definitions of terms like ”freedom” or ”security” to support a particular narrative
Credibility	Inappropriate appeals to authority or expertise	Citing experts in unrelated fields to support specific claims outside their expertise
Extension	Exaggerating or distorting an opponent’s position	Presenting extreme interpretations of policy proposals to make them appear unreasonable
False Dilemma	Presenting only two alternatives when others exist	Framing complex policy decisions as simple either/or choices
Causality	Assuming correlation implies causation	Attributing economic changes to specific policies without considering other factors
Faulty Generalization	Drawing broad conclusions from insufficient evidence	Using isolated incidents to make sweeping claims about groups or trends
Intentional	Deliberate manipulation of argument structure	Using emotional language or misleading statistics to obscure logical flaws
Red Herring	Introducing irrelevant issues to divert attention	Bringing up unrelated controversies to deflect from the main topic of discussion

5.5.4 Dataset Augmentation Strategy

Our augmentation strategy employed *GPT-4* to generate contextual articles around each fallacious statement from the original dataset. This approach was designed to create training examples that maintain the integrity of identified fallacies while embedding them within naturalistic news contexts. The process was carefully controlled to ensure that

Table 5.10: Statistical Analysis of Augmented Dataset

Metric	Value
Average Article Length (characters)	418.83
Average Article Length (words)	66.28
Average Fallacy Length (characters)	135.47
Average Fallacy Length (words)	24.24
Fallacy-to-Article Ratio	36.57%
Minimum Article Length (characters)	8
Minimum Article Length (words)	2
Maximum Article Length (characters)	956
Maximum Article Length (words)	161

generated content served its intended purpose without introducing unintended biases or additional fallacious reasoning. The augmentation process was guided by several key principles. First, the original fallacious statements needed to remain unaltered to preserve their validated classification. Second, the generated context needed to provide sufficient background information to make the fallacy identification task meaningful. Third, the surrounding text needed to maintain journalistic style and tone to ensure relevance for news article analysis. The augmentation process extends employs meticulously crafted prompts that balance multiple objectives: maintaining the original fallacious statement’s integrity, providing sufficient context for meaningful analysis, and ensuring adherence to journalistic conventions. This approach allows us to create a rich training environment while preserving the essential characteristics of the original fallacious statements.

To implement these objectives, we developed a specific prompt structure that guided the generation process:

Given a text, your task is to generate a short newspaper article, no more than 100 words. The article should include the entire text as a single, unbroken quote. The text should not be split into parts or altered in any way. It should fit naturally within the narrative of the article.

The prompt structure was specifically engineered to generate contextual articles that would naturally incorporate fallacious statements without introducing additional biases or unintended logical errors. The 100-word constraint served multiple purposes: ensuring computational efficiency, maintaining focus on the target fallacy, and facilitating consistent evaluation across the dataset, and facilitated consistent evaluation of the augmentation results. This constraint also helped maintain a clear relationship between the original fallacious statement and its contextual embedding. Analysis of the augmented dataset revealed several important characteristics that validate our approach. Table 5.10 presents key metrics that demonstrate the relationship between original fallacious statements and their generated contexts. The augmentation process resulted in a dataset where fallacious statements comprise approximately 36% of the total article length on average.

This proportion ensures that fallacies remain prominent enough for effective training while providing sufficient contextual information for meaningful analysis. The variation in article lengths, ranging from very concise to more extensive pieces, helps ensure the model can handle different writing styles and argument structures.

5.5.5 Model Selection and Fine-tuning Approach

The selection of an appropriate Large Language Model for fallacy detection required careful consideration of multiple factors, balancing performance capabilities with practical implementation constraints. Our evaluation process encompassed several state-of-the-art models, including *Vicuna-13b* [139], *Orca-13B* [140], and different versions of the *Llama2* family (7B, 13B) [141]. This comprehensive evaluation focused not only on raw performance metrics but also on the models' capacity to understand complex argumentative structures and maintain awareness of broader contextual relationships. Our hardware infrastructure significantly influenced the model selection process. The research environment provided two NVIDIA A100 GPUs with a combined video memory capacity of 40GB, establishing clear constraints on model size and computational requirements. While larger models such as *Llama2-70B* potentially offered superior performance characteristics, their resource requirements exceeded our available infrastructure. This limitation necessitated a careful balance between model capability and practical implementation. After extensive evaluation and testing, we selected *Llama2-13B* as our base model. This choice was supported by three key advantages that particularly suited our application:

Memory Efficiency: The model's architecture allows for effective deployment on our available GPU infrastructure while maintaining sufficient capacity for fine-tuning operations. This efficiency proves crucial given our hardware constraints, enabling us to implement sophisticated detection mechanisms without compromising computational capability.

Computational Performance: Initial testing demonstrated that *Llama2-13B* achieves comparable or superior performance to larger models when specifically tuned for fallacy detection while requiring significantly fewer computational resources. This balance of performance and efficiency makes it particularly suitable for our specialised task of fallacy detection in news articles.

Fine-tuning Adaptability: The model's architecture proves particularly amenable to Parameter-Efficient Fine-Tuning (PEFT) techniques, specifically Low-Rank Adaptation (LoRA). This adaptability allows for efficient specialisation to our specific task without requiring extensive computational resources while maintaining the model's core language understanding capabilities.

Table 5.11: Fine-tuning Hyperparameters

Parameter	Value	Rationale
Learning Rate	$2e - 4$	Optimized for stable convergence
Batch Size	2	Constrained by GPU memory
Optimizer	AdamW	Memory-efficient variant
Weight Decay	0.001	Prevents overfitting
Max Gradient Norm	0.3	Ensures training stability

The fine-tuning process was implemented using the PEFT library from Hugging Face alongside the Transformers library. Starting from the hyperparameters configuration suggested in the original Vicuna paper [142], we adapted them to optimise performance for our available hardware resources while maintaining training effectiveness. The most critical parameters are detailed in Table 5.11. These parameters were carefully tuned based on our hardware constraints and experimental results. For instance, the relatively small batch size of 2 was chosen to accommodate our GPU memory limitations while ensuring stable training. The selection of AdamW as our optimizer represents a compromise between memory efficiency and optimization effectiveness. This configuration proved effective in practice, allowing us to achieve fine-tuning results while maintaining reasonable computational requirements.

5.5.6 Inference Phase

The inference phase represents the practical deployment of our fine-tuned model for fallacy detection in real-world applications. This process involves transforming raw input text into structured output that identifies both the location and classification of fallacious arguments. We designed our inference pipeline to maintain high accuracy while ensuring computational efficiency, enabling practical application to news article analysis. Our implementation employs a text preprocessing stage that standardizes input format while preserving the contextual nuances essential for accurate fallacy detection. This preprocessing maintains a careful balance between normalizing the input for consistent processing and retaining critical linguistic markers and argumentative structures. The system implements robust text cleaning procedures capable of handling various input formats while ensuring that essential features for fallacy detection remain intact. Following preprocessing, the system analyzes the text using a sliding window approach, enabling the model to maintain awareness of broader context while focusing on specific segments for detailed examination. This approach proves particularly effective for identifying fallacious arguments that span multiple sentences or require understanding of broader textual context for proper classification. The window size parameters were determined through experimental validation to optimize the balance between computational efficiency and detection accuracy. The classification stage implements our taxonomy of logical fallacies, producing structured output that

specifies both the location and type of detected fallacies. This process integrates local textual features with broader contextual information to ensure accurate categorization. The system maintains a probabilistic approach to classification, assigning confidence scores to each detected fallacy to support more nuanced decision-making in practical applications.

5.5.7 Performance Analysis and Results

The evaluation of our fallacy detection approach requires a comprehensive assessment framework that can capture both the accuracy of identification and the quality of extracted fallacious statements. In this section, we first present the evaluation metrics employed to measure system performance, followed by a detailed analysis of our experimental results.

Evaluation Metrics Our assessment framework employs four complementary evaluation metrics, each designed to capture different aspects of model performance in fallacy detection. These metrics provide a multi-faceted evaluation approach that considers both semantic accuracy and structural precision in identifying fallacious statements.

The BLEU (Bilingual Evaluation Understudy) metric evaluates the precision of extracted fallacious statements through n-gram comparison:

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

where BP represents the brevity penalty, p_n denotes the n-gram precision, and w_n indicates the weights for different n-gram levels. While originally developed for machine translation evaluation, BLEU provides crucial insights into how accurately models capture the exact wording of fallacious statements in our context.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation), specifically ROUGE-L, focuses on sequence-level matching:

$$ROUGE-L = \frac{LCS(X, Y)}{\text{length}(Y)}$$

where X represents the system output and Y denotes the reference annotation. This approach proves particularly valuable in fallacy detection as it accommodates non-consecutive matches while maintaining sensitivity to word order, essential for evaluating extracted fallacious segments that might contain intervening non-fallacious text.

The METEOR (Metric for Evaluation of Translation with Explicit ORdering) metric enhances our evaluation through sophisticated semantic matching capabilities:

$$METEOR = (1 - Pen) \cdot \frac{10PR}{R + 9P}$$

where P represents precision, R denotes recall, and Pen indicates a penalty term for

Table 5.12: Performance comparison of base models

Metric	Vicuna-13b	Orca-13b	Llama2-13b
BLEU	0.41	0.44	0.48
ROUGE	0.47	0.49	0.53
METEOR	0.45	0.48	0.51
BERT-SCORE	0.88	0.90	0.92

Table 5.13: Performance improvement through fine-tuning

Metric	Llama2 Base	Llama2 Fine-Tuned
BLEU	0.48	0.73
ROUGE	0.53	0.76
METEOR	0.51	0.74
BERT-SCORE	0.92	0.95

fragmented matches. METEOR’s ability to consider synonyms and paraphrases proves especially valuable in identifying semantically equivalent fallacious arguments expressed through different linguistic constructions.

BERT-Score complements these metrics by providing deep semantic evaluation capabilities:

$$BERT\text{-}Score = F_1(cos(x_i, y_j))$$

where $cos(x_i, y_j)$ represents the cosine similarity between contextual embeddings. This metric proves particularly valuable in capturing semantic similarities that might be missed by surface-level comparison methods, offering insights into the models’ understanding of complex argumentative structures.

Experimental Results Our experimental evaluation comprised two main phases: an initial comparison of base models to identify the most suitable architecture, followed by an assessment of the improvements achieved through our fine-tuning approach. Table 5.12 presents the comparative results of our initial evaluation across three state-of-the-art language models: *Vicuna-13B* [139], *Orca-13b* [140], and *Llama2-13B* [141]:

The results demonstrate Llama2-13b’s superior performance across all evaluation metrics in its base configuration. This superiority manifests particularly in semantic understanding, as evidenced by the higher BERT-Score (0.92), indicating better grasp of contextual relationships and argumentative structures even before task-specific optimization.

Building upon these baseline results, we implemented our fine-tuning approach using the Parameter-Efficient Fine-Tuning (PEFT) library and LoRA techniques. Table 5.13 presents the performance improvements achieved through fine-tuning:

The fine-tuned model demonstrates substantial improvements across all metrics. The

BLEU score increase from 0.48 to 0.73 indicates significantly enhanced precision in identifying and extracting fallacious statements. The ROUGE score improvement to 0.76 suggests better recall in capturing the complete extent of fallacious content, while the METEOR score increase to 0.74 demonstrates enhanced semantic matching capabilities. Particularly noteworthy is the improvement in BERT-Score from an already strong 0.92 to 0.95, indicating that fine-tuning further enhanced the model’s semantic understanding capabilities while preserving its general language comprehension abilities.

These consistent improvements across all metrics validate both our choice of *Llama2-13b* as the base model and our fine-tuning methodology. The results demonstrate the effectiveness of our approach in enhancing fallacy detection capabilities while maintaining robust semantic understanding, confirming the viability of our framework for practical applications in news content analysis.

Chapter 6

Conclusions and Future Work

This thesis has addressed fundamental challenges in the development of Human Digital Twin systems, particularly focusing on three critical aspects: the organisation of conversational knowledge, efficient information processing, and response reliability. Our investigation has led to significant advancements in how digital systems can understand, process, and interact with human users while maintaining high standards of privacy and security.

The motivation for this research emerged from the growing complexity of human-digital interaction systems and the need for approaches to processing and understanding human communication. Traditional approaches to digital twin technology, while effective for physical systems, proved inadequate for capturing the nuanced nature of human knowledge and behaviour. This limitation necessitated the development of new frameworks capable of handling the complexity of human communication while maintaining computational efficiency and ensuring reliable system responses.

The research has made several significant contributions to the field of Human Digital Twin technology:

1. The development of a novel topic modelling framework that significantly reduces computational overhead while maintaining high classification accuracy. This framework, through its innovative key sentence extraction algorithm and attention-based neural architecture, enables efficient processing of large-scale conversational data without compromising semantic understanding.
2. The creation of a comprehensive HDT architecture implementing sophisticated mechanisms for knowledge acquisition and temporal processing. This architecture, centred around the Digital Brain concept, demonstrates how digital systems can maintain evolving models of user interests and preferences while ensuring privacy protection and adaptive learning capabilities.
3. The establishment of new methodologies for ensuring information reliability in HDT

systems, exemplified through our case study of fake news detection and fallacy identification. This work provides crucial insights into the parallel challenges of managing both machine-generated hallucinations and human-generated misinformation.

6.1 Key Research Outcomes

6.1.1 Advancements in Topic Classification

Our research in topic classification has produced significant advancements in both theoretical understanding and practical implementation of conversational data processing. The development of our KSE algorithm represents a fundamental breakthrough in efficient content processing, achieving a 34% reduction in training time and a 66% decrease in testing time while maintaining or improving classification accuracy. The success of this approach stems from several key innovations. First, our implementation of selective content processing through the KSE algorithm demonstrates that intelligent feature selection can actually enhance classification accuracy while reducing computational overhead. This counterintuitive finding challenges traditional assumptions about the relationship between processing complexity and classification performance. Furthermore, our comparative analysis with Large Language Models revealed surprising insights into the relative strengths of specialised versus general-purpose architectures. While LLMs like GPT-3.5 achieved impressive accuracy rates of 92%, our specialised framework maintained superior performance at 99.7% accuracy while requiring significantly fewer computational resources. This finding has important implications for developing efficient, task-specific architectures in HDT systems.

6.1.2 Human Digital Twin Framework Evolution

The development of our HDT framework, particularly the Digital Brain architecture, represents a significant advancement in human-digital interaction systems. Our implementation demonstrates how sophisticated knowledge representation and processing mechanisms can enable digital systems to maintain evolving user behaviour and preferences models while ensuring privacy protection. A key innovation in our approach lies in the implementation of temporal dynamics within the Digital Brain architecture. The system's ability to process and integrate information across multiple time scales enables more nuanced understanding of user behaviour patterns. This capability proves particularly valuable in practical applications where system responses must account for both immediate context and longer-term user preferences. The privacy-aware architecture implemented in our framework establishes new standards for protecting user data while maintaining system functionality. This work has shown how HDT systems can provide personalised assistance while preserving user privacy by carefully integrating encryption mechanisms and selective data processing. This

balance between functionality and privacy protection represents a crucial advancement in human-centric digital systems.

6.1.3 Information Reliability Framework

Our investigation into information reliability, conducted through fake news detection and fallacy identification, has produced valuable insights into ensuring trustworthy system responses. The parallel identified between AI hallucinations and human-generated misinformation suggests common patterns in how false information maintains apparent plausibility while deviating from truth. Developing our multi-faceted evaluation approach, incorporating message-based, fact-based, and context-based analysis, demonstrates the effectiveness of comprehensive verification strategies. Our experimental results show that this integrated approach achieves superior detection accuracy, while providing explainable reasoning for its classifications. Particularly noteworthy is our framework's ability to maintain high accuracy across different types of content and various deception strategies. The system's resilience to sophisticated manipulation attempts, demonstrated through our analysis of coordinated disinformation campaigns, validates our architectural decisions and parameter choices.

6.2 Future Research Directions

The contributions presented in this thesis establish foundations for several promising research directions. While our work has demonstrated significant advancements in Human Digital Twin technology, it also reveals opportunities for further development across multiple domains. This section explores key areas where future research could extend and enhance the frameworks we have developed.

6.2.1 Technical Enhancement Opportunities

While demonstrating superior performance compared to existing approaches, the topic classification framework developed in this research suggests several avenues for technical enhancement. One significant opportunity lies in developing more sophisticated sentence selection mechanisms that could further improve processing efficiency while maintaining semantic accuracy. Research in this direction might explore adaptive threshold techniques that automatically adjust to different types of content and varying document lengths. Integrating more advanced neural architectures represents another promising direction for technical development. While our current implementation of BiLSTM networks with attention mechanisms has proven highly effective, emerging architectures such as sparse attention models and efficient transformers could offer even better performance-efficiency trade-offs. Investigation into these architectures, particularly focusing on their application

to conversational data processing, could yield valuable improvements in both accuracy and computational efficiency. Furthermore, the parallel processing capabilities of our framework could be enhanced to better handle multi-modal input streams. Future research might explore architectures capable of simultaneously processing text, speech, and behavioural data while maintaining real-time performance. This enhancement would be particularly valuable for HDT systems that must integrate multiple input modalities to maintain comprehensive user understanding.

6.2.2 Knowledge Representation and Processing

Our work with the Digital Brain architecture reveals significant opportunities for advancing knowledge representation and processing capabilities. Future research could explore more sophisticated approaches to temporal knowledge integration, particularly focusing on how systems can better understand and predict evolving patterns of user behavior. This might involve developing new models for representing temporal dependencies that capture both short-term variations and long-term trends in user interests and preferences. The development of more sophisticated privacy preservation mechanisms represents another crucial area for future investigation. While our current implementation demonstrates effective privacy protection, emerging technologies such as federated learning and homomorphic encryption offer possibilities for even stronger privacy guarantees. Research into these technologies could lead to frameworks that enable more sophisticated data sharing and collaboration while maintaining robust privacy protection. Particularly promising is the potential development of adaptive knowledge structures that can automatically reorganize themselves based on evolving usage patterns. Such structures could potentially improve both system efficiency and user experience by maintaining optimal organization of information based on actual usage patterns rather than predetermined hierarchies.

6.2.3 Information Reliability and Trust

The findings from our investigation into fake news detection and fallacy identification suggest several important directions for future research in information reliability. One significant opportunity lies in developing more sophisticated approaches to understanding the relationship between different types of misleading information. This research could potentially lead to unified frameworks for detecting and mitigating both machine-generated hallucinations and human-generated misinformation. Future work might also explore more advanced trust evaluation mechanisms that can better account for contextual factors and evolving patterns of deception. This could include development of:

1. More sophisticated temporal analysis techniques for tracking source reliability over time

2. Enhanced methods for detecting coordinated manipulation attempts across multiple sources
3. Improved frameworks for evaluating the credibility of novel information sources

Additionally, the relationship between content reliability and user trust presents intriguing opportunities for future research. Understanding how users develop and maintain trust in digital systems, particularly in the context of HDT interactions, could inform the development of more effective trust management mechanisms.

6.2.4 Application Domains and Scaling

The frameworks developed in this research demonstrate potential for application across various domains beyond news content analysis. Future research might explore adaptation of these technologies to areas such as:

- Educational Systems: Investigating how HDT technology could enhance personalised learning experiences while maintaining reliable content delivery.
- Healthcare Applications: Exploring the potential for HDT systems to support health-care delivery while ensuring strict privacy protection and information reliability.
- Professional Development: Studying how HDT frameworks could support continuous learning and skill development in professional contexts.

Each of these domains presents unique challenges and opportunities that could drive further advancement of the core technologies developed in this research. Particularly important will be understanding how these systems can scale effectively while maintaining performance and reliability across different application contexts.

6.2.5 Theoretical Foundations

Finally, our work suggests several important directions for theoretical development in HDT technology. Future research might explore more formal models for understanding the relationship between human cognition and digital representation, potentially leading to more sophisticated frameworks for human-digital interaction. This could include the development of new theoretical frameworks for understanding how digital systems can effectively model and respond to human behaviour while maintaining appropriate boundaries and limitations. The theoretical foundations of privacy-aware computing in HDT contexts also merit further investigation. Research in this area might explore fundamental limits on what can be achieved while maintaining strong privacy guarantees, potentially leading to new paradigms for privacy-preserving computation in human-digital interaction systems. These future research directions collectively suggest rich opportunities for advancing the

HDT technology field. The frameworks and methodologies developed in this thesis provide solid foundations for such advancements, while the challenges and opportunities identified through our work offer clear guidance for future investigation.

Appendices

Additional Research Work

While this thesis focuses on HDTs and NLP, additional research work was conducted through various collaborations during the doctoral period. These parallel research activities, while not directly related to the main thesis topic, resulted in the following contributions:

- **A Generative Adversarial Network (GAN) Fingerprint Approach Over LTE** [143]: An innovative approach to indoor/outdoor localization using LTE signals was presented. By combining fingerprinting techniques with GANs for data augmentation, a system capable of accurate position estimation was developed while reducing the cost and complexity of traditional fingerprinting methods. The research demonstrated the potential of applying machine learning techniques to enhance wireless localization systems.
- **Robust QUIC-Based Signalling for WebRTC in Impaired Networks** [144]: The impact of using the QUIC protocol for signalling in WebRTC applications was investigated. A novel QUIC-based signalling server was proposed and demonstrated superior performance compared to traditional WebSocket-based solutions, particularly in challenging network conditions. The work showcased how QUIC's features could enhance WebRTC's reliability while maintaining efficient network resource utilisation.
- **The Impact of Network Impairments on the QoE of WebRTC Applications** [145]: Extensive subjective tests were conducted to evaluate how network impairments affect user Quality of Experience in WebRTC-based video communications. Through experimental design and analysis of user feedback, valuable insights were gained into the relationship between technical network parameters and perceived quality. The research highlighted the importance of considering human factors in communication system design.

These collaborative projects contributed to a broader understanding of modern communication systems, from protocol design to quality assessment and localization techniques.

Appendix A

Acronyms

ACR Absolute Category Rating

AI Artificial Intelligence

ASR Automatic Speech Recognition

AUC Area Under Curve

BiLSTM Bidirectional Long Short Term Memory

BoW Bag of Words

CNN Convolutional Neural Network

COVID-19 COronaVirus Disease 2019

DH Digital Human

DLP Data Loss Prevention

DT Digital Twin

GAN Graph Attention Network

GDPR General Data Protection Regulation

GN60K Global News 60K

GNN Graph Neural Network

HDT Human Digital Twin

IoT Internet of Things

KG Knowledge Graph

KGE Knowledge Graph Embboding

KNN K-Nearest Neighbours

KRL Knowledge Representation Learning

KSE Key Sentence Extractor

LDA Latent Dirichlet Allocation

LLM Large Language Model

LLMs Large Language Models

LoRA Low-Rank Adaptation

LSVM Linear SVM

LSTM Long Short Term Memory

ML Machine Learning

NBC Naive Bayes Classifier

NLP Natural Language Processing

NLU Natural Language Understanding

PDA Personal Digital Assistant

OOA On-Off Attack

PEFT Parameter-Efficient Fine-Tuning

PII Personal Identifiable Information

QoE Quality of Experience

RAG Retrieval-Augmented Generation

RDF Resource Description Framework

ReLU REctified Linear Unit

RF Random Forest

ROC Receiver Operating Characteristic

SAM Self-Assessment Manikin

SVM Support Vector Machine

TF-IDF Term Grequency–Inverse Document Frequency

TKG Temporal Knowledge Graph

WWA WhiteWashing Attack

XAI eXplainable AI

XGBoost Extreme Gradient Boosting

Acknowledges

This work has been partially funded by the Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) with the PON “Ricerca e Innovazione” 2014-2020 (PON RI) “Azione IV.5 Dottorati su tematiche green”, assigned with D.M. 1062 on 10.08.2021 and by the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 1409 published on 14.9.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU – Project Title “METATwin - Metaverse & Human Digital Twin: digital identity, Biometrics and Privacy in the future virtual worlds”, – CUP J53D23015030001- Grant Assignment Decree No. 0001382 adopted on 01/09/2023 by the Italian Ministry of University and Research (MUR).

Bio

Luigi Serreli conducts research in Natural Language Processing at the Department of Electrical and Electronic Engineering, University of Cagliari. After receiving his Master's Degree in Internet Technologies Engineering in 2021, he has focused his research on the development of advanced natural language processing systems, with particular emphasis on topic modelling and its applications in disinformation detection. His work at ATC Innovation Lab (ILab) in Athens involves the development and optimization of Large Language Models within European Union's project. Through this collaboration, he contributes to the advancement of AI-powered solutions for tackling modern challenges in digital communication and information reliability. His research interests span the intersection of natural language processing, artificial intelligence, and the Internet of Things, focusing particularly on developing intelligent systems that can effectively process and analyse textual information.

As a member of IEEE and the Italian National Consortium for Telecommunications (CNIT), he actively participates in the scientific community's efforts to advance these fields through both theoretical research and practical applications.

This work represents a contribution to the ongoing development of more sophisticated and reliable systems for natural language processing and information analysis, particularly in the context of modern digital communication challenges.

Bibliography

- [1] Michael W Grieves. “Product lifecycle management: the new paradigm for enterprises”. In: *International Journal of Product Development* 2.1-2 (2005), pp. 71–84.
- [2] Michael Grieves and John Vickers. “Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems”. In: *Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches*. Ed. by Franz-Josef Kahlen, Shannon Flumerfelt, and Anabela Alves. Cham: Springer International Publishing, 2017, pp. 85–113. ISBN: 978-3-319-38756-7. DOI: 10.1007/978-3-319-38756-7_4. URL: https://doi.org/10.1007/978-3-319-38756-7_4.
- [3] Edward Glaessgen and David Stargel. “The digital twin paradigm for future NASA and US Air Force vehicles”. In: *53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference 20th AIAA/ASME/AHS adaptive structures conference 14th AIAA*. 2012, p. 1818.
- [4] Mohsen Attaran and Bilge Gokhan Celik. “Digital Twin: Benefits, use cases, challenges, and opportunities”. In: *Decision Analytics Journal* 6 (2023), p. 100165. ISSN: 2772-6622. DOI: <https://doi.org/10.1016/j.dajour.2023.100165>. URL: <https://www.sciencedirect.com/science/article/pii/S277266222300005X>.
- [5] Mengnan Liu et al. “Review of digital twin about concepts, technologies, and industrial applications”. In: *Journal of Manufacturing Systems* 58 (2021). Digital Twin towards Smart Manufacturing and Industry 4.0, pp. 346–361. ISSN: 0278-6125. DOI: <https://doi.org/10.1016/j.jmsy.2020.06.017>. URL: <https://www.sciencedirect.com/science/article/pii/S0278612520301072>.
- [6] Baicun Wang et al. “Human Digital Twin in the context of Industry 5.0”. In: *Robotics and Computer-Integrated Manufacturing* 85 (2024), p. 102626. ISSN: 0736-5845. DOI: <https://doi.org/10.1016/j.rcim.2023.102626>. URL: <https://www.sciencedirect.com/science/article/pii/S0736584523001011>.
- [7] Ji Zhou et al. “Human–Cyber–Physical Systems (HCPSs) in the Context of New-Generation Intelligent Manufacturing”. In: *Engineering* 5.4 (2019), pp. 624–636.

- ISSN: 2095-8099. DOI: <https://doi.org/10.1016/j.eng.2019.07.015>. URL: <https://www.sciencedirect.com/science/article/pii/S2095809919306514>.
- [8] Yehor Karpichev et al. “Extended Reality for Enhanced Human-Robot Collaboration: a Human-in-the-Loop Approach”. In: *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. 2024, pp. 1991–1998. DOI: 10.1109/RO-MAN60168.2024.10731170.
 - [9] Hendrik Stern and Till Becker. “Concept and Evaluation of a Method for the Integration of Human Factors into Human-Oriented Work Design in Cyber-Physical Production Systems”. In: *Sustainability* 11.16 (2019). ISSN: 2071-1050. DOI: 10.3390/su11164508. URL: <https://www.mdpi.com/2071-1050/11/16/4508>.
 - [10] Hossein Hassani, Xu Huang, and Emmanuel Silva. “The Human Digitalisation Journey: Technology First at the Expense of Humans?” In: *Information* 12.7 (2021). ISSN: 2078-2489. DOI: 10.3390/info12070267. URL: <https://www.mdpi.com/2078-2489/12/7/267>.
 - [11] Dominik Dellermann et al. *The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems*. 2021. arXiv: 2105.03354 [cs.AI]. URL: <https://arxiv.org/abs/2105.03354>.
 - [12] Pamela Ruiz Castro et al. “Virtual simulation of human-robot collaboration workstations”. In: *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume V: Human Simulation and Virtual Environments, Work With Computing Systems (WWCS), Process Control 20*. Springer. 2019, pp. 250–261.
 - [13] Weisheng Lu et al. “Digital twin-enabled human-robot collaborative teaming towards sustainable and healthy built environments”. In: *Journal of Cleaner Production* 412 (2023), p. 137412. ISSN: 0959-6526. DOI: <https://doi.org/10.1016/j.jclepro.2023.137412>. URL: <https://www.sciencedirect.com/science/article/pii/S0959652623015706>.
 - [14] “Smart Manufacturing and Intelligent Manufacturing: A Comparative Review”. In: *Engineering* 7.6 (2021), pp. 738–757. ISSN: 2095-8099. DOI: <https://doi.org/10.1016/j.eng.2020.07.017>.
 - [15] Amira Bourechak et al. “At the Confluence of Artificial Intelligence and Edge Computing in IoT-Based Applications: A Review and New Perspectives”. In: *Sensors* 23.3 (2023). ISSN: 1424-8220. URL: <https://www.mdpi.com/1424-8220/23/3/1639>.
 - [16] Achim Buerkle et al. “An adaptive human sensor framework for human–robot collaboration”. In: *The International Journal of Advanced Manufacturing Technology* (2022), pp. 1–16.

- [17] Elias Montini et al. “An IIoT Platform For Human-Aware Factory Digital Twins”. In: *Procedia CIRP* 107 (2022). Leading manufacturing systems transformation – Proceedings of the 55th CIRP Conference on Manufacturing Systems 2022, pp. 661–667. ISSN: 2212-8271. DOI: <https://doi.org/10.1016/j.procir.2022.05.042>. URL: <https://www.sciencedirect.com/science/article/pii/S2212827122003262>.
- [18] Oluwaseun Kolade and Adebawale Owoseni. “Employment 5.0: The work of the future and the future of work”. In: *Technology in Society* 71 (2022), p. 102086. ISSN: 0160-791X. DOI: <https://doi.org/10.1016/j.techsoc.2022.102086>. URL: <https://www.sciencedirect.com/science/article/pii/S0160791X22002275>.
- [19] Duansen Shangguan et al. “A Triple Human-Digital Twin Architecture for Cyber-Physical Systems.” In: *CMES-Computer Modeling in Engineering & Sciences* 131.3 (2022).
- [20] Tsubasa Maruyama et al. “Digital Twin-Driven Human Robot Collaboration Using a Digital Human”. In: *Sensors* 21.24 (2021). ISSN: 1424-8220. DOI: [10.3390/s21248266](https://doi.org/10.3390/s21248266). URL: <https://www.mdpi.com/1424-8220/21/24/8266>.
- [21] Syed Meesam Raza Naqvi et al. “Human knowledge centered maintenance decision support in digital twin environment”. In: *Journal of Manufacturing Systems* 65 (2022), pp. 528–537. ISSN: 0278-6125. DOI: <https://doi.org/10.1016/j.jmsy.2022.10.003>. URL: <https://www.sciencedirect.com/science/article/pii/S0278612522001741>.
- [22] Meng-Lun Lee et al. “Robot-Assisted Disassembly Sequence Planning With Real-Time Human Motion Prediction”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 53.1 (2023), pp. 438–450. DOI: [10.1109/TSMC.2022.3185889](https://doi.org/10.1109/TSMC.2022.3185889).
- [23] Rong Zhang et al. “A fusion-based spiking neural network approach for predicting collaboration request in human-robot collaboration”. In: *Robotics and Computer-Integrated Manufacturing* 78 (2022), p. 102383. ISSN: 0736-5845. DOI: <https://doi.org/10.1016/j.rcim.2022.102383>. URL: <https://www.sciencedirect.com/science/article/pii/S0736584522000709>.
- [24] Baicun Wang et al. “Human digital twin (HDT) driven human-cyber-physical systems: Key technologies and applications”. In: *Chinese Journal of Mechanical Engineering* 35.1 (2022), p. 11. DOI: [10.1186/s10033-022-00680-w](https://doi.org/10.1186/s10033-022-00680-w).
- [25] Evangelia Katsoulakis et al. “Digital twins for health: a scoping review”. In: *NPJ Digital Medicine* 7.1 (2024), p. 77.
- [26] Evanthia Faliagka et al. “Trends in Digital Twin Framework Architectures for Smart Cities: A Case Study in Smart Mobility”. In: *Sensors* 24.5 (2024). DOI: [10.3390/s24051665](https://doi.org/10.3390/s24051665). URL: <https://www.mdpi.com/1424-8220/24/5/1665>.

- [27] Tambiama Madiega. “Artificial intelligence act”. In: (2021).
- [28] Peter S Park et al. “AI deception: A survey of examples, risks, and potential solutions”. In: *Patterns* 5.5 (2024).
- [29] Kazuma Kobayashi and Syed Bahauddin Alam. “Explainable, interpretable, and trustworthy AI for an intelligent digital twin: A case study on remaining useful life”. In: *Engineering Applications of Artificial Intelligence* 129 (Mar. 2024), p. 107620. ISSN: 0952-1976. DOI: 10.1016/j.engappai.2023.107620. URL: <http://dx.doi.org/10.1016/j.engappai.2023.107620>.
- [30] Elias Montini et al. “An IIoT Platform For Human-Aware Factory Digital Twins”. In: *Procedia CIRP* 107 (2022), 661â–667. ISSN: 2212-8271. DOI: 10.1016/j.procir.2022.05.042. URL: <http://dx.doi.org/10.1016/j.procir.2022.05.042>.
- [31] Luis H. Favela and Mary Jean Amon. “The ethics of human digital twins: Counterfeit people, personhood, and the right to privacy”. In: *2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI)*. 2023, pp. 1–6. DOI: 10.1109/DTPI59677.2023.10365409.
- [32] A. Arora et al. “Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization”. In: *Information and Organization* 33.3 (Sept. 2023), p. 100478. ISSN: 1471-7727. DOI: 10.1016/j.infoandorg.2023.100478. URL: <http://dx.doi.org/10.1016/j.infoandorg.2023.100478>.
- [33] S. Rabah, G. Zacharewicz, and V. Chapurlat. “Digital Twin for Services (DT4S): Conceptual Strategy”. In: *IFAC-PapersOnLine* 55.10 (2022), pp. 3256–3261. ISSN: 2405-8963. DOI: 10.1016/j.ifacol.2022.10.136. URL: <http://dx.doi.org/10.1016/j.ifacol.2022.10.136>.
- [34] Abubakar Isah et al. “Digital Twins Temporal Dependencies-Based on Time Series Using Multivariate Long Short-Term Memory”. In: *Electronics* 12.19 (Oct. 2023), p. 4187. ISSN: 2079-9292. DOI: 10.3390/electronics12194187. URL: <http://dx.doi.org/10.3390/electronics12194187>.
- [35] Phillip Schneider et al. “From Data to Dialogue: Leveraging the Structure of Knowledge Graphs for Conversational Exploratory Search”. In: *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*. 2023, pp. 609–619.
- [36] Rong Zhang et al. “A fusion-based spiking neural network approach for predicting collaboration request in human-robot collaboration”. In: *Robotics and Computer-Integrated Manufacturing* 78 (Dec. 2022), p. 102383. ISSN: 0736-5845. DOI: 10.1016/j.rcim.2022.102383. URL: <http://dx.doi.org/10.1016/j.rcim.2022.102383>.

- [37] Zhepei Wei et al. “Towards Unified Representations of Knowledge Graph and Expert Rules for Machine Learning and Reasoning”. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2022, pp. 240–253.
- [38] Franz Georg Listl et al. “Knowledge Graphs in the Digital Twin: A Systematic Literature Review About the Combination of Semantic Technologies and Simulation in Industrial Automation”. In: *arXiv preprint arXiv:2406.09042* (2024).
- [39] Keyan Cao and Chuang Zheng. “TBRm: A Time Representation Method for Industrial Knowledge Graph”. In: *Applied Sciences* 12.22 (2022), p. 11316.
- [40] Guixian Xu et al. “Research on topic detection and tracking for online news texts”. In: *IEEE access* 7 (2019), pp. 58407–58418.
- [41] Shaoor Munir et al. “Through the Looking Glass: Learning to Attribute Synthetic Text Generated by Language Models”. In: Association for Computational Linguistics, Apr. 2021, pp. 1811–1822. DOI: 10.18653/v1/2021.eacl-main.155.
- [42] S. Selva Birunda and R. Kanniga Devi. “A Review on Word Embedding Techniques for Text Classification”. In: *Innovative Data Communication Technologies and Application* (2021), pp. 267–281. DOI: 10.1007/978-981-15-9651-3_23.
- [43] Ranjan Kumar Behera et al. “Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data”. In: *Information Processing & Management* 58.1 (2021), p. 102435. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2020.102435.
- [44] Sruthi Yarkareddy, T. Sasikala, and S. Santhanalakshmi. “Sentiment Analysis Of Amazon Fine Food Reviews”. In: *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*. 2022, pp. 1242–1247. DOI: 10.1109/ICSSIT53264.2022.9716410.
- [45] Qianli Ma et al. “Deformable Self-Attention for Text Classification”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 1570–1581. DOI: 10.1109/TASLP.2021.3067210.
- [46] Bin Wang et al. “Evaluating word embedding models: methods and experimental results”. In: *APSIPA transactions on signal and information processing* 8 (2019).
- [47] Riadh Belkebir and Ahmed Guessoum. “A hybrid BSO-Chi2-SVM approach to Arabic text categorization”. In: *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*. IEEE. 2013.

- [48] Samrat Alam, Md Afnan Ul Haque, and Ashiqur Rahman. “Bengali Text Categorization Based on Deep Hybrid CNN–LSTM Network with Word Embedding”. In: *2022 International Conference on Innovations in Science, Engineering and Technology (ICISSET)*. 2022, pp. 577–582. DOI: 10.1109/ICISSET54810.2022.9775913.
- [49] Feng Wei and Uyen Trang Nguyen. “An Attention-Based Neural Network Using Human Semantic Knowledge and Its Application to Clickbait Detection”. In: *IEEE Open Journal of the Computer Society* 3 (2022), pp. 217–232. DOI: 10.1109/OJCS.2022.3213791.
- [50] Ran Wang et al. “A novel reasoning mechanism for multi-label text classification”. In: *Information Processing & Management* 58.2 (2021), p. 102441.
- [51] Wei Shi, Miao Song, and Yong Wang. “Perturbation-enhanced-based RoBERTa combined with BiLSTM model for Text classification”. In: *ICETIS 2022; 7th International Conference on Electronic Technology and Information Science*. 2022, pp. 1–5.
- [52] Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. “In-context learning for text classification with many labels”. In: *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*. Association for Computational Linguistics, 2023, pp. 173–184. DOI: 10.18653/v1/2023.genbench-1.14.
- [53] Momna Ali Shah et al. “An Automated Text Document Classification Framework using BERT”. In: *International Journal of Advanced Computer Science and Applications* 14.3 (2023). DOI: 10.14569/IJACSA.2023.0140332.
- [54] M. Kavitha and K. Akila. “Amplifying document categorization with advanced features and deep learning”. In: 2024. DOI: 10.1007/s11042-024-18483-7.
- [55] V Kumar Velu and Sendhilkumar Selvaraju. “Developing a conceptual framework for short text categorization using hybrid CNN- LSTM based Caledonian crow optimization”. In: *Expert Systems with Applications* 212 (2023), p. 118517. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2022.118517.
- [56] Ranju Mandal et al. “Empirical study of tweets topic classification using transformer-based language models”. In: *Intelligent Information and Database Systems: 13th Asian Conference, ACIIDS 2021, Phuket, Thailand, April 7–10, 2021, Proceedings 13*. Springer. 2021. DOI: 10.1007/978-3-030-73280-6_27.
- [57] Aleksa Vukotic et al. *Neo4j in action*. Manning Publications Co., 2014.
- [58] Traian Rebedea et al. “Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails”. In: *arXiv preprint arXiv:2310.10501* (2023).
- [59] Hakan Inan et al. “Llama guard: Llm-based input-output safeguard for human-ai conversations”. In: *arXiv preprint arXiv:2312.06674* (2023).

- [60] Gülnaziye Bingöl et al. “QoE Estimation of WebRTC-based Audiovisual Conversations from Facial Expressions”. In: *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. 2022. DOI: 10.1109/SITIS57111.2022.00092.
- [61] Ricardo Montalvo-Lezama, Berenice Montalvo-Lezama, and Gibran Fuentes-Pineda. “Improving Transfer Learning for Movie Trailer Genre Classification using a Dual Image and Video Transformer”. In: *Information Processing & Management* 60.3 (2023), p. 103343. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2023.103343.
- [62] Shweta Ganiger and K.M.M. Rajashekharaiyah. “Comparative Study on Keyword Extraction Algorithms for Single Extractive Document”. In: *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2018, pp. 1284–1287. DOI: 10.1109/ICCONS.2018.8663040.
- [63] Tabor Wegi Geleta and Jara Muda Haro. “Semisupervised Learning-Based Word-Sense Disambiguation Using Word Embedding for Afaan Oromoo Language”. In: vol. 2024. 2024, p. 11. DOI: 10.1155/2024/4429069.
- [64] Kevin Patel and Pushpak Bhattacharyya. “Towards Lower Bounds on Number of Dimensions for Word Embeddings”. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing*. Ed. by Greg Kondrak and Taro Watanabe. Asian Federation of Natural Language Processing, Nov. 2017.
- [65] K Schrøder. “What do news readers really want to read about? How relevance works for news audiences”. In: *Digital News Project* Feb 2019 (2019), pp. 1–36.
- [66] Gianna M Del Corso et al. “Ranking a stream of news”. In: *Proceedings of the 14th international conference on World Wide Web*. New York, NY, USA: Association for Computing Machinery, 2005, pp. 97–106. DOI: 10.1145/1060745.1060764.
- [67] Derek Greene and Pádraig Cunningham. “Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering”. In: *Proc. 23rd International Conference on Machine learning (ICML’06)*. ACM Press, 2006, pp. 377–384. DOI: 10.1145/1143844.1143892.
- [68] Ricardo Campos et al. “YAKE! Keyword extraction from single documents using multiple local features”. In: *Information Sciences* 509 (2020). DOI: 10.1016/j.ins.2019.09.013.
- [69] Talha Bin Sarwar et al. “Evaluating keyphrase extraction algorithms for finding similar news articles using lexical similarity calculation and semantic relatedness measurement by word embedding”. In: *PeerJ Computer Science* 8 (2022), e1024. DOI: 10.7717/peerj-cs.1024.

- [70] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186. DOI: 10.48550/arXiv.1810.04805.
- [71] Yijin Xiong et al. “Fusing label embedding into bert: An efficient improvement for text classification”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021, pp. 1743–1750. DOI: 10.18653/v1/2021.findings-acl.152.
- [72] Marc Brysbaert. “How many words do we read per minute? A review and meta-analysis of reading rate”. In: *Journal of Memory and Language* 109 (2019), p. 104047. ISSN: 0749-596X. DOI: 10.1016/j.jml.2019.104047.
- [73] Kumar Abhishek. “News Article Classification using a Transfer Learning Approach”. In: *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. 2022, pp. 1–6. DOI: 10.1109/ICRITO56286.2022.9965142.
- [74] Rana Reda Waly et al. “Empowering Topic Classification Accuracy by Integrating Transformer-Based with Machine Learning and Deep Learning Techniques”. In: *2023 Intelligent Methods, Systems, and Applications (IMSA)*. 2023, pp. 19–25. DOI: 10.1109/IMSA58542.2023.10217434.
- [75] Yujie Sun et al. “AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content”. In: *Humanities and Social Sciences Communications* 11.1 (2024), pp. 1–14.
- [76] Karishma Sharma et al. “Combating fake news: A survey on identification and mitigation techniques”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.3 (2019), pp. 1–42.
- [77] Xichen Zhang and Ali A Ghorbani. “An overview of online fake news: Characterization, detection, and discussion”. In: *Information Processing & Management* 57.2 (2020), p. 102025.
- [78] Gordon Pennycook and David G Rand. “The psychology of fake news”. In: *Trends in cognitive sciences* 25.5 (2021), pp. 388–402.
- [79] Alessandro Bondielli and Francesco Marcelloni. “A survey on fake news and rumour detection techniques”. In: *Information Sciences* 497 (2019), pp. 38–55.
- [80] Jooyeon Kim et al. “Leveraging the crowd to detect and reduce the spread of fake news and misinformation”. In: *Proceedings of the eleventh ACM international conference on web search and data mining*. 2018, pp. 324–332.

- [81] Rohit Mujumdar and Srijan Kumar. “HawkEye: a robust reputation system for community-based counter-misinformation”. In: *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2021, pp. 188–192.
- [82] Xinyi Zhou and Reza Zafarani. “A survey of fake news: Fundamental theories, detection methods, and opportunities”. In: *ACM Computing Surveys (CSUR)* 53.5 (2020), pp. 1–40.
- [83] Xinyan Li et al. “An empirical study on how well do COVID-19 information dashboards service user information needs”. In: *IEEE Transactions on Services Computing* (2021).
- [84] Vivek Singh et al. “Automated fake news detection using linguistic analysis and machine learning”. In: *International conference on social computing, behavioral-cultural modeling, & prediction and behavior representation in modeling and simulation*. 2017, pp. 1–3.
- [85] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. “Sentiment aware fake news detection on online social networks”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 2507–2511.
- [86] Hadeer Ahmed et al. “Detection of online fake news using n-gram analysis and machine learning techniques”. In: *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*. Springer. 2017, pp. 127–138.
- [87] Ahlem Drif, Z Ferhat Hamida, and Silvia Giordano. “Fake News Detection Method Based on Text-Features”. In: *The Ninth International Conference on Advances in Information Mining and Management*. 2019.
- [88] Archita Pathak and Rohini K Srihari. “BREAKING! Presenting fake news corpus for automated fact checking”. In: *Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop*. 2019, pp. 357–362.
- [89] Farzana Islam Adiba et al. “Effect of corpora on classification of fake news using naive Bayes classifier”. In: *International Journal of Automation, Artificial Intelligence and Machine Learning* 1 (2020), pp. 80–92.
- [90] Monther Aldwairi and Ali Alwahedi. “Detecting fake news in social media networks”. In: *Procedia Computer Science* 141 (2018), pp. 215–222.

- [91] Rutvik Vijjali et al. “Two Stage Transformer Model for COVID-19 Fake News Detection and Fact Checking”. In: *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. 2020, pp. 1–10.
- [92] Nguyen Vo and Kyumin Lee. “The rise of guardians: Fact-checking url recommendation to combat fake news”. In: *The 41st international ACM SIGIR conference on research & development in information retrieval*. 2018, pp. 275–284.
- [93] Bhavika Bhutani et al. “Fake news detection using sentiment analysis”. In: *2019 twelfth international conference on contemporary computing (IC3)*. IEEE. 2019, pp. 1–5.
- [94] Tommi Gröndahl and N Asokan. “Text analysis in adversarial settings: Does deception leave a stylistic trace?” In: *ACM Computing Surveys (CSUR)* 52.3 (2019), pp. 1–36.
- [95] Ziv Epstein, Gordon Pennycook, and David Rand. “Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources”. In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–11.
- [96] Ran Wang et al. “RumorLens: Interactive Analysis and Validation of Suspected Rumors on Social Media”. In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 2022, pp. 1–7.
- [97] Sebastian Tschiatschek et al. “Fake news detection in social networks via crowd signals”. In: *Companion proceedings of the the web conference 2018*. 2018, pp. 517–524.
- [98] Caroline Fisher. “What is meant by ‘trust’ in news media?” In: *Trust in media and journalism*. Springer, 2018, pp. 19–38.
- [99] Fabian Prochazka and Wolfgang Schweiger. “How to measure generalized trust in news media? An adaptation and test of scales”. In: *Communication Methods and Measures* 13.1 (2019), pp. 26–42.
- [100] Jesper Strömbäck et al. “News media trust and its impact on media use: Toward a framework for future research”. In: *Annals of the International Communication Association* 44 (2020), pp. 139–156.
- [101] Justin D Martin and Fouad Hassan. “News media credibility ratings and perceptions of online fake news exposure in five countries”. In: *Journalism Studies* 21.16 (2020), pp. 2215–2233.

- [102] Ivan Habernal et al. “Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018, pp. 386–396.
- [103] Zhivar Sourati et al. “Case-based reasoning with language models for classification of logical fallacies”. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 2023, pp. 5188–5196.
- [104] Max Glockner et al. “Missci: Reconstructing Fallacies in Misrepresented Science”. In: *arXiv preprint arXiv:2406.03181* (2024).
- [105] Zhivar Sourati et al. “Robust and explainable identification of logical fallacies in natural language arguments”. In: *Knowledge-Based Systems* 266 (2023), p. 110418.
- [106] Chadi Helwe et al. “MAFALDA: A Benchmark and Comprehensive Study of Fallacy Detection and Classification”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024, pp. 4810–4845.
- [107] Francisco Zanartu et al. “A technocognitive approach to detecting fallacies in climate misinformation”. In: *Scientific Reports* 14.1 (2024), p. 27647.
- [108] Yuanyuan Lei and Ruihong Huang. “Boosting Logical Fallacy Reasoning in LLMs via Logical Structure Tree”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 13157–13173.
- [109] Kai Shu et al. “Mining disinformation and fake news: Concepts, methods, and recent advancements”. In: *Disinformation, Misinformation, and Fake News in Social Media*. Springer, 2020, pp. 1–19.
- [110] Darrin Baines et al. “Defining misinformation, disinformation and malinformation: An urgent need for clarity during the COVID-19 infodemic”. In: *Discussion Papers* 20.06 (2020), pp. 20–06.
- [111] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.
- [112] James Thorne et al. “The Fact Extraction and VERification (FEVER) Shared Task”. In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. 2018, pp. 1–9.
- [113] Saurabh Agarwala et al. “Detecting Semantic Similarity Of Documents Using Natural Language Processing”. In: *Procedia Computer Science* 189 (2021), pp. 128–135.

- [114] Sonia Castelo et al. “A topic-agnostic approach for identifying fake news pages”. In: *Companion proceedings of the 2019 World Wide Web conference*. 2019, pp. 975–980.
- [115] Miguel A Alonso et al. “Sentiment analysis for fake news detection”. In: *Electronics* 10.11 (2021), p. 1348.
- [116] Mike Thelwall et al. “Sentiment strength detection in short informal text”. In: *Journal of the American society for information science and technology* 61.12 (2010), pp. 2544–2558.
- [117] Lucas Lima Vieira et al. “Analysis of the Subjectivity Level in Fake News Fragments”. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web*. 2020, pp. 233–240.
- [118] Clayton Hutto and Eric Gilbert. “Vader: A parsimonious rule-based model for sentiment analysis of social media text”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1. 2014, pp. 216–225.
- [119] Steven Loria et al. “Textblob Documentation”. In: *Release 0.15 2* (2018), p. 269.
- [120] Mehrdad Koohikamali and Anna Sidorova. “Information Re-Sharing on Social Network Sites in the Age of Fake News.” In: *Informing Science* 20 (2017).
- [121] Jean Caminha et al. “A smart trust management method to detect on-off attacks in the internet of things”. In: *Security and Communication Networks* 2018 (2018).
- [122] Ryma Abassi. “Dealing with collusion attack in a trust-based MANET”. In: *Cybernetics and Systems* 49.7-8 (2018), pp. 475–496.
- [123] Kaggle. *Fake News*. Ed. by Kaggle. 2018. URL: <https://www.kaggle.com/c/fake-news>.
- [124] Haiyi Zhang and Di Li. “Naive Bayes text classifier”. In: *2007 IEEE International Conference on Granular Computing (GRC 2007)*. IEEE. 2007, pp. 708–708.
- [125] Shan Suthaharan. “Support vector machine”. In: *Machine learning models and algorithms for big data classification*. Springer, 2016, pp. 207–235.
- [126] Zhongheng Zhang. “Introduction to machine learning: k-nearest neighbors”. In: *Annals of translational medicine* 4.11 (2016).
- [127] Gérard Biau and Erwan Scornet. “A random forest guided tour”. In: *Test* 25.2 (2016), pp. 197–227.
- [128] Yixin Nie et al. “Combining fact extraction and verification with neural semantic matching networks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 6859–6866.
- [129] Marie-Catherine De Marneffe and Christopher D Manning. *Stanford typed dependencies manual*. Tech. rep. Technical report, Stanford University, 2008.

- [130] Neill Fitzpatrick. “Media manipulation 2.0: the impact of social media on news, competition, and accuracy”. In: (2018).
- [131] David Berdik et al. “A survey on blockchain for information systems management and security”. In: *Information Processing & Management* 58.1 (2021), p. 102397.
- [132] Muhammad Saad et al. “Fighting fake news propagation with blockchains”. In: *2019 IEEE Conference on Communications and Network Security (CNS)*. IEEE. 2019, pp. 1–4.
- [133] Zeinab Shahbazi and Yung-Cheol Byun. “Fake media detection based on natural language processing and blockchain approaches”. In: *IEEE Access* 9 (2021), pp. 128442–128453.
- [134] Liqun Liu et al. “A survey for the application of blockchain technology in the media”. In: *Peer-to-Peer Networking and Applications* 14.5 (2021), pp. 3143–3165.
- [135] Alastria. *Alastria Blockchain Ecosystem*. <https://alastria.io/>.
- [136] Ravi Kiran Raman and Lav R Varshney. “Information-Theoretic Approaches to Blockchain Scalability”. In: *Handbook on Blockchain*. Springer, 2022, pp. 257–296.
- [137] Mohammad Javed Morshed Chowdhury et al. “Blockchain versus database: a critical analysis”. In: *IEEE International conf on trust, security and privacy in computing and communications*. IEEE. 2018, pp. 1348–1353.
- [138] Eve Fleisig, Rediet Abebe, and Dan Klein. “When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 6715–6726.
- [139] Wei-Lin Chiang et al. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. Mar. 2023. URL: <https://lmsys.org/blog/2023-03-30-vicuna/>.
- [140] Subhabrata Mukherjee et al. “Orca: Progressive learning from complex explanation traces of gpt-4”. In: *arXiv preprint arXiv:2306.02707* (2023).
- [141] Hugo Touvron et al. “Llama 2: Open foundation and fine-tuned chat models”. In: *arXiv preprint arXiv:2307.09288* (2023).
- [142] Lianmin Zheng et al. “Judging llm-as-a-judge with mt-bench and chatbot arena”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 46595–46623.
- [143] Luigi Serreli et al. “A Generative Adversarial Network (GAN) Fingerprint Approach Over LTE”. In: *IEEE Access* 12 (2024), pp. 82083–82094. DOI: 10.1109/ACCESS.2024.3411293.

- [144] Luigi Serreli et al. “Robust QUIC-Based Signalling for WebRTC in Impaired Networks”. In: *2023 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*. 2023, pp. 187–192. DOI: 10.1109/MeditCom58224.2023.10266652.
- [145] Gulnaziye Bingol et al. “The Impact of Network Impairments on the QoE of WebRTC applications: A Subjective study”. In: *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*. 2022, pp. 1–6. DOI: 10.1109/QoMEX55416.2022.9900882.



UNIONE EUROPEA
Fondo Sociale Europeo



Ministero dell'Università
e della Ricerca



PON
RICERCA
E INNOVAZIONE
2014 - 2020



UNIVERSITÀ DEGLI STUDI
DI CAGLIARI

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020, risorse FSE REACT-EU
Azione IV.4 “Dottorati e contratti di ricerca su tematiche dell’innovazione”
e Azione IV.5 “Dottorati su tematiche Green”