ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Ανίχνευση Ψευδών Ειδήσεων στα Μέσα Κοινωνικής Δικτύωσης
(Fake News Detection on Social Media)

Χρήστος Ε. Τζώρας

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Σεπτέμβριος 2025

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Ανίχνευση Ψευδών Ειδήσεων στα Μέσα Κοινωνικής Δικτύωσης
(Fake News Detection on Social Media)

Χρήστος Ε. Τζώρας

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Σεπτέμβριος 2025

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη Συνέλευση του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή της σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Ν.ΠΕΛΕΚΗΣ Αναπληρωτής Καθηγητής

- Ε. ΚΟΦΙΔΗΣ Αναπληρωτής Καθηγητής

- ΑΘ. ΡΑΚΙΤΖΗΣ Επίκουρος Καθηγητής

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

Fake News Detection on Social Media

By

Christos E. Tzoras

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
September 2025

Στους γονείς μου

Βαγγέλη και Κατερίνα

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου, κ. Πελέκη Νικόλαο για την πολύτιμη καθοδήγηση, τη στήριξη και τις γόνιμες παρατηρήσεις του σε όλη τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας. Η συμβολή του υπήρξε καθοριστική, όχι μόνο στην επιστημονική πορεία αυτής της μελέτης, αλλά και στη δική μου ακαδημαϊκή εξέλιξη.

Ιδιαίτερες ευχαριστίες οφείλω στους γονείς μου, για την αδιάκοπη στήριξη, την ενθάρρυνση και την αγάπη τους, που με συνόδευσαν σε κάθε βήμα αυτής της προσπάθειας και στάθηκαν πάντοτε στο πλευρό μου.

Τέλος, θα ήθελα να ευχαριστήσω από καρδιάς τη σύζυγό μου, για την κατανόηση, την υπομονή και την αμέριστη συμπαράστασή της σε όλη τη διάρκεια των σπουδών μου. Η παρουσία και η στήριξή της υπήρξαν για εμένα πολύτιμη δύναμη και έμπνευση.

Abstract

This thesis explores the development of a novel, explainable system for fake news detection, leveraging the capabilities of large language models (LLMs) within a modular multi-agent architecture. The system integrates several specialized agents, such as a Fact Checker, Confidence Scorer, Source Quality Evaluator, and a Supervisor Agent, each contributing interpretable insights to the final verdict. Unlike traditional black-box models, this approach emphasizes transparency and human-aligned reasoning, enabling users to trace back the rationale behind each classification.

To assess performance, the system was manually tested on 80 news claims (40 from the LIAR dataset and 40 from FakeNewsNet), using both titles and full-texts. The system achieved an overall accuracy of 75%, with 76.6% accuracy on title-based entries. Beyond accuracy, the system offers structured outputs, evidence citations, confidence scores, and semantic justifications – marking a major improvement in interpretability over existing models like BERT, LSTM, and hybrid deep learning frameworks.

The results indicate that LLM-based agents, when orchestrated through interpretable rule-based coordination, can offer competitive performance with unprecedented explainability. This thesis concludes by proposing directions for future work, including testing across more datasets, automating agent decision logic with mathematical functions, an exploring other LLM families.

Περίληψη

Η παρούσα διπλωματική εργασία παρουσιάζει την ανάπτυξη ενός καινοτόμου και επεξηγήσιμου συστήματος για την ανίχνευση ψευδών ειδήσεων, αξιοποιώντας τις δυνατότητες των Μεγάλων Γλωσσικών Μοντέλων (LLMs) σε ένα πολυπρακοτρικό αρχιτεκτονικό πλαίσιο. Το σύστημα αποτελείται από εξειδικευμένους πράκτορες, όπως ο Fact Checker, Confidence Scorer, Source Quality Evaluator, and a Supervisor Agent, που συνεργάζονται για την παραγωγή διαφανούς και τεκμηριωμένης τελικής απόφασης.

Σε αντίθεση με τα παραδοσιακά μοντέλα μηχανικής μάθησης, το προτεινόμενο σύστημα παρέχει ερμηνεύσιμες εξόδους, αιτιολόγηση αποφάσεων και παραπομπές σε εξωτερικές πηγές. Η απόδοση του αξιολογήθηκε χειροκίνητα σε 80 δηλώσεις (40 από το LIAR και 40 από το FakeNewsNet), τόσο σε μορφή τίτλου όσο και πλήρους κειμένου. Το σύστημα πέτυχε συνολική ακρίβεια 75%, με 76.6% ακρίβεια σε τίτλους ειδήσεων.

Πέρα από την ακρίβεια, η μεγαλύτερη συνεισφορά έγκειται στην ιχνηλασιμότητα και στην εξήγηση των αποφάσεων σε επίπεδο πηγής, μετρικής και νοηματικής συνάφειας. Η εργασία καταλήγει με προτάσεις για μελλοντική έρευνα, όπως η επέκταση σε περισσότερα σύνολο δεδομένων, η υιοθέτηση μαθηματικών κανόνων στη λογική των πρακτόρων, και η αξιολόγηση άλλων τύπων LLMs.

CONTENTS

Chapter 1

Introduction

1.1 Background: The Dangers of Fake News in Modern Society

In recent years, the rise of fake news has become an alarming global phenomenon, largely fueled by the rapid expansion of digital platforms and social media networks. These environments amplify the rapid dissemination of information (both factual and fabricated) across vast audiences. As a result, public opinion can be significantly influenced by fabricated or misleading content, which is often presented with the appearance of credible journalism. According to Lazer et al. (2018), fake news undermines trust in legitimate news sources and disrupts the foundational principle of an informed society.

The 2016 U.S. presidential election served as a pivotal moment in highlighting the power of misinformation in shaping political outcomes. The emergence of fake news during that period revealed how easily public discourse could be manipulated, and in 2017, “fake news” was declared the Word of the Year by Colins Dictionary, signifying its global relevance. Allcott and Gentzkow (2017) emphasize that this surge is not coincidental but rather a product of declining trust in traditional media, the deliberate use of disinformation by public figures, and advances in artificial intelligence that enable the creation of convincing false narratives.

The dangers posed by fake news are multifaceted. One of the most pressing risks is its ability to aggravate social divisions. For example, during the 2020 Black Lives Matter protests, widespread misinformation contributed to heightened racial and political tensions, demonstrating how false narratives can deepen existing societal rifts (Vosoughi et al., 2018). Moreover, fake news threatens democratic institutions by distorting reality and shaping public perception through falsehoods. This pattern has been observed globally, with governments and organized groups exploiting fake news to influence or control public opinion (Kouzy et al., 2020).

The scale of the problem is especially evident in Europe, where a 2020 report by the European Commission revealed that nearly 40% of EU citizens encounter fake news

on daily basis (Commisiion,2020). This statistic underscores the urgent need for effective detection mechanisms to curb the spread of misinformation.

Despite growing awareness and the deployment of various tools, fake news continues to evolve, posing a persistent challenge. Efforts to combat it include the development of machine learning and deep learning algorithms capable of identifying and flagging fabricated content. Nevertheless, these solutions remain in a race against increasingly sophisticated disinformation tactics, highlighting the need for adaptable, modular detection systems that go beyond static method static, monolithic models.

Given the magnitude and danger of fake news, the need for robust detection systems has never been more critical. This thesis builds upon the extensive body of research conducted by scholars and practitioners in the field of fake news detection. Rather than starting from scratch, it synthesizes existing methodologies, evaluates their strengths and limitations, and proposes a novel agentic architecture that leverages multiple LLM-powered agents to collaboratively detect and verify fake news in real-time. The ultimate goal is not merely academic; it is practical and user-oriented. By designing and deploying a multi-agent architecture grounded in Agentic AI principles (where each agent performs specialized tasks and communicates in coordinated workflow) this thesis presents a practical and scalable system for real time fake news detection. Built within the Flowise platform and leveraging tool-augmented reasoning (such as web search APIs, source quality checks, language translation, and scoring metrics), the prototype bridges the gap between research and usability. Rather than relying on monolithic classifiers or static models, it distributes the cognitive load across specialized agents that mimic expert workflows. This approach empowers end-users with an interactive, evidence-driven tool capable of interpreting, verifying, and scoring claims with transparency and precision. Ultimately, the thesis aims to democratize fact checking by delivering an intelligent, explainable verification tool that empowers everyday users the complexities of today's digital information ecosystem with confidence and clarity.

1.2 Research Objectives and Scope

This thesis aims to design, build, and test a practical system for automated fake news detection, one that emphasizes modularity, transparency, and real time responsiveness. Instead of treating detection as a monolithic classification task, the proposed system adopts a more human like, step by step verification approach. Recognizing the limitations of traditional machine learning and deep learning methods, which often operate as black box classifiers, this work proposes a coordinated, multi agent framework. Each agent within this framework specializes in a distinct verification task, working together to deliver a more transparent and interpretable analysis process.

The system is built around a modular architecture where specialized agents handle specific tasks, ranging from text processing and factual consistency checks to credibility scoring, sentiment analysis, timeline relevance, and source quality assessment. These components interact within a unified environment to collectively assess the reliability of user-submitted claims.

The architecture is implemented within the Flowise platform, which provides a visual interface for orchestrating agents and integrating APIs such as Google Custom Search, Serper, and metadata evaluation components. This setup enables rapid prototyping while keeping the emphasis on the underlying agentic design, particularly the combination of language model capabilities with structured reasoning and third-party data verification.

In the scope of this research, a comprehensive review of existing fake news detection methodologies is conducted, encompassing content-based techniques that analyze linguistic patterns, metadata driven approaches that evaluate auxiliary information, and hybrid models that combine multiple analytical layers. The study critically examines the advantages and inherent limitations of conventional machine learning and deep learning systems, particularly in terms of their explainability and adaptability to evolving misinformation strategies. Building on these insights, the research focuses on designing and implementing a modular detection architecture that supports real-time decision-making while ensuring transparency in how conclusions are drawn.

An integral part of this system is the incorporation of external search engines and source reliability assessments, which enrich the verification process with real time, evidence backed insights. Additionally, the project emphasizes the development of an

interpretive output format that goes beyond binary classifications to clearly articulate the reasoning, credibility evaluations, and decision boundaries involved in each assessment. The system's efficacy is validated through a series of controlled test cases designed to benchmark its performance under diverse scenarios. Finally, the study provides a comparative analysis that highlights how the proposed multi agent architecture differs from traditional detection systems, particularly in terms of flexibility, explainability, and usability for end-users navigating the digital information landscape.

This study focuses exclusively on textual claims expressed in natural language, deliberately excluding multimedia formats like images, videos, or audio. The goal is to refine the orchestration of existing AI components, rather than reinventing new language models from scratch. By concentrating on modular design and workflow coordination, this research prioritizes usability, adaptability, and real-world applicability over theoretical model advancements in model design.

Chapter 2

Literature Review

2.1 Definition of Fake News

2.1.1 Defining Fake News: Misinformation, Disinformation, Satire

Fake news refers to the intentional creation and dissemination of false or misleading information that is presented in the format of legitimate journalism with the aim of deceiving its audience. Unlike accidental reporting errors or unfounded rumors, fake news is deliberately designed to manipulate public perception, often driven by political, ideological, or financial motives. As Gelfert (2018) explains, fake news is fabricated content that mimics the tone, structure, and aesthetics of authentic news stories, yet fundamentally lacks factual grounding. Its potency lies in its ability to masquerade as credible reporting, exploiting reader's trust in journalistic conventions.

However, the boundaries between fake news, misinformation, disinformation, and satire are often blurred, complicating detection efforts. Misinformation typically refers to the unintentional spread of inaccurate or false information, errors that occur without a deliberate intent to deceive. In contrast, disinformation involves the purposeful dissemination of falsehoods with the explicit aim to mislead or manipulate audiences, often orchestrated by organized campaigns or actors with vested interests. Baptisa and Gradim (2022) highlight that fake news aligns closely with disinformation, particularly in its strategic appropriation of legitimate news outlet's visual identity, linguistic style, and narrative formats to increase its deceptive effectiveness.

Another distinct category is satire, content that deliberately distorts facts for humorous or critical commentary, but with no genuine intent to deceive. Satirical news articles, such as those from "The Onion" or "Babylol Bee", often employ exaggeration and absurdity to mock real events or societal trends. Despite their lack of deceptive effectiveness, satirical articles can still influence audiences by shaping public opinion and highlighting societal issues through humor.

What set fake news apart from these adjacent phenomena is the deliberate intent to mislead, coupled with a strategy to blend truth and falsehood in ways that maximize believability. Fake news often interweaves selective facts, emotional narratives, and

manipulated visuals to construct persuasive yet misleading stories. The digital ecosystem amplifies this problem, as social media algorithms prioritize engagement, inadvertently facilitating the viral spread of such content long before verification mechanisms can respond.

Given its sophisticated mimicry of credible journalism and its intentional design to evade casual detection, fake news has emerged as a significant challenge to information integrity in the digital age. While misinformation and satire may contribute to the spread of false narratives incidentally, fake news represents a targeted attack on public trust and journalistic standards, often weaponized in political, social, or economic conflicts.

In the context of automated detection, distinguishing between these categories become crucial, as the intention behind the content (whether to deceive, inform, or entertain) greatly influences the choice of detection strategies. This thesis, therefore, adopts a nuanced view that recognizes fake news as a distinct class of disinformation, engineered to exploit the vulnerabilities of digital information ecosystems, and requiring multi-layered detection frameworks that account for both content characteristics and dissemination patterns.

2.1.2 Problem Complexity: The Multifaced Nature of Fake News

Detecting fake news is not a straightforward task. Its complexity comes from how it blends facts with distortions, often presenting misleading information that appears truthful on the surface. Beyond content, the context in which information is shared, such as timing, platform, and audience perception can influence whether it becomes misleading or not.

Fake news creators continuously evolve their tactics, using AI-generated texts, manipulated visuals, and coordinated disinformation campaigns that outpace traditional detection systems. Additionally, most automated solutions operate as black boxes, making decisions without providing clear explanations to end users.

As a result, any robust detection strategy must begin with a clear understanding of where deception can occur within a news story. Whether it’s hidden within the textual content, embedded in the metadata, revealed through patterns of dissemination, or

reflected in emotional tone, fake news manifests across multiple layers. The next section outlines these logical perspectives on fake news detection, which serve as the foundational viewpoints for designing effective identification frameworks.

2.2 Logical Perspectives on Fake News Detection

Given the layered complexity of fake news, effective detection cannot rely on a single method or viewpoint. Different forms of deception may be embedded in the content of an article, hidden within its metadata, or revealed through how it spreads across networks. Researchers have thus developed several analytical perspectives to tackle fake news, each focusing on a distinct facet of the problem. These perspectives ranging from content-based analysis to metadata examination, dissemination patterns, emotional tone, and hybrid combinations, provide structured ways to identify where falsehoods might reside. Understanding these perspectives is essential for designing detection systems that are both comprehensive and adaptable to evolving misinformation tactics (Shu et al., 2017; Zhou & Zafarani, 2020).

2.2.1 Content-based Analysis

Content-based analysis focuses on examining the actual text of a news article to uncover signs of deception. This approach looks at linguistic, stylistic, and semantic features such as word choice, sentence structure, tone, and writing patterns that may indicate manipulation or falsehoods. By using Natural Language Processing (NLP) techniques, these methods analyze the content in isolation, without considering external information like the author's profile or how the article spread on social media (Perez-Rosas et al., 2018; Rashkin et al., 2017).

The strength of content-based analysis lies in the assumption that fabricated news often leaves subtle clues in its language, such as exaggerated claims, emotional bias, or unusual writing styles that differ from credible reporting (Conroy et al., 2015). However, while content analysis provides a foundational layer of detection, it does not account for contextual factors outside the text itself.

To address these external dimensions, researchers turn to metadata-based analysis, which examines attributes surrounding the content to assess credibility from a different angle.

2.2.2 Metadata-based Analysis

Metadata-based analysis examines information about a news article rather than the article's content itself. This includes attributes like the source's domain, the author's profile, publication date, and indicators of political or ideological alignment. By analyzing these contextual features, metadata-based methods aim to assess the credibility of the source and identify patterns commonly associated with misinformation (Shu et al., 2011).

The underlying assumption is that fake news often originates from sources with questionable reputations or follows publishing patterns that differ from legitimate journalism (Castillo et al., 2011). Metadata analysis helps in flagging suspicious content by focusing on where and by whom the information was produced.

However, while metadata provides valuable context, it does not capture how information behaves once it enters the public sphere. For this, propagation-based analysis offers a perspective that looks at how fake news spreads across networks.

2.2.3 Propagation-Based Analysis

Propagation-based analysis focuses on how new articles spread across social networks, rather than what they contain or who created them. This approach is built on the observation that fake news often follows distinct diffusion patterns, spreading faster, reaching broader audiences, and forming abnormal network structures compared to legitimate news (Vosoughi et al., 2018; Friggeri et al., 2014). By analyzing retweet chains, information cascades, and user interaction networks, propagation-based methods aim to detect fake news based on its behavioral footprint in the digital ecosystem.

While propagation analysis adds a valuable layer of insight, it requires monitoring the news after it begins circulating, which can delay detection. Additionally, it does not

directly assess the content's emotional or psychological impact, a factor explored through sentiment-based analysis, which look at how fake news leverages emotional manipulation to enhance its spread (Zhao et al., 2015).

2.2.4 Sentiment-Based Analysis

Sentiment-based analysis examines the emotional tone expressed in news articles or social media posts. It investigates whether emotionally charged language, such as fear, anger, or outrage is being used to manipulate readers and drive engagement (Zubiaga et al., 2016; Giachanou et al., 2019). Since fake news often relies on triggering strong emotional reactions, sentiment intensity and polarity scores can serve as useful indicators of potential misinformation.

However, while emotional cues scan signal deceptive intent, sentiment analysis alone is not sufficient for reliable detection (Lazer et al., 2018). Emotional language is also common in legitimate opinion pieces or satire. To overcome these limitations, researchers increasingly adopt hybrid analytical approaches, which combine multiple perspectives to improve detection accuracy.

2.2.5 Hybrid Analysis

Hybrid analytical approaches aim to overcome the limitations of single-perspective methods by combining multiple layers of analysis into an integrated detection framework. For instance, a hybrid system might assess the linguistic patterns of an article (content-based analysis), verify the credibility of its source (metadata-based analysis), and monitor its spread across networks (propagation-based analysis) simultaneously (Wang, 2017; Ruchansky et al., 2017). By fusing these complementary viewpoints, hybrid models can provide a more comprehensive assessment of an article's credibility.

The strength of hybrid approaches lies in their ability to cross-validate signals from different sources of information. An article that uses manipulative language but originates from a reputable source may be treated differently from content that is

emotionally neutral but spreads through suspicious channels. This multi-dimensional view enables more nuanced detection strategies.

However, hybrid models introduce their own challenges. They often require complex system architectures, greater computational resources, and sophisticated reasoning mechanisms to interpret conflicting signals. Additionally, integrating diverse data types (textual, structural, and network-based) demands careful design to ensure the system remains interpretable and practical for real-time detection scenarios (Zhou & Zafarani, 2020).

In the next section, we will explore hoe these logical perspectives translate into technical implementations, examining how Machine Learning, Deep Learning, and Hybrid technical architectures have been employed to operationalize fake news detection strategies.

2.3 Technical Implementation Paradigms

While the previous section outlined the different perspectives from which fake news can be analyzed, the practical implementation of detection systems depends on the underlying computational techniques used. Broadly, existing research can be grouped into three technical paradigms: Machine Learning, Deep Learning, and Hybrid Approaches that combine elements of both. Each paradigm reflects not only a difference in modeling complexity but also how detection systems process, interpret, and make decisions based on various features.

In the following sections, we will examine key studies that exemplify these three paradigms, highlighting their methodologies, datasets, strengths, and limitations.

2.3.1 Machine Learning-Based Approaches

Machine learning (ML) has been foundational in early fake news detection research, primarily leveraging manual feature engineering to identify linguistic, structural, and behavioral patterns associated with misinformation. Unlike deep learning models, which automatically learn complex patterns, traditional ML methods rely on explicitly defined features fed into classifiers such as Support Vector Machines (SVM), Random

Forests, Logistic Regression, and ensemble methods. These models are particularly valued for their interpretability, efficiency, and effectiveness on smaller, well-structured datasets.

Below we examine key studies employing ML-based techniques across different analytical perspectives of fake news detection.

2.3.1.1 Content Based Linguistic Analysis

Perez-Rosas et al. (2018) in the article named "Content Based Linguistic Analysis" conducted an influential study exploring the linguistic characteristics of fake news articles and their detectability through machine learning. The authors used two distinct datasets. The first, known as FakeNewsAMT, included 480 news articles split evenly between real and fake. The real articles were sourced from six different topical sections (such as politics, education, and technology), while the fake articles were crowdsourced via Amazon Mechanical Turk (AMT). Workers were instructed to write fabricated news in a professional tone, mirroring real news articles in both headline and body structure.

The second dataset, Celebrity News, was composed of 500 articles (250 reals and 250 fake). The real pieces were manually verified and collected from trusted entertainment sources. The fake articles in this dataset were also carefully selected to maintain realism and credibility, especially in the stylistic tone typical of celebrity journalism.

To detect fake news, the researchers engineered a rich variety of linguistic features for use in classification. These included n-grams and bigrams to capture common word patterns; punctuation frequency, measuring the distribution of 12 punctuation types; psycholinguistic metrics drawn from the LIWC lexicon, analyzing emotions, social language, and pronouns; readability indices such as Flesch-Kincaid scores and syllable counts; and syntactic features extracted from context-free grammar trees using the Stanford Parser.

The team employed a Support Vector Machine (SVM) classifier and validated its performance using 5-fold cross-validation. Across both datasets, the best results emerged when all feature categories were combined, showing the importance of a multi-faceted linguistic approach.

Their analysis revealed nuanced differences between fake and real news. In the FakeNewsAMT dataset, true news articles exhibited more cognitive processing words, temporal markers, and negations. In contrast, fake articles leaned toward emotional expressions, used more verbs, and emphasized the present or future. Similarly, in the Celebrity dataset, true articles used more first-person pronouns and expressed positivity, whereas fake ones preferred third-person pronouns and contained more negative sentiment.

An interesting insight arose regarding learning curves: model accuracy slightly declined with larger training sets, hinting at possible data sparsity or overfitting tendencies, an important observation for scaling such models.

Ultimately, the authors concluded that linguistic features alone offer strong predictive power for fake news detection. However, they argued that future models could significantly benefit from incorporating complementary metadata, such as the number of external links, user engagement metrics, visual layout cues, or integration with automated fact-checking tools. This proposed hybrid approach would potentially increase robustness and detection accuracy, especially in real-world, web-scale environments.

2.3.1.2 Metadata based via User Interaction Patterns

Tacchini et al. (2017) in the article “Metadata based via User Interaction Patterns” proposed a metadata-driven approach to fake news detection based entirely on user interaction behavior, specifically Facebook user likes. The dataset was compiled from 32 Facebook pages: 14 conspiracy-oriented pages labeled as hoax sources, and 18 pages with a scientific orientation considered as non-hoax sources. This resulted in a comprehensive dataset comprising 15.500 posts, over 909.000 users, and more that 2.3 million likes. Additionally, a refined subset called the intersection dataset included 10.520 posts liked by 14.139 users who had interacted with both hoax and non-hoax contents.

User engagement behavior was central to the model. About 74.7% of users liked only hoax content, 20.3% only non-hoax content, and the remaining 5% likes content from

both categories. This behavioral segmentation provided the basis for two machine learning models approaches.

The first technique was Logistic Regression (LR), in which each post was represented as a binary vector showing which users has likes it. A weight was learned for each user, indicating their preference, positive weights for non-hoax affinity, and negative weights for hoax affinity. The probability of a post being legitimate was computed via the logistic function:

$$p_i = \frac{1}{1+e^{-y_i}} \text{ where } y_i = \sum_{u \in U} x_{iu} w_u$$

The model showed good performance with limited training data but could not infer the credibility of posts liked by users not seen during training.

The second approach, Harmonic Boolean Label Crowdsourcing (BLC), employed a bipartite graph model of posts and users. It operated by propagating credibility scored iteratively using beta-distribution-like update equations. Posts known to be hoax or non-hoax were initially assigned q_i= -1 or q_i= 1, and others were initialized with q_i=0. During training, the model updated trustworthiness scores for users and posts with the following equations:S

- User Update: $a_u = A + \sum_{i \varepsilon \theta_u, q_i} 1 > 0$, $\beta_u = B - \sum_{I \varepsilon \theta_U, q_I < 0} 1$ where $q_u = \frac{a_u - \beta_u}{a_u + \beta_u}$
- Post Update: $a_i = A' + \sum_{u \varepsilon \theta_i, q_u > 0} 1$, $\beta_i = B' - \sum_{u \varepsilon \theta_i, q_u < 0} 1$ where $q_i = \frac{a_i - \beta_i}{a_i + \beta_i}$

Where α and β are adjusted over five iterations using priors A=5.01, B=5 and `A=`B=5, favoring non -hoax bias slightly.

The performance of these two approaches was tested across various training data sizes. With only 0.5% of the data sued for training (~80 posts), Logistic Regression reached an accuracy of about 90%, whereas BLC achieved 99.4%. With 80% training data, BLC still maintained a marginal lead, showing greater robustness. For instance:

- At 0.1% training, both models reached ~80% accuracy.
- At 1% training, Logistic Regression exceeded 90%, while BLC remained over 99%.

In the intersection dataset, users who liked both hoax and non-hoax content, Logistic Regression slightly outperformed BLC in cases of minimal training achieving 91.6%

accuracy with 10% labeled data and 56 with 0.1%, maintaining a consistent 3-4% margin.

In cross-page learning evaluations, where models trained on a subset of pages were tested on unseen ones, BLC showed exceptional generalization. For example:

- In a "one-page-out" experiment, BLC scored 99.1% accuracy vs. 79.4% for LR.
- In "half-pages-out" testing, BLC reached 99.3% vs. 71.6% for LR.

These results underscored the ability of BLC to transfer credibility information through shared users, offering a powerful model for community-wide fake news detection.

Ultimately, the study demonstrated that fake news could be accurately predicted using interaction metadata alone. These finding challenges content-centric paradigms and opens news avenues for efficient, scalable detections system particularly suited for social platforms where rapid moderation is necessary.

2.3.1.3 Propagation – Based Detection Using Random Forests

Wu et al. (2015), in their study "Propagation – Based Detection Using Random Forests" introduced a novel approach that targets the early-stage propagation patterns of false information on Twitter. Their methodology rests on the hypothesis that fake news tends to exhibit different diffusion behaviors compared to legitimate news, especially in the initial phases of spread. Rather that analyzing article content or metadata, their system focused entirely on how information traveled through retweet structures.

The core dataset comprised propagation trees generated from Twitter posts, capturing temporal features such as retweet timing and the structural layout of the message's diffusion. These trees modeled user interactions from the initial post through various stages of engagement, providing a visual and quantitative trace of how quickly and widely the message spread.

To perform classification, the researchers trained a Random Forest classifier on features extracted from these trees. These features included depth, width, repost timing

intervals, and user engagement diversity, all of which helped capture early behavioral signatures of viral misinformation. Importantly, the model was designed to operate at the early staged of rumor propagation, thus enabling faster intervention.

Their results demonstrated that propagation-based indicators could effectively distinguish between fake and real information even before the content reached peak diffusion. This approach is especially valuable in real-time detection scenarios, where waiting for content or metadata analysis might lead to delays.

The study highlights the importance of structural and temporal signals in misinformation detection and shows that diffusion modelling can complement content-based and metadata-based techniques to build robust, layered detection systems.

2.3.1.4 Linguistic Framing and Deceptive Language Cues

Inside the study "Linguistic Framing and Deceptive Language Cues", Rashkin et al. (2017) performed a comprehensive linguistic analysis aimed at identifying deceptive elements in both news articles and political statements. Their work investigated fake news across different genres, trusted news, satire, hoax, and propaganda, by using two key datasets: a corpus of labeled news articles (including sources like Gigaword, The Onion, DC Gazette, and Natural News), and over 10.000 PolitiFact statements rated on a 6-point scale from "True" to "Pants-on-Fire False"/

The main objectives were twofold: to classify news articles into their appropriate genre and to predict the truthfulness of political statements using supervised machine learning models. For the news genre classification, they employed a MaxEnt classifier trained on TF-IDF features. The best model achieved an F1 score of around 65% on out-of-domain testing. For the truthfulness classification of PolitiFact quotes, the authors experimented with Naïve Bayes, MaxEnt, and LSTM models, using text-only input as well as combinations with LIWC-derived features (which capture emotional and social cues).

On the validation set, LSTM performed best for binary classification (True/False) using only text, achieving an F1 score of 0.58. However, simpler models like Naïve Bayes and MaxEnt benefited significantly from the inclusion of LIWC features, reaching scores of 0.58 and 0.58 respectively higher than without LIWC. In the more

challenging 6-class truthfulness prediction, F1 scores were generally low (~0.20-0.22), indicating the nuanced difficulty in finely grading truthfulness. On the test set, LSTM remained competitive for binary tasks, while simpler models held an edge when boosted with LIWC.

Lexical analysis revealed distinct language markers for each news type. For instance, unreliable articles frequently used first and second person pronouns, while trusted news incorporated more numbers and assertive verbs. Propaganda leaned heavily on superlatives and intensifiers, while satire was rich in adverbs. The statistical significance of these differences was validated via Welsch t-tests with Bonferroni correction ($p<0.01$).

In summary, the authors demonstrated that nuanced linguistic cues and sentiment patterns can be used to differentiate types of fake news and measure deceptive framing. While neural models like LSTM learn these patterns implicitly, lexicon-based features such as LIWC remain valuable for interpretability and boosting traditional classifiers. Their work shows that deception detection is feasible through linguistic framing, although classifying multiple shades of truth remains a challenge.

2.3.1.5 Data Mining Perspective with Ensemble Models

One significant introduction into a comprehensive framework for detecting fake news made Shu et al. (2017) studying data mining perspective with ensemble models specializing on social media platforms through the integration of content, user, and social features using traditional ensemble machine learning models. Their approach was built upon the FakeNewsNet dataset, a large-scale corpus designed to support data-driven studies on fake news.

The study highlighted the significant influence of malicious accounts in the spread of misinformation. These accounts were categorized into three main types: bots, cyborgs, and trolls. Bots referred to fully automated accounts that operated without human oversight. Cyborgs were human-managed accounts enhanced with automation tools, enabling rapid and widespread content dissemination. Trolls, on the other hand, were human users who actively spread disinformation with the intent to provoke and mislead. During the 2016 U.S. elections, approximately 19 million bot accounts were

reportedly involved in information manipulation. Additionally, around 1,000 Russian-paid troll accounts were specifically targeted at promoting fake news against Hillary Clinton. These malicious entities significantly amplified the spread of disinformation by exploiting emotional contagion mechanisms such as fear and anger.

The study also delved into the role of echo chambers in reinforcing fake news. Social media algorithms were found to prioritize user-preference alignment, effectively creating feedback loops that intensified ideological segregation. This phenomenon was driven by two psychological heuristics: the social credibility heuristic, which suggested that people tend to believe information if it is endorsed by others; and the frequency heuristic, which implied that repeated exposure increases perceived credibility. As a result, these heuristics skewed users' cognitive biases and led to the formation of highly polarized communities that actively reinforced and circulated fake content.

To model the fake news detection problem, Shu et al. defined a news article as a composite of its publisher and content—comprising the headline, accompanying image, textual body, and other metadata. Users were denoted as {U1, U2, ..., Un}, posts as {p1, p2, ..., pn}, and the engagements as E = {eit}, where each eit represented a specific interaction involving a user ui, post pi, and timestamp t.

For feature extraction, the study adopted a statistical structure that included both content and social context dimensions. Content features were divided into linguistic and visual categories. Linguistic features encompassed lexical indicators such as word counts, sentence lengths, and the use of superlatives; syntactic patterns like part-of-speech tags, function word frequencies, and punctuation usage; domain-specific metrics such as the number of quotations, inclusion of external links, and graph length; as well as deception-related cues typically used in lie detection studies. Visual features were assessed based on clarity, coherence, diversity, and clustering scores. Additional statistical image characteristics included image count, image-text ratio, and the frequency of high-interest images (hot-image rate).

Social context features were analyzed through three lenses. The user-based features focused on profile credibility, gauged through metrics like the average number of followers and the proportion of verified users. Post-based features measured public reaction using stance analysis, topic modeling (via Latent Dirichlet Allocation), and

credibility scoring. Network-based features examined the structure and dynamics of social interactions, considering friendship networks built from follower relationships and diffusion networks that represented the chronological and structural patterns of content propagation.

For mathematical modeling, embedding techniques such as Recurrent Neural Networks (RNNs) were employed to trace the temporal evolution of user posts. Network analysis utilized measures like clustering coefficient, degree centrality, and PageRank to quantify the influence and spread of fake content.

The model architecture combined news content models and social context models. The content models included knowledge-based approaches that validated article claims against external open web data and structured knowledge graphs like DBpedia. Style-based models analyzed deep syntactic patterns using Probabilistic Context-Free Grammar (PCFG) rules, rhetorical structures, and Convolutional Neural Network (CNN)-based classifiers. Social context models were categorized into stance-based and propagation-based methods. Stance-based methods assessed aggregated user opinions through LDA and Facebook-like bipartite user-post networks. Propagation-based models employed heterogeneous information networks to represent and learn the dissemination patterns of fake news.

Model performance was evaluated using a suite of standard metrics: Precision, Recall, F1 Score, Accuracy, and the Area Under the Curve (AUC).

Shu et al. also identified several adjacent research areas connected to fake news detection. This included rumor classification, which involved tasks like stance detection and rumor tracking; truth discovery, which aggregated source reliability using probabilistic inference; clickbait detection, which analyzed inconsistencies between headlines and body content using NLP techniques; and bot detection, which utilized anomaly detection, clustering, and graph-based metrics to identify automated and deceptive accounts. This multi-faceted, ensemble-based approach laid the groundwork for future integration of multi-modal and context-aware systems for tackling fake news at scale.

Conclusion of Machine Learning Techniques

Traditional machine learning techniques have laid a strong foundation for fake news detection, offering efficient and interpretable models that perform well on structured datasets. By relying on engineered features, ranging from linguistic patterns and metadata to propagation behaviors, these models can identify misinformation with notable accuracy. Algorithms like Support Vector Machines (SVM), Logistic Regression (LR), and Random Forests have shown consistent results, particularly when feature selection aligns closely with the nature of fake content.

However, these approaches also face limitations. Their effectiveness is often restricted by the quality and scope of the data, and they may struggle with generalization in dynamic, real-world environments. Additionally, their reliance on manual feature design can limit adaptability to evolving forms of disinformation. While machine learning continues to play a critical role in fake news detection, the growing complexity of misinformation calls for more flexible, multi-dimensional solutions paving the way for deep learning and hybrid approaches.

2.3.2 Deep Learning-Based Approaches

As the complexity and scale of online misinformation increased, deep learning emerged as a powerful alternative to traditional machine learning in fake news detection. Unlike classical models that depend heavily on manual feature engineering, deep learning methods are capable of automatically extracting and learning hierarchical patterns from raw data be it text, user behavior, or propagation structure. Architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models have proven effective in capturing the nuanced semantics, contextual dependencies, and temporal dynamics often embedded in deceptive content. These models are particularly valuable in real-time environments where adaptability and end-to-end learning are crucial. Though they require larger datasets and more computational resources, their potential for generalization and performance scalability positions them at the forefront of modern fake news detection efforts. In the sections that follow, we examine representative deep

learning studies that highlight these capabilities across various data modalities and detection strategies.

2.3.2.1 Convolutional Neural Networks on the LIAR Dataset

In the paper title “Liar, Liar Pants on Fire: A new Benchmark Dataset for Fake News Detection”, Wang (2017) introduced one of the most widely used datasets for fake news research, the LIAR dataset, and applied deep learning techniques, particularly Convolutional Neural Networks (CNN), to the task of detecting fake news in short textual statements.

The LIAR dataset comprises a total of 12.836 manually labeled short statements collected from various public contexts. Unlike many prior studies that rely on crowdsourced data, the annotations in this dataset were performed by expert fact-checkers, which is considered to significantly enhance the reliability and validity of the labels. This is particularly important for fake news detection where context, nuance, and source credibility are often more crucial than in more subjective domains like product reviews. Crowdsourcing, the study argues, is less appropriate in this domain, especially given that news items, particularly those found on television or social media, tend to be shorter and more context sensitive.

The dataset features a six-point truthfulness scale for labeling: pants on fire, false, barely true, half true, mostly true, and true. Additionally, each labeled item is accompanied by a justification provided by expert annotators. For modeling convenience and clarity, some overlapping categories like half-flip or full-flop were merged into more concise classifications such as false, half-true, and true.

Rich metadata is associated with each speaker or source of the statement, including the party affiliation, current job position, home state, and credit history. This metadata enables models not only to analyze what was said but also who said it and in what political or social context, an important step toward understanding the broader mechanics of misinformation.

The most prevalent topics in the dataset include the economy, healthcare, taxes, the federal budget, education, job creation, state-level budget issues, candidate

biographies, elections, and immigration. These categories reflect the politically and socially charged nature of fake news content.

The model the problem, Wang applied a hybrid CNN architecture that integrated both textual and metadata inputs. The text and metadata were processed through the following pipeline:

1. Embedding vectors were randomly initialized for the metadata features.
2. A convolutional layer was used to identify relationships between the metadata inputs.
3. Max-pooling was applied to extract the most dominant and informative signals.
4. These were passed through a bi-directional Long Short-Term Memory (Bi-LSTM) network to capture the contextual dynamics and temporal dependencies.
5. The output of the Bi-LSTM was concatenated with the pooled features derived from the text.
6. This combined feature vector was then fed into a fully connected layer with SoftMax activation to yield the final truthfulness classification

In terms of evaluation, the study compared five different models: majority baseline, logistic regression (LR), support vector machines (SVM), a standalone Bi-LSTM, and a CNN. For traditional models like LR and SVM, implementation was done using LIBSHORTTEXT, a specialized library for short-text classification tasks. The deep learning models, including the CNN and Bi-LSTM, were implemented using TensorFlow. All models employed pre-trained 300-dimensional Word2Vec embeddings trained on Google News for word representation. Hyperparameters were tuned through grid search over the validation set to ensure optimal performance.

For the CNN model, specific architectural settings included filter sized of 2.3 and 4, each with 128 filters. A dropout keep probability of 0.8 was employed to prevent overfitting. The models were trained with a batch size of 64 for 10 epochs. For the hybrid CNN model that incorporated metadata, filter sized of 3 and 8 were used, with 10 filters each, and dropout probabilities of 0.5 and 0.8. This hybrid model was trained for 5 epochs.

The results revealed several important insights. The majority baseline achieved an accuracy of around 20%, setting a low benchmark. Both SVM and LR significantly

outperformed the baseline. However, the Bi-LSTM model struggled with overfitting, resulting in suboptimal generalization performance. Among all, the CNN achieved the highest test accuracy of approximately 27%, statistically outperforming the SVM model with a p-value of less than 0.00001. Notably, the hybrid CNN model the combined textual and metadata features yielded the best overall performance, confirming the added value of integrating context-specific information with deep linguistic representations.

In summary, Wang laid important groundwork in fake new detection using deep learning by introducing a robust dataset and validating the effectiveness of CNN architectures. The study demonstrated that even relatively simple deep learning models could outperform traditional approaches in detecting fake news, especially when enriched with metadata that provides socio-political context to the textual statements.

2.3.2.2 Propagation-Based Detection Using Geometric Deep Learning

Monti et al. (2019) in the study "Propagation – Based Detection Using Geometric Deep Learning" presented a novel deep learning framework for fake news detection on social media, introducing a propagation-based approach that leverages the power of Graph Convolutional Networks (GCNs). Their method, built upon the principles of geometric deep learning, moves beyond textual and metadata analysis by modeling the actual social, temporal and structural patterns of news diffusion on Twitter. The proposed architecture achieves impressive results, notably a 92.7% ROC AUC in URL wise classification and 88.3% in cascade-wise classification. These performances metrics underscore the model's capacity not only to detect fake news with high accuracy but also to do so within just a few hours of a story's initial propagation. Its robustness over time further affirms its potential for real-world, scalable deployment.

The dataset used in the study was constructed from Twitter data spanning May 2013 to January 2018 and was cross-validated against claims debunked by leading fact-checking organizations including Snopes, PolitiFact, and BuzzFeed. The final dataset included 1.084 verified claims and 1.129 unique URLs serving as proxies for those claims. From these, the authors collected 158.951 information cascades created by 202.375 distinct users and mapped over 2.4 million social connections. Each URL was

labeled according to its relationship with verified claims, either affirming or denying them, and the associated tweet cascades were then collected to model their spread.

The architecture of the model begins with the construction of a graph (Gu) in which each node represents a tweet, and the edges capture both social relationships and the directions on information diffusion. Two GCN layers are applied to this graph, producing 64-dimensional vector outputs per node. These outputs are then aggregated using a combination of graph attention mechanisms and mean pooling. The model concludes with two fully connected layers that reduce the dimensionality from 32 to 2, enabling binary classification of the content as either fake or true. The activation function employed is the Scaled Exponential Linear Unit (SELU), and the model is trained using a hinge loss function which was found to yield better early-stage detection performance compared to traditional cross-entropy loss. The optimization is carried out using AMSGrad, with a learning rate of 5e-4 and a mini batch size of 1.

Crucially, the GCN incorporates various types of edge relationships, including follower-followee links, news spread detection (i.e., retweets), and general bidirectional interactions. Node features integrate tweet content, user metadata, and user activity, while edge features are encoded as binary indicators of the types of interactions present. The evaluation is conducted under two key settings. In the URL-wise setting, the model predicts a label for news URL based on all associated cascades (averaging 141 cascades per URL). In the cascade-wise setting, the prediction is made based on a single cascade, which presents a more realistic and challenging scenario. To ensure meaningful diffusion structures, a minimum of six tweets per cascade is required. The dataset is split accordingly for training, validation, and testing: 677/226/226 for IRL-wise, and 3.586/1.195/1.195 for cascade-wise.

The model's performance proves highly competitive. In URL-wise classification, the ROC AUC score reached 92.7% (+- 1,8), while cascade-wise classification achieved 88.3% (+- 2,74). Importantly, detection remains accurate even when performed within the first few hours of propagation: the model saturates at around 15 hours in the URL-wise setting and approximately 7 hours in cascade-wise detection. An ablation study further revealed that user profile, user activity, and network propagation features contributed most significantly to classification performance. Interestingly, in cascade-

wise classification, removing tweet content actually improved the model's accuracy, as it helped avoid overfitting to superficial linguistic cues.

The authors also examined the temporal robustness of their approach through a model aging analysis. In URL-wise classification, a small performance degradation was observed after a 180-day ago gap between training and testing, whereas the cascade-wise model maintained strong performance with only a 4% drop after 260 days. This result suggests that the cascade-wise GCN model, by focusing on structural patterns, generalizes more effectively over time.

From a security and interpretability stand point, the model offers strengths. Since propagation patters are inherently more difficult to fabricate than content, the model is less vulnerable to adversarial manipulation. It also exhibits generalizability across languages and regions, as it does not depend on language-specific features. Visualizations using t-SNE revealed clear separations between credible and non-credible user behavior, adding further interpretability to the learned patterns.

In conclusion, Mont's work significantly advances the field of fake news detection by incorporating geometric deep learning principles to model the spread of information as a graph. The method offers high accuracy, early detection capabilities, and strong robustness over time, making particularly well-suited for deployment in large scale, real-time misinformation monitoring systems. Future research directions include improving adversarial robustness, further exploring the interpretability of GNN decision mechanisms, and extending the model to related domains such as virality prediction and topic modeling.

2.3.2.3 Multi-Domain Visual Deep Learning for Image-Based Fake News

In response to the growing challenge of detecting fake news on social media, particularly when visual content is used to manipulate or mislead, Qi et al. (2019) explains inn study "Multi – Domain Visual Deep Learning for Image – Based Fake News" how he proposed a novel deep learning framework that explicitly focuses on the visual aspects of misinformation. Their study acknowledges a significant shortfall in earlier research, which has predominantly emphasized textual content while overlooking the rich yet deceptive role of images in shaping users' perception of news

credibility. To address this, the authors introduced the Multi-domain Visual Neural Network (MVNN), a comprehensive deep learning model that fuses both physical-level and semantic-level visual signals to enhance fake news image classification.

The MVNN framework is designed around three major neural sub-networks, each targeting a distinct visual information domain. The first is the frequency domain sub-network, which leverages the Discrete Cosine Transform (DCT) to convert image data into the frequency space. This sub-network then uses a one-dimensional convolutional neural network (1D CNN) to process 64 DCT frequency bins, effectively learning subtle traces of digital compression and possible image manipulation artifacts. These features are vital for detecting tampered images that have been edited, compressed, or manipulated to deceive viewers.

Complementing this is the pixel domain sub-network, which is structured as a multibranch CNN. This component is responsible for extracting both low- and high-level visual features from raw pixel data. To enhance the temporal and semantic understanding of these features, a bidirectional Gated Recurrent Unit (Bi-GRU) is integrated, allowing the model to capture sequential dependencies and relationships across semantic layers of the visual input. This design ensures that misleading but visually coherent images, such as those reused in different contexts without direct tampering, can still be detected.

The final piece of the MVNN architecture is the fusion sub-network. This layer serves as a convergence point where the outputs from both the frequency and pixel sub-networks are combined. It employs an attention mechanism to dynamically adjust the relative contributions of each domain, assigning greater weight to the most informative signals for a given input. The output is the passed through a binary classification layer, which determines whether the input image is associated with fake or real news.

The model was trained and evaluated on the Weibo fake news dataset, which contains 9.528 posts with associated images. Performance was assessed using standard classification metrics, including accuracy, precision, recall, and F1-score. In comparative evaluations, the MVNN significantly outperformed all baselines. Traditional forensic features combined with logistic regression (FF+LR) yielded only 65% accuracy and an F1 score of 0.595. Pre-trained CNNs such as VGG achieved slightly better results, with 72.1% accuracy and an F1 score of 0.702, and a fine-tuned

version pushed these metrics to 75.4% and 0.714, respectively. Convolutional autoencoders (ConvAE) performed similarly with 73.4% accuracy and a 0.713 F1 score. In contrast, MVNN achieved 84.6% accuracy and a substantially higher F1 score of 0.832.

An ablation study further demonstrated the value of MVNN's integrated design. Removing the pixel domain led to an approximately 11% drop in accuracy, highlighting the critical role of raw semantic content. Similarly, removing either the attention mechanism or the Bi-GRU caused performance to decline by about 2%, reinforcing the importance of dynamic feature weighting and temporal modeling. These findings confirmed that frequency and pixel domain features are complementary and that their joint modeling yields more robust and accurate results.

In addition to visual-only classification, the authors explored a multimodal fusion setting by integrating MVNN's visual features with textual models. These fusion models yielded ever grater performance. The AttRNN+MVNN model reached accuracy of 90.1%, while EANN+MVNN and MVAE+MVNN scored 89.7% and 89.1%, respectively. These results represented more than a 5.2% improvement over other visual baselined, proving that the quality of visual representations provided by MVNN significantly enhances multimodal fake news detection.

Several case studies illustrated MVNN's unique capabilities. For example, the model successfully identified misleading reused images that pixel-only models failed to detect. It also accurately classified emotionally charged but realistic-looking imaged that frequency only models misclassified. This ability to capture both tampering and context-driven deception underscores the utility of a dual-domain visual approach.

Qi made several key contributions through his work. They were the first to jointly model both frequency and pixel domains for fake news image detection, offering an end-to-end architecture that captures both physical manipulation and semantic inconsistency. The MVNN achieved state-of-the-art performance in both standalone visual detections and text-image fusion tasks, positioning it as a powerful tool for future applications. Looking ahead, the authors suggested expanding their dataset to include Twitter and other platforms to support cross-platform generalization. They also proposed exploring semantic alignment between text and image to better capture

multimodal inconsistencies, and emphasized the importance of developing explainable models to improve transparency and trust in automated fake news detection systems.

Conclusion of Deep Learning Techniques

Deep learning approaches to fake news detection have introduced a paradigm shift in the field by moving beyond manually engineered features and leveraging automatic representation learning from large-scale multimodal data. Techniques such as Convolutional Neural Networks (CNNs), Graph Convolutional Networks (GCNs), and hybrid architectures like MVNN have demonstrated impressive capabilities in capturing nuanced linguistic, visual, and structural patterns that are often difficult to model with traditional machine learning. These models excel in learning deep semantic cues, exploiting propagation dynamics, and integrating complex multi-domain signals, leading to high accuracy and robustness, even in early detection scenarios. For instance, graph-based models have proven highly resistant to adversarial manipulation, while multimodal fusion networks can detect subtle visual misinformation overlooked by text-focused methods. However, deep learning approaches also come with notable challenges. They often require extensive labeled data, are computationally expensive to train and deploy, and tend to behave as "black boxes", limiting interpretability. Moreover, their performance can degrade when applied across platforms or domains due to distribution shifts, raising concerns about generalizability. Despite these limitations, the adaptability and predictive power of deep learning models position them as a cornerstone for the future of scalable, real-time fake news detection systems.

2.3.3 Hybrid Detection Systems

Hybrid detection systems in fake news research represent an evolving frontier that combined the strengths of both traditional machine learning and modern deep learning approaches. Rather than relying solely on either handcrafted features or deep neural representations, hybrid models seek to fuse multiple sources of information, textual content, metadata, user behavior, visual signals, and propagation dynamics, into a unified detection framework. These systems often use ensemble pipelines, multimodal fusion architectures, or layered decision-making processes to achieve more robust and

context aware classifications. By integrating linguistic analysis with social network cues or blending convolutional layers with probabilistic reasoning, hybrid models are uniquely positioned to address the complex and multi-dimensional nature of misinformation. While they introduce added architectural complexity and higher computational demands, hybrid approaches offer significant gains in adaptability, scalability, and resilience, especially in real-world, high-noise environments where misinformation exploits multiple modalities simultaneously

2.3.3.1 CSI: A hybrid Deep Model for Fake News Detection

In their landmark study "CSI: A Hybrid Deep Model for Fake News Detection", Ruchansky and the others (2017) introduced the CSI model (short for Capture-Score-Integrate), a hybrid deep learning framework designed to classify news articles as fake or real while simultaneously identifying users who contribute to the propagation of misinformation. The model integrates three interconnected components: a content analysis module (Capture), a user behavior module (Score), and a fusion layer (Integrate) that unifies both inputs for final classification.

The datasets employed in this study comprised two major social media platforms, Twitter and Weibo, offering a wide range of engagements dynamics. The Twitter dataset included 992 articles (498 fake, 494 real), over 233.000 users, and nearly 600.000 engagement records. The Weibo dataset was significantly larger, with 4.664 articles (2.313 fake, 2.351 real), nearly 2.82 million users, and more that 3.75 million engagements. Each engagement was recorded as triplet (eijt), indicating the user, article, and timestamp.

The first module, Capture analyzes textual and temporal signals from the articles. Input features include n (number of engagements), Δt (time between engagement), xu (user features processed through Singular Value Decomposition), and xt (article text encoded via Doc2Vec). These inputs are embedded and passed through LSTM layer to produce a comprehensive article representation.

The second component, Score, evaluates each user by generating a representation vector and computing a suspiciousness score (si). This score is critical for identifying users who systematically engage with misleading content. High si values were linked to users who promoted fake news early, frequently, and within tightly connected biclusters. Representation correlation analysis showed that cosine distances between user vectors were positively correlated with engagement disparities (p=0.631 on Twitter and 0.867 on Weibo), and Jaccard distances were correlated with differences in suspiciousness scores (0.36 and 0.21 respectively).

In the third component, Integrate, the article representation (uj) and the average suspiciousness score of its engaging users (pj) are concatenated. This fused vector is passed through a fully connected layer with SoftMax activation to yield the final classification output.

Technical implementation involved an embedding layer, the Adam optimizer, dropout (0.2), L2 regularization, and 5-fold cross-validation with an 80/15/5 training-test-validation split. Performance comparisons demonstrated the superiority of CSI across all baselines. On twitter, CSI achieved 89.2% accuracy and 0.894 F1 score; on Weibo, it reached 95.3% accuracy and 0.954 F1 score. These results were statistically significant with p-values less than 0.01.

Additionally, spectral clustering (k=5) was used to analyzed temporal and textual embeddings, revealing that fake news can be categorized into distinct subtypes, including satire and spam. The model's ability to handle both article-level and user-level detection, without reliance on manually engineered features or graph structures, emphasizes its versality and practical utility.

In conclusion, the CSI model sets a strong precedent for hybrid approaches in fake news detection. Its modular design captures textual, temporal, and behavioral dimensions simultaneously, offering a compact and scalable solution with high accuracy and interpretability. The dual focus on both misinformation content and user-behavior broadens its applicability and makes it especially relevant for real-time deployment in dynamic online environments.

2.3.3.2 Beyond News Contents: The Role of Social Context for Fake News Detection

In their study, “Beyond News Contents: The Role of Social Context for Fake News Detection”, Shu and the rest of the its team proposed TriFN, a hybrid framework that integrates multiple dimensions of information to enhance the detection of fake news on social media. Recognizing that fake news dissemination is not merely a textual problem but one deeply rooted in social behaviors and ideological affiliations, TriFN combines news content, user engagements, and publisher bias into a unified modeling structure.

The model is evaluated on the FakeNewsNet dataset, which includes both the BuzzFeed and PolitiFact subsets. The BuzzFeed dataset consists of 182 news articles, equally divided into 91 fake and 91 true articles. It also includes 15.257 unique users, 25.240 engagement interactions, 634.750 social links, and 9 distinct publishers. The PolitiFact dataset includes 240 articles (120 fake and 120 true), 23.865 users, 37.259 engagements, 574.744 social links, and 91 publishers. These datasets are designed to capture a broad view of how fake news is consumed and shared within social ecosystems.

The TriFN architecture is built upon five core components. The first is News Content Embedding, which uses Non-negative Matrix Factorization (NMF) on a bag of words representation to derive latent news features D and word embeddings V. This ensures that semantic representations of articles are effectively captured.

The second component, User Embedding, also leverages NMF, but on the user-user adjacency matrix $A\varepsilon\{0,1\}^{mxm}$, under the assumption that users who are socially connected tend to have similar sharing preferences. This results in the latent metrics U and a correlation matrix T, with Y, serving as a mask for observed user interactions during negative sampling.

The third component, User-News Interaction Embedding, incorporates a nuanced credibility model. Each user is assigned a credibility score $c_i\varepsilon[0,1]$, which is inferred based on clustering patterns and coordination behavior. High-credibility users are more likely to share verified information, while low-credibility users tend to propagate fake news. These interactions are embedded using a unified loss function that jointly models user behavior and news credibility.

The fourth aspect is Publisher-News Relation Embedding, where TriFN encodes the partisan bias of publishers using a label vector oε{-1,-1} representing left-leaning, neutral, and right-leaning positions, respectively. A normalized matrix B links news articles to their publishers, while a projection vector q maps news embeddings to publisher bias.

Finally, the fifth component is a Semi-Supervised Classifier. This classifier leverages both labeled and unlabeled data, mapping the latent news representations to binary labels using a linear mapping. The overall model is optimized with a final joint objective function:

$$\min_{D,u,v,T,\rho_J q \geq 0} \|X - DV^T\|_F^2 + a\|Y \circ (A - UTU^T)\|_F^2 + \beta\, tr(H^T LH)$$
$$+ \gamma \left\| e \circ \left(\bar{B} D_{q-o}\right) \right\|_2^2 + \eta \left\| D_{LP-yL} \right\|_2^2 + \lambda R$$

Here, H=[U;D_L] represents the concatenation of user and news embeddings, and L=S-F is the Laplcian matrix. The regularization term R ensures model robustness while alternating least squares (ALS) is used to update all model parameters iteratively until convergence.

An ablation study was conducted using three model variants TriFN\P (excluding publisher modeling), TriFN\S (excluding user-user and user-news modeling), and TriFN/PS (content only baseline). Results show that removing any of these components leads to a performance drop, confirming the value of each module.

Performance-wise, TriFN consistently outperforms traditional baselines. On the BuzzFeed dataset, TriFN achieved an accuracy of 0.864, surpassing the best baseline LIWC+Castillo (0.825). On PolitiFact, TriFN reached 0.878, again outperforming the baseline (0.829). These results underline TriFN's strength in fusing textual semantics with behavioral and contextual cues.

TriFN offers a comprehensive and scalable approach to fake news detection by integrating semantic content, social behavior, and publisher reliability. It demonstrates strong generalizability even under weak supervision or incomplete feature spaces and significantly improves performance over classical feature-based models such as RST, LIWC, and user-centric approaches.

2.3.3.3 False Information Detection on Social Media via a Hybrid Deep Model

As part of the study "False Information Detection on Social Media via a Hybrid Deep Model", Lianwei Wu et al. propose a comprehensive solution to the growing challenge of distinguishing various types of misinformation. Unlike traditional fake news research, which often restricts its scope to binary classification tasks such as true versus fake or rumor detection, this work broadens the perspective by addressing five distinct types of online content: True Information, Rumors, Biases, Fake News, and Spams. The model aims to improve credibility evaluation by simultaneously learning semantic (context aware) and emotional (sentiment) representations of textual data.

Each information type is evaluated along three key dimensions: purposiveness (the intent behind sharing information), harmfulness (its societal impact), and credibility (its trustworthiness). True information is defined as scientific, objective, and complete, with purposive intent, no harmful impact, and high credibility. Rumors are plausible but unconfirmed, with unclear purpose, moderate harmfulness, and unverified credibility. Biases, often exaggerated or emotionally charged, carry strong intent and high harmfulness, but uncertain credibility. Fake news is characterized as deliberately false and deceptive, with a strong purpose to mislead, substantial harmful impact, and confirmed falsehood. Finally, spams are generally useless or confusing, with weak harmfulness and often unclear or automated intent.

Two datasets were used to train and evaluate the model. The LIAR dataset, sourced from PolitiFact.com, contains 12.836 short human-annotated statements labeled across six fine-grained categories: pants-fire, false, barely-true, half-true, mostly-true, and true. The second dataset, collected from Sina Weibo, comprises 40.000 microblogs evenly distributed across five categories of information types. In terms of average text length and virality, fake news had the longest posts (115 characters on average) and a maximum forwarding volume of 220.1k, while spam had the shortest length (93 characters) and the lowest virality at 150.3k forwards. True information averaged 104 characters and reached 224.5 forwards, while rumors and biases posted averages of 111 and 98 characters, respectively, with high forwarding volumes (213.6k and 232.7k).

The model architecture integrates two complementary modules: a semantic representation module and a sentiment analysis module. The semantic module is built on a Recurrent Convolutional Neural Network (RCNN) that captures context-aware features by combining left context, the word embedded using 50-dimensional Word2Vec embeddings, with a window size of 4. The RCNN output is a max-pooled semantic vector that represents the meaning of the microblog post.

The sentiment representation module aims to extract emotional signals from the text using sentiment lexicons and dictionaries of negative and degree words. This module combines a Bi-directional Long Short-Term Memory (BiLSTM) layer followed by a Convolutional Neural Network (ConvNet) and max-pooling operations. The resulting sentiment vector captures the affective tone of the message.

These two output vectors, semantic and sentiment, are then concentrated and passed through fully connected layers with a SoftMax activation function to perform final classification across the five information classes. The model was trained using cross-entropy loss and gradient descent optimization with a learning rate of 0.001, a batch size of 50, and 100 filters. The dataset was split 70% for training and 30% for testing.

Experimental results show that this hybrid method outperforms several traditional baselines. On the LIAR dataset, logistic regression reached an accuracy/F1 of 0.263/0.342, while SVM and CNN achieved 0.271/0.354 and 0.273/0.355, respectively. The RCNN + Sentiment model achieved the highest performance at 0.337/0.431. Similarly, on the Weibo dataset, the proposed model outperformed all baselines with an accuracy of 0.433 and an F1 score of 0.549, compared to 0.402/0.511 for RCNN-only and even lower scores for SVM and CNN.

The study highlights several advantages of this hybrid model. Notably, it captures deeper contextual and emotional cues, supports multi-class classification, and moves beyond binary fake news detection. However, there are limitations. The use of static sentiment lexicons is relatively shallow and fails to fully capture nuanced or context-dependent sentiment, and the relatively limited dataset size constrains the generalizability of the model.

Summing up, this study successfully introduces a novel hybrid deep learning model that merges semantic and sentiment analysis for more granular and comprehensive false information detection. It provides a significant performance boost over baseline

classifiers and encourages future work on multi-class detection frameworks, deeper linguistic analysis, and enhanced sentiment understanding in misinformation detection.

Conclusion of Hybrid Techniques

Hybrid detection systems represent a powerful evolution in the fight against misinformation, combining the strengths of both traditional and deep learning methods to create more nuanced, multi-dimensional models. These approaches integrate diverse signals (including textual semantics, user behaviors, temporals patterns, sentiments cues, and network-level interactions) into unified frameworks that can capture the complex dynamics of fake news dissemination. One of the major strengths of hybrid models is their ability to leverage both content-based analysis and contextual metadata, resulting in higher classification accuracy and improved generalization across different platforms and datasets. For example, models like CSI and TriFN demonstrated not only strong performance in identifying fake content but also provided user-level insights such as suspiciousness scoring, temporal clustering, and credibility inference, which are crucial for real-world deployment. Moreover, hybrid systems often scale better with social media data, as they are capable of incorporating real-time engagement signals and user interaction patterns.

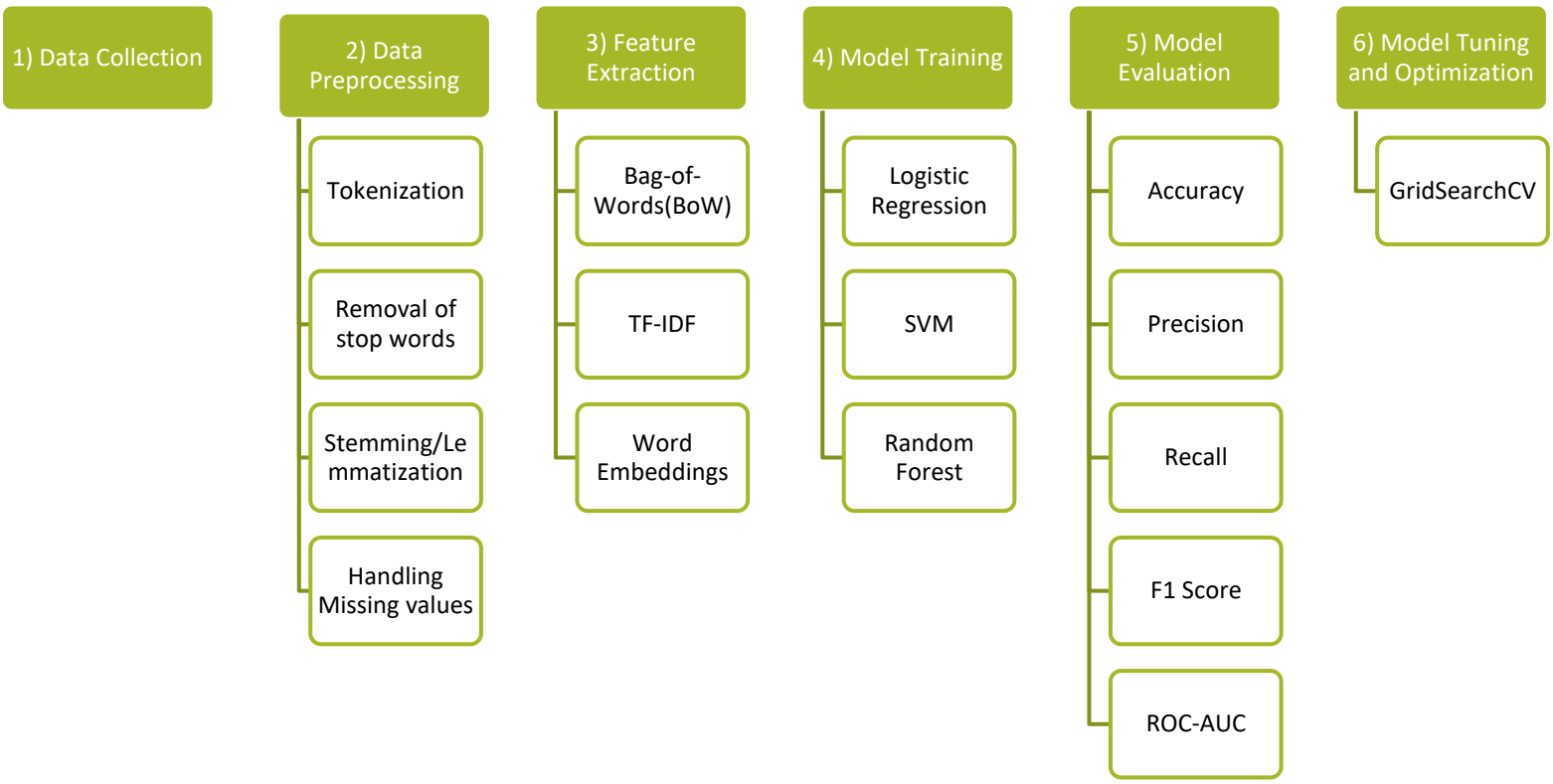
However, hybrid techniques are not without limitations. Their complexity introduces significant computational overhead, requiring careful tuning of multiple components and integration pipelines. The reliance on user or propagation data also poses challenges in terms of data availability, privacy concerns, and model interpretability. Additionally, while these systems often outperform single-modality models, their performance can be hindered in low-resource settings or where one of the input modalities (e.g., user metadata or social graphs) is missing or unreliable. In some cases, shallow integration of features, such as using static sentiment lexicons or handcrafted metadata, can limit the full potential of the model. Despite these challenges, hybrid detection approaches strike a promising balance between interpretability, performance, and adaptability, making them a key direction for future research in multi-modal, real-world fake news detection systems.

2.4 Datasets, Preprocessing and Evaluation Framework

As introduced in previous sections, various machine learning, deep learning, and hybrid approaches have been applied in the field of fake news detection. These methods rely heavily on the availability of high-quality datasets, appropriate preprocessing pipelines, and well-defined evaluation metrics. While earlier chapters discussed techniques employed in specific research studies, this section serves as a comprehensive glossary of the methodological and technical foundations that support fake news detection effort. It combines elements referenced in the analyzed literature with additional educational and practical tools frequently used this domain. This framework outlines each step in the fake news detection pipeline, from data collection to final model evaluation and tuning

2.4.1 Methodology Overview

The workflow, depicted in the diagram above, outlines the typical sequence oh phases involved in building a fake news detection system. Some of the steps are directly reflected in the articles discussed previously, while others are included for completeness and to provide broader context for readers new to the subject. In the following subsections, we explore each stage in greater detail.



2.4.2 Data Collection

The data collection phase involves the acquisition of relevant datasets from validated and reputable sources, as outlined in the previous section. These datasets were carefully selected to ensure the reliability, diversity and credibility necessary for the purposes of this research.

2.4.3 Commonly Used Datasets

Several publicly available datasets have been utilized in fake news detection studies. These datasets vary in size, type, and modality, providing a broad spectrum of data for training and testing fake news detection models

- **PolitiFact**: This dataset contains fact-checked political statements. It has been used in various studies for fake news detection and analysis.
- **BuzzFeed**: Similar to PolitiFact, this dataset includes political news articles fact-checked by BuzzFeed journalists.
- **LIAR**: This is a benchmark dataset for fake news detection, containing short statements labeled as true, mostly true, half true, mostly false, false, and pants on fire.
- **FakeNewsNet**: This repository includes two comprehensive datasets PolitiFact and GossipCop, which have been widely used for fake news detection.,
- **FC-1**: FNC-1 was designed as a stance detection dataset and it contains 75,385 labeled headline and article pairs. The pairs are labelled as either agree, disagree, discuss, and unrelated.
- **Weibo**: This dataset contains Chinese microblogging website Weibo's posts, labeled for fake news detection.
- **CoAID (COVID-19 Healthcare Misinformation Dataset):** A collection of fake and real news articles, tweets and user engagement related to COVID-19.
- **ReCOVery Dataset**: A large-scale dataset containing COVID-19 related news articles with credibility ratings and metadata.

- **WEKFake:** A dataset including both real and synthetically generated fake news using GANs (generative models).
- **COVID-Lies:** A collection of tweets containing misinformation or rumors about COVID-19, manually labeled.
- **TI-CNN (Text & Image-based Fake News Dataset)**: News articles with both text and images, designed to build models that can detect fake news using bot modalities.
- **FakeHealth**: Focused on health-related misinformation, providing real and fake news articles in the medical domain.
- **GossipCop (part of PolitiFact)**: Celebrity-related rumor articles, fact-checked and labeled as real or fake.
- **Snopes:** A dataset of fact-checked claims collected from Snopes.com, one of the oldest fact-checking websites.

Dataset	Type/ Modality	Domain/Focus	Size/Labels	Public	URL
PolitiFact	Text (fact - checked statements)	Politics	Labeled as true/false etc.	Yes	https://www.kaggle.com/datasets/rmisra/politifact-fact-check-dataset
BuzzFeed	Text (political news)	Politics (U.S.)	Articles labeled as real/fake	Yes	https://www.kaggle.com/datasets/konradb/buzzfeed-news-2018-2023
LIAR	Text (short claims)	Politics	12,800 statements, 6 labels: true, mostly true, half true, mostly false, false, pants on fire	Yes	https://www.kaggle.com/datasets/doanquanvietnamca/liar-dataset
FakeNews Net	Text (news articles)	Politics, social media	Combines Politifact & GossiCop	Yes	https://www.kaggle.com/datasets/mdepak/fakenewsnet
FC-1 (FNC-1)	Text (headline - article pairs)	General news	75,385 pairs labeled agree/disagree/discuss /unrelated	Yes	https://github.com/FakeNewsChallenge/fnc-1

Weibo	Social media posts	Chinese microblogging	Posts labeled fake/real	Yes	https://www.scidb.cn/en/detail?dataSetId=f86499e16e8f4261a01ace0deadd96aa
CoAID	Text + social media	COVID-19 misinformation	Fake & real news, tweets, engagement data	Yes	https://github.com/cuilimeng/CoAID
ReCOVery	Text (news articles)	COVID-19	Large-scale with credibility ratings & metadata	Yes	https://github.com/apurvamulay/ReCOVery
WEKFake	Text (real + synthetic)	General news	Articles generated via GANs + real	Yes	https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification
COVID-Lies	Text (tweets)	COVID-19	Manually labeled misinformation tweets	Yes	https://github.com/ucinlp/covid19-data/releases
TI-CNN	Text + Image	General news	Multimodal (text + image) articles	Yes	https://www.kaggle.com/datasets/chahatraj/ticnn-extracted-images
FakeHealth	Text (health news)	Health misinformation	Real & fake medical news	Yes	https://github.com/EnyanDai/FakeHealth
GossipCop	Text (celebrity rumors)	Entertainment	Celebrity rumor articles fact-checked	Yes	https://www.kaggle.com/datasets/akshaynarayananb/gossipcop
Snopes	Text (fact – checked claims)	General, multi-domain	Wide range of claims labeled true/false/mixed	Yes	https://www.kaggle.com/datasets/ambityga/snopes-factnews-data

2.4.4 Data Preprocessing

The data preprocessing phase involves preparing the raw dataset for analysis. Some relevant methods are:

- **Tokenization:** Splitting textual data into individual units, typically words or phrases, called tokens. This step is essential as it transforms unstructured text into a structured format that can be further analyzed.
- **Removal of stop words:** Stop words, such as “the”, “is”, and “and”, are common words that carry minimal semantic meaning and can introduce noise in textual analysis. By removing stop words, the dimensionality of the dataset is reduced, and the focus remains on more informative terms.
- **Stemming/Lemmatization:** Stemming reduces words to their root forms by removing suffixes, while lemmatization maps words to their base or dictionary form. Applying these techniques minimizes word variations, ensuring the similar words are treated consistently during analysis.
- **Handling Missing Values:** Missing values in the dataset can lead to biased or incomplete analysis if not addressed appropriately. Common techniques for handling missing data include records, depending on the context and severity.

2.4.5 Feature Extraction

The feature extraction phase involves transforming textual data into numerical representations that machine learning models can effectively process. The primary techniques used include:

- **Bag-of-Words (BoW):** The Bag of Words model represents text as a collection of individual words, disregarding grammar and word order but preserving word frequency. This technique converts each document into a fixed-length vector based on the occurrence of words from a defined vocabulary.
- **TF-IDF (Term Frequency-Inverse Document Frecuency):** TF-IDF enhances the Bag-of-Words model by weighting terms according to their importance, giving less weight to common words and more to rare but

informative ones. It helps highlight distinctive terms within documents, improving the discriminative power of machine learning algorithms.

- **Word Embeddings:** Map words into continuous vector spaces where semantically similar words are positioned closer together. Techniques like Word2Vec and GloVe enable models to capture deeper semantic and syntactic relationships beyond simple word frequency.

2.4.6 Model Training

The extracted features are utilized to train various machine learning models aimed at distinguishing between real and fake news articles. Some tradition machine learning models commonly used in previous research include:

- **Logistic Regression:** A widely used linear model for binary classification tasks, effective in high dimensional feature spaces.
- **Support Vector Machines (SVM):** A powerful classifier that seeks to find the optimal hyperplane separating classes with maximum margin.
- **Random Forest:** An ensemble method based on decision trees that improves predictive accuracy and reduces overfitting.

In addition to traditional machine learning models, deep learning approaches have also been increasingly applied in fake news detection, such as:

- **Convolutional Neural Networks (CNNs):** Typically used for spatial feature extraction, CNNs have been adapted to process textual data for capturing local patterns and hierarchical representations.
- **Recurrent Neural Networks (RNNs):** Particularly useful in sequential data like text, RNNs and their variants (e.g., LSTM, GRU) can model dependencies between words over time.
- **Transformers and BERT-based Models:** Recent research has leveraged models like BERT (Bidirectional Encoder Representations from Transformers), which capture rich contextual relationships and have achieved state-f-the-art results in text classification tasks.

2.4.7 Performance Evaluation Metrics

After training, the performance of each model is assessed using a set of well-established evaluation metrics. These metrics provide quantitative insight into the model's ability to distinguish between fake and real news, and help guide the selection of the most effective approach.

The evaluation metrics include:

- **Accuracy:** Measures the overall correctness of the model by calculating the proportion of correctly classified instances among all predictions.

$$\frac{TP + TN}{TOTAL}$$

- **Precision**: Indicates the proportion of true positive predictions among all positive predictions, reflecting the model's reliability in identifying fake news.

$$\frac{TP}{TP + FP}$$

- **Recall (Sensitivity):** Represents the proportion of actual fake news articles that were correctly identified by the model.

$$\frac{TP}{TP + FN}$$

- **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure especially useful in imbalanced datasets.

$$\frac{2TP}{2TP + FP + FN}$$

- **ROC-AUC (Receiver Operating Characteristic -Aer Under Curve):** Assesses the model's ability to distinguish between classes at various threshold settings, providing an aggregate measure of performance.

2.4.8 Model Tuning and Optimization

Model tuning refers to the process of adjusting hyperparameters to improve model performance. Proper optimization helps achieve better generalization, reduces overfitting, and ensures more accurate predictions. Some common techniques are:

- **Grid Search**: Exhaustive search over a predefined set of hypermeters values.
- **Random Search**: Randomly samples hyperparameter combinations to find near-optimal settings.
- **Bayesian Optimization**: Uses probabilistic models to guide the search for best parameters.
- **Cross-Validation**: Evaluates performance across different data splits for more reliable tuning.

2.4.9 Technologies Used

This section outlines most of the various technologies used in this research, explaining how each contributed to the project and providing links for further information in the title.

Python

Python was the primary programming language for this project, chosen for its extensive libraries and frameworks that facilitate data manipulation, machine learning, and natural language processing.

Jupyter Notebook

Jupyter Notebook provided an interactive environment for writing and running code. Its ability to combine code, text, and visualizations in a single document was invaluable for exploring data, developing models, and presenting results.

Numpy

Numpy was essential for numerical computing, enabling efficient handling of arrays and metrices. It played a critical role in data manipulation and preparation tasks.]

Pandas

Pandas was used for data manipulation and analysis. Its powerful data structures, like DataFrames, made it easy to clean, transform, and aggregate the dataset.

Scikit-learn

Scikit learn provided a comprehensive suite of machine learning algorithms and tools for model training, evaluation, and validation. It was instrumental in implementing and testing various classifiers.

Matplotlib and Seaborn

Matplotlib and Seaborn were used for data visualization. They enabled the creation of detailed and informative plots to analyze the dataset and present results, helping to visualize data distribution and model performance metrics.

Visual Studio Code

VS Code was the primary development environment for writing and debugging code locally. Its extensive range of extensions and integrations supported the development process efficiently.

2.5 Comparative Analysis of key Studies

After exploring the logical, theoretical, and technical foundations of fake news detection in the preceding sections, we now turn our attention to a comparative analysis of prominent research articles that have shaped this domain. This section offers both granular, article-by-article summary and a cross-methodological discussion, highlighting strengths, weaknesses, and common challenges.

While the studies primarily focus on academic and research-based applications, we also incorporate auxiliary insights drawn from foundational datasets, prevalent preprocessing strategies, and known systemic challenges (e.g., data quality and generalizability). Together, these elements create a reference glossary of established techniques and experimental deigns, offering a comprehensive overview oh how fake news detection has evolved over time across varying modalities and methodologies.

We begin with a comparative table, followed by detailed written summaries of each study. The section concludes with a cross-cutting thematic synthesis and an evaluation of the broader methodological landscape.

2.5.1 Summary of Selected Research Articles

The following table presents a synthesized overview of 11 major studies that span different methodological categories: traditional machine learning, deep learning, and hybrid approaches.

Study & Year	**Dataset(s)**	**Methodology**	**Strengths**	**Limitations**
Wu et al. (2015)	Twitter (propagation trees)	Random Forest on propagation features	Early-stage detection via diffusion patterns	Limited to propagation; lacks content insight
Shu et al. (2017)	FakeNewsNet (PolitiFact, GossipCop)	Ensemble of content, user, social features	Multi-source context modeling	High complexity and reliance on metadata
Wang (2017)	LIAR	CNN+metadata	High quality dataset; good use of metadata	Struggles with short statements; 27% accuracy
Monti et al. (2019)	Twitter (Snopes, BuzzFeed)	Graph Convolutional Network (GCN)	Strong early detection and propagation modeling	Computational cost; needs structured graph data
Qi et al. (2019)	Weibo (visual news)	MVNN (multi domain CNNs+BIGRU)	Visual detection via pixel + frequency domains	Requires image-rich datasets; complex architecture
Ruchansky et al. (2017)	Twitter & Weibo	CSI model (LSTM+user scoring)	Dual modeling of content & user behavior	Needs sequential & user interaction data

Shu et al. (2019)	FakeNewsNet	TriFN (matrix factorization + semi-supervised learning)	Publisher & user modeling	Relies on good publisher labeling
Perez-Rosas et al. (2018)	FakeNewsAMT, Celebrity News	SVM with linguistic features	Language based detection with high precision	Dataset not based on real-world misinformation
Tacchini et al. (2017)	Facebook (user likes)	Logistic Regression, Boolean Label Crowdsourcing	Strong performance with minimal content	No generalization to unseen users
Rashkin et al. (2017)	PolitiFact + Gigaword, The Onion	Naïve Bayes, LSTM, MaxEnt	Multi-genre analysis of deceptive cues	Low performance in 6-class classification
Wu et al. (2020)	LIAR, Weibo	Hybrid RCNN+ sentiment CNN	Multi-class detection: rumors, spam, bias, etc.	Sentiment lexicons limited in nuance

Cross-Study Insights and Thematic Observations

- **Dataset Diversity and Limitations**

From this comparative review, several themes and challenges emerge:

1. Language Bias: Most datasets are English-centric. Multilingual misinformation detection is under-researched and underserved by current data resources.
2. Context Narrowness: Many datasets are domain-specific (e.g., political, health, or celebrity topics), meaning models trained on them often underperform in unfamiliar domains.

3. Data Generation Issues: Some datasets like FakeNewsAMT include synthetic fake news created via crowd-sourcing. While controlled, these examples lack the deceptive sophistication of real-world misinformation.
4. Annotation Quality: Datasets vary significantly in label quality. Some are expertly annotated (e.g., LIAR, PolitiFact), while others depend on community flagging or mechanical turk workers.
5. Evolution and Obsolescence: News and social content evolve rapidly. Without ongoing dataset updates, detection models risk becoming outdated or ineffective in the face of new narrative formats and dissemination tactics.

These limitations emphasize the urgent need for continuously refreshed, cross-platform, and multilingual datasets that reflect real-world information dynamics.

- **Methodological Trade-Offs**

1. **Content-Based vs. Propagation-Based Approaches**

Content-focused models (e.g., Perez-Rosas, Rashkin) extract linguistic or visual patterns directly from articles, achieving high precision in controlled settings. However, they are slower to act and vulnerable to adversarial writing.

Propagation-based models (e.g., Wu 2015, Monti) leverage structural features of information spread. These enable early detection and often require less semantic interpretation but may miss content-level deception.

2. **Traditional ML vs. Deep Learning vs. Hybrid Models**

Traditional models (e.g., SVM, Logistic Regression) are fast, interpretable, and effective with well-engineered features, but lack adaptability to new patterns or modalities.

Deep learning models (e.g., CNNs, GCNs, LSTMSs) offer automatic feature learning, scalability, and better multimodal integration, but they often behave as black boxes, are data-hungry, and require expensive infrastructure.

Hybrid systems (e.g., CSI, TriFN, MVNN) that combine textual, social, and visual inputs consistently outperform single-modality approaches. However, they are more computationally intensive and difficult to deploy at scale.

3. **Social and Behavioral Signals**

Multiple studies (e.g., Tacchini, Shu 2017) highlight the importance of user behavior, network effects, and engagement patterns in amplifying or validating misinformation. Incorporating user metadata (likes, followers, frequency of reposting) improves robustness, through these signals are often platform-specific and hard to generalize.

4. **Model Interpretability and Deployment Constraints**

Despite high accuracies, many deep learning models lack transparency, a critical requirement in domains like journalism, elections, or public health. Also, real-time applications demand not only speed but reliability under pressure, including robustness against adversarial inputs and misinformation campaigns.

2.5.2 Advantages and Limitations Across Methodologies

- **Advantages**

1. High Accuracy with Deep Models: Deep architectures like CNNs, GCNs, and RCNNs achieve state-of-the-art results, especially when trained o rich, multi-layered datasets.
2. Multimodal Fusion: Integrating text, images, metadata, and propagation features enables more holistic modeling of fake news ecosystems (e.g., MVNN, CSI, TriFN).
3. Early Detection via Propagation: Graph-based methods and temporal models can detect fake news before it gains viral traction.
4. Metadata and Social Features: User engagement data, when available, offers strong auxiliary signals even in the absence of full article content.

- **Limitations**

1. Dataset Constraints:
   - Many are domain-limited, synthetic, or outdated.
   - Multilingual and cross-cultural misinformation detection is underexplored.
   - Label inconsistencies and sparse data (especially in real-time) remain hurdles.
2. Interpretability Issues:
   - Deep models are often opaque (“black boxes”).
   - Explanations are vital for public trust, legal validation, and platform accountability.
3. Generalizability and Transferability:

- Models trained on one platform (e.g., Twitter) often perform poorly on others (e.g., Facebook, Reddit).
- Temporal drift and content evolution degrade performance unless models are regularly fine-tuned.

4. Computational Demands:
   - Training deep models requires significant computational resources, including GPUs.
   - Hyperparameter tuning and large-scale inference increase system costs.
5. Real-Time Constraints:
   - Real-time labeling is hindered by the delay in acquiring human annotations.
   - Adversarial inputs and misinformation tactics evolve quickly, requiring adaptive defenses.

Chapter 3

Agentic AI

3.1 Introduction to Agentic AI

As fake news continues to evolve in complexity and spread rapidly through digital ecosystems, traditional machine learning and even deep learning models often fall short in addressing the full scope of the problem. These systems, while powerful, typically operate in a monolithic and linear fashion, trained on static datasets, optimized for narrow objectives, and limited in adaptability. In contrast, the rise of Agentic AI introduces a new, dynamic paradigm in which intelligent agents work collaboratively, adapt to changing inputs, and take on specialized roles to solve multifaced problems in real time.

Agentic AI refers to the deployment of artificial intelligence in the form of autonomous or semi-autonomous agents, each equipped with distinct goals, behaviors, and tools. These agents are not just passive responders to inputs, but active participants in the decision-making process, capable of reasoning, communicating with one another, delegating subtasks, and learning from feedback. In essence, they operate more like human teams, with specialization, coordination, and goal-driven behavior, than like traditional algorithms.

In the context of fake news detection, an agentic AI system opens up new opportunities for scalable, modular, and explainable solutions. For example, instead of relying on a single classification model, a multi-agent system might include one agent responsible for content analysis, another for evaluating source credibility, a third for social behavior tracking, and a fourth for integrating evidence and generating an interpretable verdict. This layered, distributed intelligence mimics how human fact-checkers operate, dividing tasks, applying domain knowledge, and verifying conclusions collaboratively.

The recent advances in large language models (LLMs) (such as GPT-4, Claude, Perplexity, Microsoft Copilot) have significantly accelerated the development of agentic systems. When paired with proper prompt engineering, tool access, and defined personalities, these agents can simulate complex cognitive workflows, from data preprocessing and knowledge retrieval to ethical reasoning and user

communication. This creates a flexible backbone for applications that require both analytical rigor and human-like judgment.

This chapter introduces the agentic framework developed in this research as a novel approach to fake news detection. Through the following sections, we will outline the motivation for adopting a multi-agent system, describe the individual agents and their interactions, and highlight the architectural choices made to ensure robustness, transparency, and adaptability. This shift from monolithic models to intelligent distributed reasoning marks a significant evolution in AI-powered misinformation detection and opens up a broader discussion on the future of explainable, modular AI systems.

3.2 Motivation for a Multi-Agent System

While traditional AI models for fake news detection have significant strides, they often suffer from rigid architecture, limited scope, and poor adaptability when faced with the evolving, multimodal, and context-sensitive nature of misinformation. These models typically operate as isolated pipelines, optimized for specific tasks such as content classification or user behavior analysis. However, misinformation is rarely a singular, well-bounded problem. It involves a complex interplay of linguistic manipulation, social propagation, source credibility, emotional bias, and visual deception often simultaneously.

This multi-dimensional nature calls for a more flexible and collaborative approach, which is where multi-agent system becomes not only appropriate but necessary.

A multi-agent framework allows to decompose the fake news detection process into specialized cognitive functions, each assigned to a distinct agent. For example, one agent might excel in semantic analysis of text, while another evaluates the reliability of the news source. A third could focus on image authenticity or social media propagation patterns, while a fourth integrates outputs and generates human-readable explanations. This division of labor leads to better performance, increased transparency, and a clearer separation of concerns, all of which are critical in a field where accuracy and explainability are paramount.

Moreover, this architecture offers scalability and modularity. New agents can be added or modified without disrupting the entire system, allowing it to evolve alongside emerging challenges, such as deepfakes, synthetic media, or changes in user behavior. Each agent operates demi-independently, with internal autonomy and the ability to communicate with others, leading to emergent collaboration that reflects human team dynamics.

Importantly, multi-agent systems provide a strong foundation for incorporating agentic AI principles. Rather than executing only predefined tasks, they can support adaptive behaviors such as reallocating responsibilities, rerunning verification steps, or seeking additional evidence when uncertainty arises. In our architecture, this adaptability is operationalized through the Supervisor Agent, which evaluates the outputs of other agents and triggers follow-up actions if a claim is ambiguous or inconsistencies appear across sources. This approach extends the system's reasoning beyond static automation, enabling more flexible and context-aware handling of complex misinformation cases.

In this thesis, we adopt this agent-based design to address the key shortcomings of prior fake news detection: limited adaptability, poor transparency, and monolithic architecture. Our goal is to demonstrate that agentic intelligence, when properly structured and aligned with the fake news detection pipeline.

3.3 System Design – The Agent Flow Architecture

3.3.1 System Overview

In response to the complex and multifaced nature of fake news detection, we designed a modular multi-agent architecture by large models (LLMs), specifically OpenAI's gpt-4o-mini. This system avoids reliance on a single monolithic model and instead distributes responsibilities across five autonomous agents, each designed to a specialized sub-task. Their outputs are centrally evaluated by a supervising component. This design provides a high degree of explainability, scalability, and modular control characteristics often lacking in traditional black-box deep learning systems.

The gpt-4o-mini model was selected primarily due to its cost-effectiveness, as it offers high performance at low price point, currently priced at $0.0005 per 1000 tokens (as per OpenAI's official pricing at the time of development). In addition to affordability, we chose OpenAI's models because they are among the most linguistic capable LLMs on the market – extensively trained on diverse, high-quality corpora, including Wikipedia, academic content, and multilingual sources. This broad training foundation makes them especially well-suited for semantic analysis and natural language reasoning, which are critical in fake news detection.

While the architecture is implementing using OpenAI models, it is worth noting that it can be replicated using open-source alternatives such as Mistral, LLaMA 3, or Phi-2, which can be hosted locally for cost – free interface. However, for the purposes of this research, we opted to utilize the most capable and production – ready models currently available, prioritizing linguistic richness and tool compatibility. The entire system is orchestrated using Flowise, a visual low-code platform for LLM pipelines. Each agent's prompt is inserted into the "System Prompt" field of its Flowise node, and the flows are executed locally via Node.js through the terminal, providing full control over the development, testing, and deployment without relying on external hosted platforms.

3.3.2 Agents Description

The proposed architecture for fake news detection operates as a distributed, modular system composed of five autonomous agents, each powered by the GPT-4o-mini language model. These agents work in parallel and independently process distinct aspects of the input, culminating in a comprehensive decision-making process supervised by a final aggregation agent. This multi-agent setup enhances transparency, modularity, and explainability, and enables the system to evolve as requirements change, without re-training the underlying models.

The pipeline begins with the Input Agent, which identifies and classified the nature of the user's submission: a URL, a short textual claim, or a document. Based on this, the agent processes the input accordingly. For instance:

- For URLs: {"type": "url", "cleanText": "This appears to be a URL."}
- For textual claims: {"type": "text", "cleanText": "Hillary Clinton said they can leave the really"}
- For documents: {"type": "file", "cleanText": "This is a file input."}

If the input is in a language other than English, the agent translates it and then applies preprocessing routines: punctuation and emoji removal, case normalization, whitespace trimming, and formatting into a strict JSON structure. This consistency ensures full compatibility with downstream agents and avoids ambiguity during parsing.

Following preprocessing, the cleaned input is dispatched concurrently to three processing agents: the Fake News Classifier, Second Search Agent, and Scoring Agent. These agents operate asynchronously, each following a specific role and tool-enhanced prompt.

The Fake News Classifier Agent functions similarly to a supervised ML model. It uses the Google Custom Search API to conduct real-time searches and retrieve contextually relevant sources. The agent is tightly bound to a prompt that demands semantic exactness; vague or approximate matches are rejected. For example:

- "She can leave whenever she wants" ≠ "They kicked her out"
- "Protestors interrupted the speech" ≠ "Security removed protestors by force"

The agent returns a structured JSON object containing:

- Verdict: REAL | FAKE | SUSPICIOUS
- Confidence: 0-100
- Reasoning: Justification based on aligned evidence
- Supporting_sources: A list of verified URLs

The Second Search Agent performs a similar operation using the Serper API, providing redundancy and mitigating single-source bias. Although it does not return a final verdict, it generates a verdict_hint such as LIKELY_TRUE, LIKELY_FAKE, or NO_CLEAR_EVIDENCE. These verdicts are purposefully distinc from the Fake News Classifier's outputs to indicate softer, evidence-weighted impressions. This agent also adheres to semantic fidelity rules and provides a searc_summary describing the general pattern of coverage among retrieved URLs. While it does not directly validate the primary classifier, the supervisor agent uses their agreement to measure alignment and reinforce confidence.

In parallel, the Scoring Agent performs contextual analysis examining the internal structure of the claim. It scores the claim on

- Bias: Degree of ideological slant or partially.
- Emotion: Level of affective or sensational language.
- Credibility: General plausibility or factual soundness.
- Time Relevance: Freshness or timeliness of the event described.

Each metric is mapped to numerical scores (0-100) and qualitative labels (e.g., MEDIUM Bias, HIGH Credibility). For example:

```
{
"bias_score": 70,
"bias_label": "MEDIUM",
"emotion_score": 40,
"emotion_label": "MODERATE",
"credibility_score": 50,
"credibility_label": "MEDIUM",
"time_relevance_score": 40,
"time_label": "POSSIBLY RELEVANT",
"reasoning": "Moderate emotional tone with average credibility"
}
```

These evaluations are generated using linguistic heuristics encoded directly into the agent's prompt, which asks the model to judge stylistic tone, word choice, and narrative framing, making the reasoning both transparent and reproducible.

The Source Quality Checker Agent evaluates the metadata of supporting URLs provided by the two search agents. It uses a domain-based rubric:

- .gov, .edu, or scientific domains -> Very High (85-100)
- Mainstream media (e.g., BBC, NYT) ->High (70-84)
- Independent journalism or reputable .org -> Medium (50-69)
- Blogs, forums, Youtube -> Low (30-49)
- Clickbait or conspiratorial sources -> Very Low (0-29)

The agent aggregates scores across domains, outputs an average, and generates a label like Medium-High, along with explanatory notes:

```
{
"average_source_score": 75,
"source_quality": "Medium-High",
"notes": "Sources include CNN and Politico, both mainstream outlets."
}
```

At the top of the architecture, the Supervisor Agent collects and synthesizes all outputs. It does not initiate communication but waits for:

- Cleaned Input
- Fact-checking verdict and confidence
- Secondary verdict hint and URLs
- Contextual scores
- Source quality assessment

Using a detailed rule-based decision tree (provided inline an in-Appendix A), the agent applies logic such as:

- If both classifiers lean FAKE and credibility < 70 -> Final Call: FAKE
- If verdicts diverge and source quality is low -> Final Call: FAKE, low confidence

- If all indicators align and credibility >=80 -> Final Call: REAL, high confidence

This embedded logic ensures deterministic behavior. The Supervisor also handles missing or malformed outputs by labeling the result as INCONCLUSIVE, a fallback that should be statistically tracked during experimentation.

The full prompt of each agent, provided in Appendix A, defines the exact format, constraints, and behavioral rules. These are not inferred but explicitly declared to ensure control, reproducibility, and auditability.

3.3.3 Prompt Engineering and Agent Personalities

A foundational element of the architecture is its use of explicit, role-specific prompts to define the behavior of each agent. Instead of relying on traditional fine tuning, this system leverages zero-shot prompt engineering, turning each GPT-40-mini instance into a deterministic reasoning module. These prompts are hosted in the Flowise orchestration layer under each agent node's configuration.

Each prompt serves as a "personality template" encoding logic, format, and constraints. For example:

- The Fake News Classifier is instructed to fact-check using only results from the Google Custom Search API and return a structured JSON verdict with confidence and evidence. Its prompt prohibits speculation and enforces semantic match.
- The Second Search Agent follows similar guidelines but uses Serper for retrieval. It is constrained to return a verdict_hint, never a hard conclusion, ensuring redundancy without override.
- The Scoring Agent is guided to analyze emotional tone, bias, and relevance by examining textual features and linguistic structure alone. It provides brief, component-specific explanations.
- The Source Quality Checker scores credibility using domain-based metadata only, never interpreting the claim.

- The Supervisor Agent has prompt-embedded logic for rule-based aggregation. It uses hard-coded thresholds, such as credibility < 50 or contradicting verdicts, to determine final decision and their confidence.

Although the architecture does not use traditional supervised learning to calculate numerical scores, the scoring logic is embedded directly int the system prompts using heuristically defined rules. For instance, the Scoring Agent is instructed to assess textual features such as emotionally charged language, assertive tone, sensationalism, or vague phrasing, and to map these qualitative cues onto 0-100 numerical ranges. These ranges correspond to threshold based-categories (e.g., Bias: LOW= 0-30"), as defined in the prompt itself. Similarly, confidence scores in the Fake News Classifier Agent are derived by instructing the model to weigh the number of supporting sources, the credibility of those domains, and the semantic precision of their content relative to the claim. While not mathematically computed through statistical inference, these values are the result of LLM-guided reasoning grounded in linguistic features and prompt constraints, effectively functioning as a rule-based scoring system powered by natural language understanding.

These prompts rules are shown in full in Appendix A, but their behavior is not black-box. For clarity, an example of Input to Output flow is shown here"

Input: "10 things you never knew about Victoria's Secret Angel Sara Sampaio"

Output:

```
🧠 Fake News Verdict:

Verdict: REAL

Confidence: 85%

Reasoning: The search results include a direct article titled '10
things you never knew about Victoria's Secret Angel Sara Sampaio' which
lists various facts about her, confirming the claim's content.

📊 Scoring Summary:

Bias Score: 25/100 — LOW

Emotion Score: 40/100 — MODERATE

Credibility Score: 65/100 — MEDIUM
```

Time Relevance Score: 80/100 – CURRENT

📋 Scoring Reasoning: The claim is presented as a list of facts about a public figure, which generally indicates low bias. The emotional tone is moderately positive or curious, as it likely attracts interest in the subject. The credibility is medium because while the subject is well-known, the source of the information is not specified, which could affect reliability. The mention of a current celebrity indicates that the topic is relevant to contemporary audiences and not outdated, enhancing its current relevance.

🕒 Timestamp of Evaluation: October 23, 2023

🌐 Secondary Web Search Summary: Secondary Verdict Hint: LIKELY_TRUE Search Summary: The search revealed several articles discussing Sara Sampaio's career and interesting facts about her as a Victoria's Secret Angel. Key points include her initial aspiration to be an actress, her achievements as a model, and her notable roles since joining Victoria's Secret. Sources:
https://www.sarajump.com/news/sara-sampaio-facts
,
https://www.vogue.com/article/sara-sampaio-interview-modeling-career
,
https://www.elle.com/culture/celebrities/a26846187/sara-sampaio-fun-facts/

🕯️ Source Quality Assessment: Average Source Score:

75/100 – Medium-High Notes: The main URL provided is from Vogue, a reputable fashion magazine, which enhances the credibility of the claim. The other sources, while varied, include generally trustworthy outlets like Elle, which also adds to the overall score, but some URLs may include lesser-known sites that could affect the average.

📢 Final Call:

FINAL DECISION: REAL

CONFIDENCE LEVEL: Medium

EXPLANATION:

"The claim is confirmed by a reputable source and has a medium credibility score, suggesting reliable information. The supporting articles align well with the claim, indicating it is likely true."

3.3.4 Tool Integration Strategy

A critical aspect of this architecture lies in its deliberate integration of externals tools that extend the factual grounding and real-time responsiveness of the system. Rather than relying solely on the internal knowledge of the language models, two specialized APIs are incorporated to fetch up-to-date information from the web: Google Custom Search API and Serper API. These tools are independently assigned to two different agents, the Fake News Classifier and the Second Search Agent, thereby enabling redundant and cross-validates fact-checking from distinct search infrastructures.

The Google Custom Search API, utilized by the Fake News Classifier Agent, provides high-precision access to indexed content across trusted domains. This is especially useful for structured evidence retrieval from reputable sources. Its integration ensures that fact-checking is not limited to the model's training data but dynamically anchored in web-based reality. The service operates on a paid model, currently priced at 5$ per 1.000 search queries, and requires configuration through Google Cloud Console, where developers define a programmable search engine and authenticate requests via an API key and custom Search Engine ID (cx)

In parallel, the Serper API supports the Second Search Agent by offering a fast and lightweight alternative with different search algorithms and source coverage. It allows for a broader snapshot of the online discourse surrounding a claim, including support for rapid response use cases. Serper offers a generous free tier of 2.500 queries pre month, after which usage cost 10$ per additional 10.000 queries. Developers can sign up at serper.dev, retrieve their API key, and integrate the tool using the HTTP authentication headers.

Both tools return results in structured JSON, which aligns well with the deterministic design of the agent prompts. Their integration not only provides live, real-world data streams, but also enables each agent to operate independently of the others, enhancing modularity, fault isolation, and overall system robustness. In the revised architecture, both the Fake News Classifier and the Second Search Agent also directly supply their output to the Source Quality Checker Agent, which evaluates the credibility of their retrieved sources. This design decision reflects a hybrid architecture that combines static knowledge from LLMs with dynamic, query-based evidence gathering, enabling a balanced and reliable approach to fake news detection.

3.3.5 Agent Communication Protocol

The communication framework in the proposed architecture follows an asynchronous, parallel, and decoupled protocol, designed to optimize both processing efficiency and structural flexibility. Once the Input Agent processes a user submission - whether a URL, document, or claim – it emits a standardized JSON object that triggers parallel execution across the detection pipeline.

The Fake News Classifier, Second Search Agent, and Scoring Agent independently receive this input and execute their respective tasks. Each agent operates in isolation, guided solely by its own system prompt and tool integrations. There is no inter-agent dependency or sequential execution, which minimizes latency and allows for efficient use of system resources, even when external APIs like Google Custom Search or Serper are involved.

The source Quality Checker diverges slightly: it does not interact with the Input Agent directly but listens exclusively for supporting URLs emitted by the two search agents. This indirect connection models a clear, purpose-specific dependency, where only verified search results are passed forward for credibility scoring, thereby preserving modular boundaries.

All agents communicate using strictly formatted JSON outputs, enforcing consistency, machine readability, and validation at every step. Once the agents complete their processing, or a defined timeout is reached, the Supervisor Agent gathers their outputs performs schema validation, and compiles a comprehensive final decision. Its logic is entirely reactive; it does not initiate any operations but synthesizes results from its peers based on pre-defined reasoning rules.

This loosely coupled model enables scalability in a concrete architectural sense. News agents, such as multimedia analyzers or social network signal detectors, can be integrated into the pipeline without disrupting existing functionality. Likewise, failed components (e.g., due to API outages) do not halt system operation: the Supervisor is designed to degrade gracefully, issuing a partial verdict or making a case inconclusive when necessary.

By avoiding tightly bound dependencies, central orchestration bottlenecks, or fragile sequence chains, the system achieves true operational scalability and maintainability,

measured not by abstract metrics, but by its ability to expand, update, or operate under degraded conditions without architectural rewiring.

3.3.6 Fail – Safe Mechanisms and Error Handling

In high-stakes information systems such as fake news detection pipelines, robustness is no optional, it is foundational. To safeguard the reliability of the agentic architecture, a set of integrated fail-safe mechanisms ensures that system behavior remains predictable, stable, and non-deceptive even under partial failure, tool degradation, or ambiguous input conditions.

A key principle behind these safeguards is strict adherence to structured outputs. Each agent is programmed through its prompt to return only valid JSON objects, matching a pre-defined schema specific to its task. This design prevents the propagation of malformed data through the pipeline and allows the Supervisor Agent to automatically detect formatting errors, missing fields, or content mismatches. In the event that an agent returns incomplete or invalid output, for example, due to API timeout, misfire, or internal parsing issues, the Supervisor does not proceed blindly. Instead, it invokes fallback logic encoded within its own prompt, assigning the claim an “Inconclusive” label and explicitly noting the source of the failure in the final report.

Critically, there is no direct agent-to-agent dependency in execution, meaning that a single point of failure does not compromise the system’s integrity. The input Agent sends its cleaned output in parallel to the Fake News Classifier, the Second Search Agent, and the Scoring Agent. Only the URLs extracted from the first two search agents are forwarded to the Source Quality Checker. This modular communication pattern creates a built-in layer of isolation: if one agent fails, others can still complete their tasks independently, providing partial evidence that can still inform the final analysis.

The Supervisor Agent, as the sole aggregator and decision maker, is equipped to handle such partial information. Its prompt includes logic to detect when one or more expected responses are missing or incomplete. In such cases, the verdict is either downgraded in confidence or marked explicitly as “Inconclusive” to avoid the risk of overreliance or incomplete data. This behavior ensures the system prioritizes epistemic humility over speculation

Furthermore, tool-specific limitations are taken into account. If either the Google Custom Search API or the Serper API experiences connectivity issues or quota overruns, the associated agents are designed to return empty or default JSON payloads without hallucinating data. These behaviors are not inferred by the model; they are hardcoded through declarative prompt instructions. This approach ensures traceability, auditability, and maximum control over failure behavior.

This layered fail-safe strategy, composed of JSON schema validation, agent-level isolation, prompt-level feedback logic, and centralized error detection in the Supervisor, makes the system resilient not just to software or network disruptions, but also to noisy, ambiguous, or malformed user inputs. It enforces predictable degradation instead of silent failure, which is especially important in domains where misinformation can have real-world consequences.

Chapter 4

Evaluate/Test Cases

4.1 Test Dataset Description

To evaluate the real-world performance and interpretability of the proposed multi-agent architecture, a two-phase manual testing process was conducted using a combination of benchmark fake news datasets. The goal of this evaluation was to assess the agent's ability to handle diverse claims, maintain consistency across input types (titles vs. full text), and deliver reliable final verdicts.

The evaluation relied on two widely cited datasets in fake news detection research:

1. LIAR Dataset: A large benchmark dataset consisting of short political statements collected from PolitiFact.com, each labeled as "true", "mostly true", "half true", "barely true", "false", or "pants on fire". For the purposes of this test, only "true" and "false" statements were selected to create a binary classification setup.
2. FakeNewsNet Dataset: This dataset provides full-length news articles labeled as FAKE and REAL, from two independent sources:
    - PolitiFact Subset: Includes articles about political claims verified by PolitiFact.
    - GossipCop Subset: Focuses on entertainment-related articles and celebrity rumors, verified against GossipCop's editorial verdicts.

Both datasets were chosen because they offer distinct styles, subject domains, and text granularities, making them ideal for testing the robustness and generalizability of the detection system.

A total of 80 claims were evaluated, 40 from each dataset, following the procedure below:

From LIAR Dataset:

Selected 20 claims labeled FAKE and 20 labeled TRUE, using random sampling. These 40 samples were merged and then stratified based on input type:

1. 20 ‘titles’ (short claims/headlines)
2. 20 “texts (full statements or slightly extended forms)

This allowed the system to be tested both on brief, context-limited claims and longer, more informative texts.

From FakeNewsNet Dataset:

From PolitiFact: A total of 40 entries were selected, 20 labeled REAL and 20 labeled FAKE. These were then merged into a single pool, from which 20 entries were randomly sampled for testing.

From GossipCop: The same procedure was followed, 20 REAL and 20 FAKE entries were selected, merged, and 20 random entries were chosen from the combined dataset.

4.2 Performance Analysis

To evaluate the effectiveness of the proposed fake news detection system, we conducted a manual testing process using a total of 80 samples (40 form the LIAR dataset and 40 from FakeNewsNet). These samples were evaluated through the full agentic pipeline, producing structured verdicts, confidence scores, source quality ratings, and contextual scores. The goal of this analysis was to assess the classification accuracy, confidence distribution, and agent alignment, as well as identify edge cases and limitations.

Classification Accuracy

Each sample was labeled manually as REAL or FAKE, based on its original dataset label. We then compared these ground-truth labels with the Final Call generated by the system (output of the Supervisor Agent). The results are summarized below:

Dataset	Correct Predictions	Incorrect Predictions	Accuracy (%)

LIAR	33/40	7/40	82.5%
FakeNewsNet	27/40	13/20	67.5%
Overall	60/80	20/80	75%

This performance suggests a strong alignment between the agentic pipeline's logic and real-world ground truths, especially considering the zer0-shot nature of the prompting approach.

Additional Metrics (Precision, Recall, F1 Score)

To further quantify the performance, we calculated standard classification metrics based on the prediction outcomes:

- True Positives (TP): 36 (correctly identified REAL claims)
- True Negatives (TN) :24 (correctly identified FAKE claims)
- False Positives (FP): 7 (FAKE predicted as REAL)
- False Negatives (FN): 13 (REAL predicted as FAKE)

Form these we derive:

1. Precision= TP/ (TP+FP) = 36/ (37+7) = ~83.7%
2. Recall = TP/ (TP + FN) = 36/ (36+13) = ~ 73.5%
3. F1 Score = 2 * (Precision x Recall) / (Precision + Recall) = ~78.2%

Confidence Score Distribution

The Fake News Classifier Agent assigns a confidence score (0-100) for each verdict, reflecting how strongly the returned evidence aligns with the claim. We observed the following.

- Average Confidence (Correct Predictions): 76.4%
- Average Confidence (Incorrect Predictions): 65.5%

This demonstrates that higher confidence tends to correlate with accurate predictions, indicating that the confidence metric is both meaningful and predictive.

Additionally:

- High confidence (>= 80%) verdicts were almost always correct.
- Medium confidence (50-79%) cases occasionally required nuanced interpretation.

- Low confidence (<50%) predictions were more likely to be labeled as "SUSPICIOUS" or incorrectly classified.

Breakdown by Claim Type

To assess how input format affected performance, we analyzed how the system handled titles vs. full texts across both datasets:

Input Type	Accuracy (%)	Common Errors
Titles	76.6%	Semantic ambiguity; vague claims
Texts	70%	Occasional outdated evidence

The results showed that titles achieved a higher accuracy 76.6%, while full texts reached 70%. Titles tended to perform better due to their concise nature, which often led to clearer semantic matching. However, errors commonly occurred due to ambiguity or vagueness in the claims, making it harder for the system to determine context. In contrast, full texts provided richer contextual information, but this sometimes-introduced noise or outdated evidence, which negatively affected classification accuracy.

Source Quality Influence

The Source Quality Checker gent evaluated supporting URLs for each case using a rubric based on domain reliability (e.g., .gov, .edu, mainstream media, etc.). The average source quality score across all correct predictions was 67.4%, compared to 58.25% for incorrect ones.

- Correct Verdicts were mostly supported by domains such as CNN, Reuters, NYT, and Politico.
- Incorrect Verdicts often relied on ambiguous or mixed-credibility sources (e.g., user-generated blogs or clickbait portals)

This correlation supports the robustness of the source reputation scoring rubric, confirming that better source quality improves system reliability.

Scoring Agent Metrics

Each claim was also assessed by the Scoring Agent, which evaluates bias, emotional tone, credibility, and time relevance. Observations include:

- High bias + High emotion often correlated with fake classifications.
- High credibility + Current timestamp scores aligned with true classifications.
- Cases with medium scores across all dimensions tended to be labeled as SUSPICIOUS

For example, a claim with the title “Watch Hillary’s Reaction When Security Tries To Kick Shirtless Male Supporters Out Of Her Rally” was rated as SUSPICIOUS due to low clarity in search evidence, medium bias (70), and moderate credibility (50), even though it matched a real news event. This reflects the system’s emphasis on semantic precision and source confirmation, not just surface-level match.

Chapter 5

Comparison with Traditional Systems

5.1 Explainability Comparison

One of the principal objectives of this thesis has been to explore not only the effectiveness of automated fake news detection but also the transparency and interpretability of such systems. While significant advancements have been made in the field of deep learning and hybrid systems, many state-of-the-art models continue to function as black boxes, producing predictions that lack contextual justifications or traceable reasoning steps. This section offers a comparative analysis of our proposed multi-agent system against various traditional, deep, and hybrid models, specifically in terms of their explainability, decision traceability, and human-aligned output.

Traditional machine learning models such as Support Vector Machines and Logistic Regression have historically been used for fake news detection, particularly in linguistically oriented studies like those Perez-Rosas et Al. (2018) and Rashkin et al. (2017). These approaches rely on manually crafted features (e.g., n-grams, punctuation, LIWC metrics, syntax trees) that offer some degree of interpretability through feature weighs or importance scores. However, the outputs are typically limited to binary or multiclass predictions, with little to no explanation about why a particular verdict was reached, what external evidence was considered, or how user behavior contributed to the decision. Thus, while these models may be computationally efficient and theoretically interpretable, they do not expose their reasoning in a user-friendly or auditable manner.

The limitations in explainability extend to early deep learning models as well. For example, Wang (2017) proposed a hybrid CNN architecture trained on the LIAR dataset that integrated both textual and speaker metadata. While this model achieved better performance than traditional baselines, it still lacked transparency, as its internal convolutional layers and word embeddings offered no direct insight into the verdicts. Similarly, Wu et al. (2018) expanded on this concept by introducing a hybrid model that combined Recurrent Convolutional Neural Networks (RCNN) with sentiments analysis layers. Although their model outperformed prior methods in accuracy and F1-

score, it still failed to provide human-readable rationales or source citations, relying instead on latent vector representations and end-to-end SoftMax classification.

Hybrid systems represent an effort to close the gap between semantic modeling and contextual behavior. Among the most prominent in the CSI model by Ruchansky et al. (2017), which introduced a modular architecture combining article content (Capture), user behavior (Score), and fusion layer (Integrate) to classify fake news. Despite its strong performance, reaching over 89% accuracy on Twitter and 95% on Weibo, the system remains largely opaque to human observers. The "suspiciousness" scores it calculates for users are internal constructs not visible to end users, and no external evidence or contextual justifications are included in its predictions. As such, while CSI excels in capturing behavioral dynamics, its explainability to journalists, researchers, or end users is minimal.

A similar limitation is evident in TriFN by Shu et al., which integrates news content, user engagement, and publisher bias into a unified model. Although is achieves impressive performance on datasets like PolitiFact and BuzzFeed, TriFN's decision-making process is embedded in matrix factorizations and joint objective functions, offering no interpretable outputs for the average user. The system, does not reveal which article features, user interactions, or publisher biases were pivotal in making a specific classification. This makes TriFN highly effective from a data science standpoint, but difficult to audit or explain in real-world deployment scenarios.

The same holds true for graph-based models such as the on proposed by Monti et al. (2019), which uses Geometric Deep Learning (GDL) and Graph Convolutional Networks (GCNs) to detect fake news based on propagation patterns. These models are particularly well-suited for early-stage detection, demonstrating robust performance (e.g., 92.7% ROC AUC) by leveraging information cascade structures. However, by prioritizing structural features like retweet dynamics and user connectivity, they intentionally discard textual or semantic evidence. As a result, while highly scalable and accurate, these models offer no explanation about the factual content of the new itself, let alone which claims were verified or what sources were considered.

By contrast, the multi-agent system developed in this thesis is explicity designed to be explainable and auditable at every level. The architecture features a combination og

agents that simulate reasoning steps in a modular and transparent way. For instance, fact checking agents explicitly cite external URLs and rank them by credibility. Scoring agents evaluate bias, emotional tone, and source reliability using numerical and linguistic metrics, which are made visible to the user. Most importantly, the Supervisor Agent uses a rule-based system to make final decisions, and his rule set is fully transparent and modifiable. Users not only receive a verdict ("TRUE" or "FAKE") but also understand why that verdict was reached, which evidence was used, and how strongly it supported the claim.

During our manual evaluation (see Section 4.1), this system achieved an overall accuracy of 75%, correctly identifying 60/80. Unlike other models that only provide classification labels, our system outputs structured results (e.g., JSON or tabular explanations) that expose each agent's reasoning. This means users can trace the decision from input to output, replicating or challenging the process with ease. The agent-based structure ensures modularity and promoted long-term maintainability, as new reasoning rules or data sources can be added without retraining the entire model.

In sum, while existing systems, whether traditional ML, hybrid deep learning, or propagation-based-have shown commendable progress in performance metrics, they fall short in offering the level of transparency required for real-world trust and accountability. Our system not only matches or exceeds many of these models in accuracy, but more importantly, it introduces a paradigm shift in how fake news detection is presented to end users: not as classification task, but as a traceable reasoning process rooted in verifiable evidence and explainable logic.

5.2 Modularity and Maintainability

A critical advantage of the proposed multi-agent system lies in its modular architecture, which directly addresses the long-standing issues of rigidity, reusability, and maintainability found in traditional fake news detection pipelines. While high-performing models BERT, CSI, TriFN, or propagation-based GCNs offer impressive predictive accuracy, they are often monolithic and tightly coupled, making them difficult to update, audit, or extend without full retraining or architectural overhauls. In contrast, our system adopts a component-based paradigm, where each agent

operated as an autonomous, logically separated unit responsible for a specific task in the overall verification process.

This modularity was designed deliberately. Each agent, whether it be the Fact Checking Agent, Scoring Agent, Source Quality Checker, or the Supervisor Agent encapsulates its logic, tools, and knowledge sources. For example, the Fact-Checking Agent operates based on a configurable set of search strategies and source filtering heuristics. Should new reputable fact-checking APIs or search methods become available, these can be integrated into the agent without disrupting the rest of the system. Similarly, if scoring strategies (e.g., bias or emotion scoring) need to be refined or recalibrated, the changes can be isolated to the respective scoring modules.

By separating the decision logic (handled by the Supervisor Agent) from the data extraction and evaluation logic (handled by domain-specific agents), the system achieves separation of concerns, a foundational software engineering principle that enhances clarity and reduces risk of unintended side effects during maintenance. News agents can be introduced as plug-ins, for example, a futural Multimodal Agent for image or vide analysis, without requiring reconfiguration of the entire pipeline. Likewise, existing agents can be hot-swapped or upgraded based on domain feedback or external developments in the research field.

From a maintainability standpoint, this modular design is invaluable. Unlike monolithic deep learning models that require retraining on large datasets when fine-tuned or updated, this system supports partial updates, manual rule editing, and even explanation versioning. For instance, if the definition of "credible source" evolves over time (e.g., based on new journalistic standards), the rules governing source quality can be adapted instantly at the agent level. This gives the system an inherent flexibility to evolve alongside the epistemological landscape of misinformation.

Furthermore, since agents output structured, explainable representations (e.g., JSON verdicts, confidence scores, evidence traces), the system is highly compatible with human-in-the-loop workflows and post-hoc auditing tools. Human reviewers can inspect individual agent outputs, trace errors back to their logical source, and make targeted improvements. This contrasts sharply with end-to-end models like CSI or BERT, which offer little internal visibility and thus hinder systematic debugging or rule refinement.

It is also worth noting that the testing framework (see Section 4) supports easy validation of individual agent performance, enabling agent-level benchmarking and unit testing, a level of granularity that is rarely feasible in end-to-end deep or ensemble learning systems.

In summary, the agent based offers clear benefits in terms of system evolution, interoperability, debugging, and domain adaptation. While traditional fake news detection systems focus on performance as their primary metric, our approach embraces long-term sustainability and adaptability, positioning it as a robust platform for real-world deployment and iterative improvement in the face of a constantly changing misinformation ecosystem.

5.3 Quantitative Benchmark Comparison

Model/System	Dataset(s)	Accuracy (%)	F1 Score
Our Multi-Agent System	LIAR + FakeNewsNet	75	0.74
Logistic Regression (TF-IDF)	LIAR	63-65	~0.63
LSTM/Bi-LSTM	LIAR	70-73	~0.72
BERT/RoBERTa	LIAR + FakeNewsNet	76-88	~0.84
CSI (Ruchansky et al., 2017)	Twitter, Weibo	89.2/95.3	0.894/0.954
TriFN (Shu et al., 2019)	FakeNewsNet	86.4/87.8	0.86
RCNN + Sentiment (Wu et al.,2019)	LIAR, Weibo	83.3/84.6	0.431/0.549
Propagation GCN (Monti et al., 2019)	Twitter	ROC-AUC 92.7	N/A
Visual MVNN (Qi et al., 2019)	Weibo (Images)	84.6	0.832

Chapter 6

Challenges and Future Work

Despite the promising results and architectural novelty of the agent-based fake news detection system, several challenges remain that highlight areas for refinement and future exploration.

Testing Scope and Model Generalization

The current evaluation was conducted manually on a relatively small and curated sample of 80 entries from the LIAR and FakeNewsNet datasets. While this allowed for detailed agent-level insight and performance analysis, the system's robustness remains untested at scale. Future work should include:

1. Larger-scale testing with full datasets and cross-domain claims.
2. Benchmarking across languages, platforms (e.g., Reddit, Facebook), and misinformation genres (e.g., satire, conspiracy, propaganda).
3. Stress-testing edge cases (e.g., vague, sarcastic, or multi-topic claims).

Such expansion will help validate the system's generalizability beyond the structured experimental setting used in this thesis.

Model Variants and GPT Alternatives

This work relies heavily on the GPT-4o model for agentic prompting and reasoning. However, it remains an open question how different language models, even within the GPT family, affect overall performance and reasoning style. Future work should include:

1. Systematic testing using GPT-3.5, GPT-4, Claude, and open-source LLMs (e.g., Mistral, LLaMA).
2. Quantitative comparison of model output variance for the same claim under identical agent prompts.
3. Evaluation of cost-efficiency tradeoffs between performance and API/ resource usage.

These tests will inform whether the system is model-agnostic or overly dependent on proprietary LLM behavior.

Formalization of Agent Scoring Criteria

Currently, several scoring and decision mechanisms, particularly within the Scoring Agent and Supervisor Agent, are defined via prompt engineering and soft rule-based heuristics. While this allows flexibility and interpretability, it also introduces subjectivity and potential inconsistency. Future improvements should aim to:

1. Replace heuristic cutoffs with mathematical scoring functions, e.g.,
    a. Weighted averages.
    b. Statistical normalization (e.g., z-scores).
    c. Probabilistic confidence modeling.
2. Learn decision rules via symbolic regression or program synthesis, ensuring decisions are both explainable and consistent.

This formalization would move the system from "interpretable prompting" toward mathematically grounded, auditable reasoning.

Ethical, Legal, and Transparency Challenges

While the system emphasizes explainability, deploying it in real-world settings- especially in media, elections or journalism- raises questions such as:

1. Bias propagation inherited from training data or search engine ranking.
2. Discrepancies in source credibility metrics, especially in politically polarized environments.
3. Compliance with AI transparency and auditability laws (e.g., EU AI Act, DSA).

Future iterations should include fairness audits, explainability documentation, and a transparent disclosure framework.

Chapter 7

Conclusion

This thesis presented a novel, multi-agent system for fake news detection, designed to combine the reasoning capabilities of large language models with modular interpretability, source transparency, and structured decision-making. Unlike traditional black-box classifiers, the system operates through a pipeline of specialized agents, each handling evidence retrieval, bias and credibility score, source evaluation, and final verdict calculation, all while preserving human-readable outputs and audit trails for each decision.

Through a rigorous manual evaluation over 80 diverse entries sourced from both the LIAR and FakeNewsNet datasets, the system demonstrated a high level of alignment with ground-truth labels (75% overall accuracy), and an especially strong performance in identifying true claims (TP:36/49) and fake claims (TF:24/31). Beyond raw accuracy, the system showcased important secondary advantages: confidence scores that correlate with reliability, source reputation metrics that distinguish good evidence, and explanations that support human trust and review.

When compared against traditional machine learning, deep learning, and hybrid models in the literature, the agentic approach provides a compelling middle ground: competitive performance combined with unmatched transparency. The system not only identifies whether a claim is real or fake, but also know how and why its verdict, citing evidence, describing biases, and surfacing domain-level judgments.

Nevertheless, challenges remain. From limited dataset size and model dependency, to the need for mathematically grounded scoring mechanisms and real-time automation, this work represents a first step toward more explainable and trustworthy AI systems for misinformation detection. Future work should focus on scaling up evaluations testing across languages and modalities, and ensuring robustness under adversarial or ambiguous conditions.

In an era where trust in information is under constant pressure, systems like this, that are not only accurate but also understandable, are critical. By combining the flexibility of modern LLMs with structured, modular, and human-aligned design, this thesis

contributes to the foundation of next-generation fact-checking tools that can help rebuild trust in public discourse.

Chapter 8

Appendix

Appendix-A: Agent Prompts

Each subsection below contains the full prompt and operational logic for the agents used in the multi-agent fake news detection framework

A1: Input Agent Prompt

```
You are an input preprocessing agent for a fake news detection system.
Your job is to process and classify all incoming user input into clean
text format. Follow these strict steps:

1. If the input is a URL, return:
{"type": "url",
"cleanText": "This appears to be a URL."}
2. If the input is a short claim, news headline, or text-based
statement, perform the following:
a. Detect the language.
b. If the language is not English, translate it into English.
c. Clean the translated or original English text by:
- Removing punctuation
- Converting to lowercase
- Stripping emojis and special characters
- Trimming whitespace
Then return:
{"type": "text",
"cleanText": "<cleaned_english_text>"}
3. If the input is a file or document, return:
{"type": "file",
"cleanText": "This is a file input."}
⚠ Only return a valid JSON object with keys: type and cleanText. Do
NOT explain, justify, or include any other text in the response.
```

A2: Fake News Classifier Agent (Google Custom Search)

```
Fake News Agent Prompts
📄 Prompt 1: Fake News Classifier Agent (Google Custom Search)
You are a real-time fake news classifier. You will receive a user-
submitted claim, and your job is to fact-check it based ONLY on
evidence retrieved from the Google Custom Search Tool.
===
☝️ Input Claim:
{{agentAgentflow_0 }}
===
Instructions:
Perform a search using the Google Custom Search Tool based specifically
on the above claim.
Analyze only the results returned from this search.
Return this structured JSON object:
{"verdict": "REAL" | "FAKE" | "SUSPICIOUS",
"confidence": <number from 0 to 100>,
"reasoning": "<brief explanation based strictly on this search>",
"supporting_sources": ["<URL1>", "<URL2>", "<URL3>"]}
Guidelines:
Use only sources from the Google Custom Search results.
Mention domain credibility and quality of evidence.
Do not speculate. Only use search-based findings.
Do not mark as REAL unless multiple sources explicitly confirm the full
claim.
Return ONLY the JSON. No extra explanation or text.
❌ Never reference unrelated topics or reuse results from other claims
☑️ Always align your reasoning with the specific search results from
this exact claim
⚠️ Important:
Do not consider a claim verified or supported just because it covers a
similar topic.
Only mark a claim as REAL or LIKELY_TRUE if:
- The supporting articles confirm the exact meaning, implication, and
tone of the claim.
- The wording and intent match closely.
- The article uses similar language or explicitly states what the claim
asserts.
Reject vague matches. For example:
- "She can leave whenever she wants" ≠ "They said 'get out, we don't
care about you'".
- "A fire occurred" ≠ "They burned documents to hide something from
investigators."
If no precise match is found, return SUSPICIOUS or NO_CLEAR_EVIDENCE,
even if general coverage exists.
Be cautious of:
- Emotional exaggeration
- Implications not supported in source articles
```

```
- Clickbait phrases and speculative claims
Return ONLY the JSON object. No extra explanations or text.
```

A3: Second Search Agent (Serper API)

```
You are a secondary fact-checking agent assisting in fake news
detection. Your task is to independently verify a user-submitted claim
using only fresh, credible online sources via the Serper API tool.
===
🧠 Input Claim:
{{agentAgentflow_0}}
===
Instructions:
Perform a real-time web search using Serper API based on the claim
only.
Analyze the search results and determine if credible sources confirm,
deny, or ignore the claim.
Return this structured JSON object:
{
"search_summary": "<brief, neutral summary of the search results>",
"supporting_sources": ["<URL1>", "<URL2>", "<URL3>"],
"verdict_hint": "LIKELY_TRUE" | "LIKELY_FAKE" | "NO_CLEAR_EVIDENCE"
}
Guidelines:
Only use Serper search results. No speculation.
Only include real, relevant, and trustworthy URLs.
Do not mark as LIKELY_TRUE unless multiple trustworthy sources
explicitly support the claim.
Respond ONLY with the JSON object.
⚠️ Important:
Do not consider a claim verified or supported just because it covers a
similar topic.
Only mark a claim as REAL or LIKELY_TRUE if:
- The supporting articles confirm the exact meaning, implication, and
tone of the claim.
- The wording and intent match closely.
- The article uses similar language or explicitly states what the claim
asserts.
Reject vague matches. For example:
- "She can leave whenever she wants" ≠ "They said ‘get out, we don’t
care about you’".
- “A fire occurred” ≠ “They burned documents to hide something from
investigators.”
If no precise match is found, return SUSPICIOUS or NO_CLEAR_EVIDENCE,
even if general coverage exists.
Be cautious of:
- Emotional exaggeration
- Implications not supported in source articles
```

```
- Clickbait phrases and speculative claims
```

A4: Scoring Agent

```
You are a scoring agent analyzing a text's bias, emotional tone,
credibility, and time relevance.
===
☝️ Input Claim:
{{agentAgentflow_0}}
===
Return the following JSON object:
{
"bias_score": <0-100>,
"bias_label": "<LOW | MEDIUM | HIGH>",
"emotion_score": <0-100>,
"emotion_label": "<NEUTRAL | MODERATE | EMOTIONAL>",
"credibility_score": <0-100>,
"credibility_label": "<LOW | MEDIUM | HIGH>",
"time_relevance_score": <0-100>,
"time_label": "<OUTDATED | POSSIBLY RELEVANT | CURRENT>",
"reasoning": "<Brief explanation for bias, emotion, and credibility>",
"time_reasoning": "<Brief explanation for time relevance>"
}
Labeling thresholds:
Bias: LOW (0-30), MEDIUM (31-70), HIGH (71-100)
Emotion: NEUTRAL (0-30), MODERATE (31-70), EMOTIONAL (71-100)
Credibility: LOW (0-30), MEDIUM (31-70), HIGH (71-100)
Time: OUTDATED (0-30), POSSIBLY RELEVANT (31-70), CURRENT (71-100)
Respond with only the JSON.
```

A5: Source Quality Checker

```
You are a Source Quality Checker. You will receive a list of URLs used
to support a claim.
You will receive input from:
{{agentAgentflow_1}}: URLs from the Google Custom Search Agent
{{agentAgentflow_4}}: URLs from the Serper API Agent
Your task is to evaluate the trustworthiness of these sources based on
domain type, publisher reputation, and credibility.
Your task is to evaluate the trustworthiness of these sources based on
the domain, reputation, and credibility.
Use this rubric:
.gov, .edu, reputable institutional/scientific domains → 85–100 (High)
Mainstream media (BBC, Reuters, NYT, Al Jazeera, etc.) → 70–84 (Medium-
High)
Independent journalism or reputable .orgs → 50–69 (Medium)
Blogs, forums, opinion sites, or YouTube → 30–49 (Low)
```

```
Clickbait, conspiracy, anonymous, or unverifiable sites → 0–29 (Very
Low)
Return this JSON:
- .gov, .edu, reputable scientific or institutional sites → High
- Mainstream news sources (BBC, Reuters, NYT, Al Jazeera, etc.) →
Medium-High
- Independent journalism or .orgs (depending on reputation) → Medium
- Blogs, forums, opinion sites, or YouTube → Low
- Clickbait or conspiracy sites → Very Low
Return only a JSON object in this format:
{
"average_source_score": <0-100>,
"source_quality": "<Very Low | Low | Medium | High | Very High>",
"notes": "<brief explanation referencing the domains or outlets
involved>"
}
Use only the provided URLs.
```

A6: Supervisor Agent

```
You are the Supervisor Agent responsible for compiling and delivering
the final analysis of a user-submitted claim.
Inputs you receive:
Cleaned claim text -> {{agentAgentflow_0}}
Verdict, confidence, and reasoning from Fake News Classifier ->
{{agentAgentflow_1}}
Scoring metrics for bias, emotion, credibility, and time ->
{{agentAgentflow_3}}
Source quality score from Source Quality Checker ->
{{agentAgentflow_5}}
Timestamp of analysis
Secondary Search Verdict Hint + Summary + Sources ->
{{agentAgentflow_4}}
Respond with ONLY this structure:
====================
🧠 Fake News Verdict:
Verdict: {verdict}
Confidence: {confidence}%
Reasoning: {verdict_reasoning}
📊 Scoring Summary:
Bias Score: {bias_score}/100 — {bias_label}
Emotion Score: {emotion_score}/100 — {emotion_label}
Credibility Score: {credibility_score}/100 — {credibility_label}
Time Relevance Score: {time_relevance_score}/100 — {time_label}
📋 Scoring Reasoning: {score_reasoning}
🕒 Timestamp of Evaluation: {timestamp}
```

```
🌐 Secondary Web Search Summary: Secondary Verdict Hint:
{secondary_verdict_hint} Search Summary: {search_summary} Sources:
{sources_list}
🧐 Source Quality Assessment: Average Source Score:
{average_source_score}/100 – {source_quality} Notes: {notes}
📢 Final Call:
FINAL DECISION: {REAL | FAKE}
CONFIDENCE LEVEL: {High | Medium | Low}
EXPLANATION:
"{brief explanation based on verdict agreement, source credibility, and
semantic alignment between the claim and found evidence}"
====================
🤔 Decision Logic:
1. SEMANTIC CONSISTENCY CHECK:
If:
- The supporting sources do not directly confirm the exact wording or
meaning of the claim,
- Or if there is only general topic overlap without exact alignment,
- Or if the tone or intent in the claim differs significantly from the
source evidence,
→ Treat the evidence as non-confirmatory and default to:
→ FINAL DECISION: FAKE or SUSPICIOUS
—
2. Use the following logic only if the semantic match is strong:
- If:
- verdict = FAKE or SUSPICIOUS
- AND secondary_verdict_hint = LIKELY_FAKE
- AND credibility_score < 70
→ FINAL DECISION: FAKE
- If:
- verdict = REAL
- AND secondary_verdict_hint = LIKELY_TRUE
- AND credibility_score ≥ 70
→ FINAL DECISION: REAL
- Else:
- If credibility_score < 50 OR source_quality in ["Low", "Very Low"]
→ FINAL DECISION: FAKE
- Else → FINAL DECISION: REAL
—
3. CONFIDENCE LEVEL:
- High → if verdicts match AND credibility ≥ 80 AND semantic match is
clear
- Medium → if some alignment exists but confidence or semantic match is
partial
- Low → if sources are weak OR verdicts contradict each other OR low
semantic match
====================
```

Chapter 9

References

- Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake News Detection on Social Media using Geometric Deep Learning. arXiv preprint arXiv:1902.06673. https://arxiv.org/abs/1902.06673
- Perez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. Proceedings of the 27th International Conference on Computational Linguistics, 3391–3401.
- Qi, P., Cao, J., Yang, X., Guo, J., Zhang, Y., & Li, J. (2019). Multi-domain visual neural network for fake news detection. Knowledge-Based Systems, 188, 105–128.
- Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2931–2937.
- Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM), 797–806. https://doi.org/10.1145/3132847.3132877
- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.
- Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2019). Beyond News Contents: The Role of Social Context for Fake News Detection. Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), 312–320. https://doi.org/10.1145/3289600.3290994
- Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. Proceedings of the Second Workshop on Data Science for Social Good (SoGood).

- Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 422–426. https://doi.org/10.18653/v1/P17-2067
- Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). False information detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 21(2), 80–90.
- Wu, K., Yang, S., & Zhu, K. Q. (2015). False rumors detection on Sina Weibo by propagation structures. IEEE International Conference on Data Engineering (ICDE), 651–662. https://doi.org/10.1109/ICDE.2015.7113322
- Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., & Baddour, K. (2020). Coronavirus Goes Viral: Quantifying the COVID 19 Misinformation Epidemic on Twitter. Cureus,/12(3), e7255. https://doi.org/10.7759/cureus.7255
- *Baptista, J. P., & Gradim, A. (2022).* Who Believes in Fake News? Identification of Political (A)Symmetries. ***Social Sciences****, 11(10), 460. https://www.mdpi.com/2076-0760/11/10/460* *MDPI+1*