AUTOMATED MULTI-STEP FACT-CHECKING ON SOCIAL MEDIA THROUGH CLAIM DETECTION, SIMPLIFICATION, AND VERIFICATION

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

MEGHA SUNDRIYAL

(PHD20009)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI - 110020

AUGUST 2025

THESIS CERTIFICATE

This is to certify that the thesis titled **Automated Multi-Step Fact-Checking on Social Media through Claim Detection, Simplification, and Verification**, submitted by **Megha Sundriyal** (PhD20009), to the Indraprastha Institute of Information Technology Delhi, for the partial award of the degree of **Doctor of Philosophy**, is a bonafide record of the research work done by her under my supervision. In my opinion, the thesis has reached the standard, fulfilling the requirements of the regulations relating to the degree. The contents of this thesis, in full or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

Dr. Tanmoy Chakraborty
Thesis Supervisor
Associate Professor
Dept. of Electrical Engineering
IIT Delhi, 110016

Dr. Md Shad Akhtar
Thesis Supervisor
Assistant Professor
Dept. of Computer Science and Engineering
IIIT Delhi, 110020

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to those who have supported me throughout my thesis journey. First and foremost, I am deeply thankful to my PhD advisors, Dr. Tanmoy Chakraborty and Dr. Md Shad Akhtar, whose meticulous guidance and patience have been instrumental in my professional and personal growth. Each interaction with them inspired intriguing hypotheses that kept me focused on my academic goals. Their commitment to fostering an inclusive environment for open discussions has significantly enriched my research experience. I would also like to acknowledge the financial support provided by the University Grants Commission (UGC), India, through the UGC-NET Junior Research Fellowship (JRF) and Senior Research Fellowship (SRF), which enabled me to carry out my research during the course of my PhD.

I sincerely thank my esteemed committee members, Prof. Mukesh Mohania and Dr. Raghava Mutharaju, for their valuable contributions to my annual progress report. Their insightful feedback enhanced the quality of my work and gave me a clearer perspective on my research trajectory. The constructive criticism and encouragement I received during our discussions were pivotal in refining my ideas and methodologies, empowering me to address challenges more effectively.

My sincere gratitude goes to Prof. Preslav Nakov, who was an exceptional mentor during my research visit to Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), UAE. His exemplary work ethic and emphasis on critical thinking motivated me to strive for excellence. Prof. Nakov's extensive expertise in natural language processing and misinformation detection provided me with invaluable insights that significantly enriched my research experience. I am extremely thankful to all of the members of Prof. Nakov's lab for their warmth and friendliness, which made me feel truly welcome as a member of their team. Their warmth and enthusiasm created an inclusive environment that considerably enhanced my experience. The lively discussions and shared views provided vital advice and encouragement, making my time at MBZUAI extremely memorable.

Throughout my PhD, I had tremendous support and cooperation from the Laboratory for Computational Social Systems (LCS2) research group. Starting my PhD during COVID-19 required adapting to an online setting, but my fellow labmates' support and joint effort made all the difference. Their willingness to share their experiences and insights created a professional yet friendly atmosphere that

greatly facilitated my learning and growth. I am especially grateful to my fellows in the lab – Shivani Kumar, Yash Atri, Aseem Srivastava, Sarah Masud, Abdullah Mazhar, Aditya Bhardwaj, and Zeba Afroz – for supporting me and presenting in a professional and friendly environment in the lab. I'd like to thank all of my co-authors, who were always up for a discussion and had interesting perspectives on the problems at hand. Furthermore, I want to convey my heartfelt gratitude to peers from various lab groups in IIIT-Delhi, who provided additional interdisciplinary perspectives on my research journey.

This journey would not have been possible without the constant support of my family and friends, whose love and encouragement served as my guiding light. My parents, Krishan Kumar Sundriyal and Tulsi Sundriyal, have always believed in my abilities and instilled in me the principles of perseverance and hard work. I couldn't have completed this journey without their love and support, which has been a safety net for my aspirations. I am also deeply thankful to my younger brother, Mayank Sundriyal, who has always been my biggest cheerleader. My family's unconditional love has ensured that I can confidently pursue my dreams, and I adore them more than anything. I would like to extend my sincere thanks to Himanshu Singh, who has been a continuos source of emotional support and encouragement throughout this journey. His faith in me, ability to lift my spirits during the toughest moments, and constant cheer have been invaluable. Last but not the least, I am immensely grateful to my friends and colleagues – Gunjan Singh, Mansi Goel, Gaurav Rai, Avi Gupta, Anam Fatima, Daminee Saini, Aman Kumar, Tanuj, Krishna, and Uditi Kansal – for their companionship and understanding. Their presence helped me sustain a healthy work-life balance throughout this demanding journey. Each one has contributed uniquely to my experience, and I am truly thankful for their warmth, friendship, and encouragement.

Reaching this milestone would not have been possible without the contributions and encouragement of all those who have been part of this journey, directly or indirectly. This acknowledgment is a humble token of my deep and heartfelt gratitude to everyone who has supported and stood by me along the way.

Megha Sundriyal

Delhi, August 2025

ABSTRACT

The spread of misinformation on social media platforms endangers public discourse and societal trust. Addressing this issue requires robust mechanisms for identifying and verifying the accuracy of information circulated online. This thesis focuses on improving fact-checkers' capabilities by creating a comprehensive framework for efficient misinformation management via identifying, simplifying, and verifying claims on social media. This research endeavour is multifaceted and involves (a) comprehending the nature and context of claims made on social media, (b) creating algorithms to detect these claims automatically, (c) simplifying complex or noisy claims to enable more accessible analysis, (d) determining which claims are worthy of verification, and (e) using sophisticated computational techniques to verify the accuracy of these claims. We employ natural language processing and machine learning algorithms to parse and interpret textual content from social media platforms. Using cutting-edge models, we automatically detect and distil claims from massive amounts of data, addressing the scale challenge inherent in social media ecosystems. Furthermore, we use novel claim simplification processes to convert verbose or ambiguous statements into explicit, concise claims that can be verified. In the final stage, we verify identified claims, thus providing a reliable method to affirm or refute claims. The efficacy of our methods is demonstrated through extensive experiments on real-world data, showing significant improvements in the speed and accuracy of fact-checking operations. This thesis contributes to fact-checking by offering a scalable, efficient solution for combating misinformation on social media and equipping fact-checkers with tools critical in the fight against information distortion. The methodologies presented herein lay the groundwork for future research and practical applications in digital information verification.

PUBLICATIONS

Conferences

C1. Shreya Gupta, Parantak Singh, **Megha Sundriyal**, Md Shad Akhtar, Tanmoy Chakraborty, "LESA: Linguistic Encapsulation and Semantic Amalgamation Based Generalised Claim Detection from Online Content," *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021.*

C2. **Megha Sundriyal**, Parantak Singh, Md Shad Akhtar, Shubhashis Sengupta, Tanmoy Chakraborty, "DESYR: Definition and Syntactic Representation Based Claim Detection on the Web," *Proceedings of the 30th ACM International Conference on Information and Knowledge Management, 2021.*

C3. **Megha Sundriyal**, Atharva Kulkarni, Vaibhav Pulastya, Md. Shad Akhtar, Tanmoy Chakraborty, "Empowering the Fact-checkers! Automatic Identification of Claim Spans on Twitter," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022.*

C4. **Megha Sundriyal**, Ganeshan Malhotra, Md Shad Akhtar, Shubhashis Sengupta, Andrew Fano, Tanmoy Chakraborty, "Document Retrieval and Claim Verification to Mitigate COVID-19 Misinformation," *Proceedings of Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations, 2022*.

C5. **Megha Sundriyal**, Tanmoy Chakraborty, Preslav Nakov, "From Chaos to Clarity: Claim Normalization to Empower Fact-Checking," *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023*.

C6. Shubham Mittal, **Megha Sundriyal**, and Preslav Nakov, "Lost in Translation, Found in Spans: Identifying Claims in Multilingual Social Media," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023*.

C7. Arghodeep Nandi, **Megha Sundriyal**, Euna Mehnaz Khan, Jikai Sun, Emily Vraga, Jaideep Srivastava, and Tanmoy Chakraborty, "The Psychology of Falsehood: A Human-Centric Survey of Misinformation Detection," *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025*.

C8. Wenchao Dong, **Megha Sundriyal**, Seongchan Park, Jaehong Kim, Meeyoung Cha, Tanmoy Chakraborty, and Wonjae Lee, "Parallel Communities Across the Surface Web and the Dark Web," *Findings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025*.

Peer-reviewed Journals

J1. Mohit Bhardwaj, **Megha Sundriyal**, Manjot Bedi, Md. Shad Akhtar, and Tanmoy Chakraborty, "HostileNet: Multilabel Hostile Post Detection in Hindi," *IEEE Transactions on Computational Social Systems, 2023*.

J2. **Megha Sundriyal**, Md. Shad Akhtar, and Tanmoy Chakraborty, "Leveraging Social Discourse to Measure Check-Worthiness of Claims for Fact-Checking," *IEEE Transactions on Artificial Intelligence, 2025*.

Overview Papers

O1. **Megha Sundriyal**, Md. Shad Akhtar, and Tanmoy Chakraborty, "Overview of the CLAIMSCAN-2023: Uncovering Truth in Social Media through Claim Detection and Identification of Claim Spans," *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, 2023*.

O2. **Megha Sundriyal**, Tanmoy Chakraborty, and Preslav Nakov "Overview of the CLEF-2025 Check-That! Lab Task 2 on Claim Normalization," *Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025*.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Since the dawn of human civilization, people have relied on the information they gather to thrive on this planet. This distinctive trait to share knowledge, whether through simple cave paintings in the Stone Age or complex encrypted messages in the digital age, sets us apart from all other living beings (3; 4). Over time, information sharing has evolved into a vital tool for survival, guiding our decisions based on the data we collect (5). Thus, the accuracy and reliability of information significantly impact individual decision-making and collectively shape society. In today's era, where information flows more freely and rapidly than ever, falsification or fabrication of facts has become a powerful tool to sway public opinion and shape societal narratives. This manipulation of information, often under the guise of truth, constitutes a form of *misinformation* that can distort perceptions, influence behaviour, and undermine trust in critical systems (6; 7). The spread of misinformation in society can lead to grave consequences, including public confusion, loss of trust in institutions, and changes in society's behavioural and psychological patterns. Addressing this issue requires a multi-dimensional approach that includes critical thinking, media literacy, and fact-checking initiatives.

While misinformation has become a widespread problem in the digital age, its rapid dissemination did not occur overnight. In fact, it has long been a persistent issue throughout history, dating back to ancient storytelling traditions where truth was often mingled with exaggeration to captivate audiences (8; 9). This blend of fact and fiction laid the groundwork for the many forms of misinformation we see today. To comprehend all aspects and complexity of the problem, it's essential to look back at how misinformation has evolved over time.

1.1 Evolution of Misinformation

Misinformation has evolved alongside human communication, tracing back to ancient societies, where it often spread through oral traditions and word of mouth (10). Figure 1.1 depicts numerous significant examples from various periods, demonstrating how misinformation has evolved over time. In ancient times, stories were retold and sometimes unintentionally altered, leading to myths and misconceptions. A notable example is the campaign by Octavian against Antony in 63 BC (11). Octavian waged a campaign against Antony that was designed to smear his reputation. He painted Antony as a womanizer and puppet of Cleopatra. This campaign was achieved through *'short, sharp slogans written upon coins in the style of archaic Tweets'* (12). Eventually, these tactics contributed to Octavian's rise as the first Roman Emperor, illustrating misinformation's potential to reshape political landscapes. Ancient rulers in Egypt, Greece, and Rome employed similar primitive forms of blending truth with fabrications to sway public opinion and maintain power.

As civilizations progressed, so did the sophistication and reach of information evolved with new communication mediums. The invention of the printing press in 1440 by Johannes Gutenberg enabled information – and misinformation – to spread more rapidly than ever before (13). During the *Reformation* of the 16th century, both Catholics and Protestants used printed materials containing misinformation to disparage each other's beliefs (14). To influence new converts or hold onto their existing followers, propagandists from both groups tried to disseminate texts pertaining to church theology. These printed materials occasionally served as guidelines for lay people to resort to when determining how to conduct themselves in the church and society. Another striking example of misinformation occurred in August 1835 when *The New York Sun* published an article series entitled *'Great Astronomical Discoveries'* about

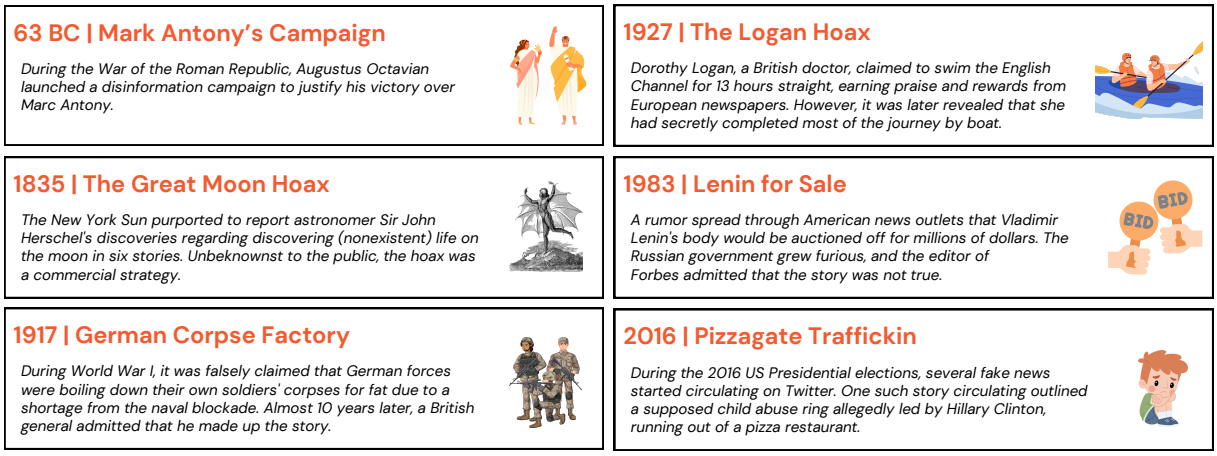


Figure 1.1: Historical context of information disorders with key examples of misinformation across different periods.

the discovery of life on the moon, with complete illustrations of humanoid bat-creatures and bearded blue unicorns (15). The stories were presented as factual reports from a prominent astronomer, Andrew Grant, supposedly working with the famous British astronomer Sir John Herschel. The newspaper's circulation went from 8,000 to 19,000 copies, making it the world's best-selling at the time. However, in early September 1835, shortly after the initial articles were published, the hoax was exposed when other newspapers began investigating the claims, and it called it as *'The Great Moon Hoax'* (16).

Later, in the 18th and 19th centuries, print media frequently employed sensationalized news stories to boost sales. This practice was later termed as *'yellow journalism'* (17). These sensationalized news articles contributed largely to the Spanish-American War in 1898, establishing misinformation as a significant instrument in modern journalism (18). The 20th century witnessed more refined misinformation tactics, particularly during global conflicts, where the power of fabricated narratives and strategic deception became entirely apparent. During World Wars, propaganda proved important for boosting national morale and confusing or destabilizing rivals. For instance, Nazi Germany orchestrated a highly coordinated campaign of anti-Semitic misinformation, using fabricated narratives to justify atrocities and manipulate public sentiment (19). In a similar vein, the Allied forces used misleading military techniques, including false intelligence and misdirection, to confuse the enemy and achieve strategic advantages on the battle-ground (20; 21; 22). The Cold War ushered in a new era of systematic misinformation, with intelligence organizations around the world using misinformation as a psychological weapon (23). One of the most prominent examples of this was the Soviet KGB's use of *'active measures'* (24). It was a series of covert operations aimed at transmitting false or misleading information worldwide to sow dissension, weaken governments, and generate distrust among friends. These strategies were not confined to political arenas; they also influenced public perceptions and foreign relations, leaving a long-term impact on global affairs.

The advent of radio and television further transformed information dissemination, resulting in an interconnected world where content could be instantly broadcast to a global audience. While connectivity provided various positives, it also fuelled the spread of misinformation. In today's digital age, where trust in news sources has become polarised, many consumers feel entitled to choose or create their own *facts* and *beliefs*. One such prominent example is Orson Welles' 1938 radio adaptation of H.G. Wells's science fiction novel called *'War of the Worlds'*. Delivered in breaking news format, the broadcast was so convincingly produced that many listeners mistakenly accepted the fictional narrative as fact. This resulted in widespread panic across the country as some fled their homes believing a Martian invasion was underway (25).

Over the years, the methods of misinformation propagation have developed vastly, from engraved coins to more sophisticated distortion in print and broadcast media. These early incidents paved the way for the more widespread and rapid transmission of misinformation we witness today.

1.2 Misinformation on Social Media

Social media has revolutionized how we access and consume information, making it a primary source for millions of people around the world (26; 27). Platforms like Facebook, Twitter, and Instagram have democratized the flow of information, allowing people to share perspectives widely. Though effective in promoting free speech and fostering worldwide connectivity, they also present perfect breeding grounds for misinformation (28; 29). Unfortunately, the speed at which information spreads on these platforms enables false narratives to spread like wildfire, reaching millions in hours.

The potential of digital media to shape beliefs, as evident in incidents like the *War of the Worlds* broadcast, now operates on a much larger scale. A recent example of this occurred when a tourist submersible lost contact during a dive to view the Titanic wreck, capturing the attention of millions worldwide (30). On June 25th, a TikTok video surfaced, claiming to feature the passengers' final screams. In just ten days, the video amassed 4.9 million views, falsely portraying the screams as the last moments of the five victims. In reality, the audio was taken from the video game *'Five Nights at Freddy's'*. Despite the misinformation, the TikTok clip went viral, quickly surpassing the spread of verified facts. The rise of social media as a news distributor has significant implications for journalism, blurring the lines between factual reporting and misinformation. False information has often hampered the major elections. As the US election approached in 2016, international media reports revealed a profitable troll farm run by teenagers in the small town of Veles in the Former Yugoslav Republic of Macedonia (31). It was discovered that over 100 pro-Trump websites pushing fabricated news were registered in Veles, with one operator earning $16,000 in the final three months of the campaign. The content included viral fake stories about the Pope endorsing candidate Donald Trump. In the final weeks of the election campaign, President Obama talked at length about the *'digital gold rush'* experienced by Veles' fake news' farm. In an October 2016 interview, Colombian hacker Andrés Sepúlveda explained how he manipulated social media and interfered in elections across Latin America from 2006-2014 (32). Sepulveda claimed to have led a team of hackers that *'stole campaign strategies, manipulated social media to create false waves of enthusiasm and derision, and installed spyware in opposition offices'* during the 2012 Mexican presidential election.

Throughout history, times of public health emergencies have surged the spread of misinformation and bogus health remedies (33; 34). In moments of fear and uncertainty, people seem to be more prone to misinformation that promises quick fixes, even when these solutions are not backed by science. The COVID-19 pandemic was the first of its kind health emergency during the modern age of technology (35). When the entire world went into lockdown, the digital media was active. The sheer volume of information generated during the COVID-19 pandemic was stunning. A jumble of facts, fiction, half-truths, rumours, and conspiracy theories parade daily on news outlets, cable television, talk radio, and social media. The challenge of divining wheat from chaff became all the more critical in matters of life or death. Concern over the flood of misleading information about COVID-19 led the World Health Organization (WHO) to declare the outbreak an *'infodemic'* (36). Today, governments, tech companies, and several fact-checking organizations are trying to combat misinformation by strengthening regulations, adjusting algorithms, and educating the public. However, as technology advances, especially with AI tools that can produce realistic deepfakes, the battle against misinformation remains challenging. The efforts to build critical thinking and public awareness should be promoted to fight this evolving threat.

1.3 Rise of Fact-Checking

While the emergence of social media has brought unprecedented challenges in verifying information, fact-checking itself is not new. In fact, it has been around for almost a century. In 1923, Briton Hadden and Henry Luce launched *Time Magazine*, completely changing the use and goal of facts (37). Their ground-breaking periodical publication initially called *Facts*, aimed to condense intricate tales into brief reports and analyses. They pioneered the concept of a specialized research department – present-day fact-checking – to guarantee that every written word was verifiably accurate. As a formalized process, fact-checking began in the early 20th century, particularly in the 1920s (38). It was primarily *ante-hoc* that aimed to identify errors so that the text could be corrected before dissemination rather than after publication. This process aimed to identify and correct errors or inconsistencies in the content before it went to print or was broadcast. If inaccuracies were found, the information could either be corrected or, in some cases, the story could be rejected entirely. This method was essential for maintaining media outlets' credibility, ensuring that misinformation did not reach the public domain (39). At this time, fact-checking was primarily the responsibility of editors and journalists working directly in the editorial process, particularly in traditional print media. These individuals would manually review articles, verify facts, and cross-reference sources.

Amid the growth of technology, *independent* fact-checking emerged as a pivotal safeguard against misinformation. In the early 2000s, the first external fact-checking organization arose (40). These independent fact-checking organizations focused on *post-hoc* fact-checking, where they verified content after it had already been published. As social media platforms grew in prominence, the pace and reach of misinformation also escalated. With the click of a button, false content could go viral, and separating fact from fiction became increasingly difficult. While the fundamentals of independent fact-checking remained the same, the landscape had drastically changed. Several prominent fact-checking organizations, such as FactCheck.org[1], PolitiFact[2] in the U.S., and Full Fact[3] in the U.K., emerged to verify social media claims in real-time. These organizations utilized various methods, including research, expert consultation, and data analysis, to assess the truthfulness of statements and correct falsehoods. A crucial turning point in the ascent of fact-checking occurred in 2009 when the esteemed Pulitzer Prize was given to PolitiFact in the national reporting category (41). As the number of fact-checking organizations increased, bringing them together under a unified framework became crucial. Thus, in 2015, the Poynter Institute launched the International Fact-Checking Network (IFCN)[4] to set a code of ethics for fact-checking organizations. The IFCN's role was to evaluate and assess fact-checking organisations to ensure adherence to ethical guidelines and bestow certifications upon publishers who successfully passed the assessment. These certifications were valid for a duration of one year, with fact-checkers being subject to annual re-examination to maintain their accreditation. As of November 2024, IFCN listed 170 organizations as members (42). Around the same time, the Duke Reporters' Lab launched ClaimReview[5], a markup system for fact-checking articles. They collaborated with the IFCN; subsequently, Google News Initiative bid them a grant of $200,000 (43). The tool enabled search engines and readers to find non-partisan fact-checks and organized them into structured data for automated fact-checking.

Challenges of Manual Fact-Checking. Despite the enormous growth in the number of fact-checking organizations, they face significant challenges that hinder their ability to address misinformation. The multifaceted challenges faced by them range from the difficulties in fact-checking, the pressures of rapid

[1] https://www.factcheck.org/
[2] https://www.politifact.com/
[3] https://fullfact.org/
[4] https://www.poynter.org/ifcn/
[5] https://schema.org/ClaimReview

news cycles, and the influence of social media platforms and algorithms on information distribution (44; 45; 46). One of the most pressing issues is the *Sheer Volume of Information* circulating on the internet and social media platforms. Professional fact-checkers are tasked with sifting through vast amounts of data to verify facts, which is both time-consuming and resource-intensive. The sheer volume of content generated on social media platforms far exceeds the capacity of human fact-checkers to review and verify. This results in a backlog of claims that remain unchecked or inadequately addressed. Additionally, the *Rapid 24/7 cycle* at which misinformation spreads intensifies these difficulties. The manual nature of traditional fact-checking methods also means a lag between the creation of misinformation and its correction. Furthermore, fact-checking efforts are often hampered by resource constraints, as many organizations operate with limited staff and funding. These challenges highlight the need for more scalable and efficient solutions to combat misinformation effectively. The sheer volume of daily content on social media platforms and their algorithmic structures has further facilitated the rapid spread of misleading and false information. These platforms are designed to maximize user engagement, often prioritizing sensational or controversial content that garners more attention. This emphasis on engagement often leads to the viral spread of misinformation as users share, retweet, and like misleading posts. The speed and scale at which information travels on social media are unparalleled, creating an environment where false claims can gain traction and reach a vast audience before they can be effectively addressed. The accessibility and ease with which misinformation can be created and disseminated contribute to a digital ecosystem where verifying the authenticity of information becomes increasingly challenging.

Need for Automated Fact-Checking. The furore over misinformation has exacerbated the urgent need for more efficient fact-checking. Despite the presence of several fact-checking organizations, scrutinizing even a single statement requires significant time and resources. A professional fact-checker might take several hours or days on any given claim (47; 48). Thus, relying solely on manual fact-checking is insufficient, highlighting the need for automated fact-checking. Many researchers have tackled automated fact-checking with various innovative approaches (49; 50; 51; 52). These systems primarily focus on retrieving evidence to verify the claims' truthfulness. However, this task often relies on human intelligence and may not always be feasible through automated means alone. Fact-checking is a nuanced and subjective task requiring human judgment to make informed decisions. As a result, it is critical to bridge the gap between automated and manual fact-checking processes. In this thesis, we aim to re-conceptualize fact-checking as a comprehensive framework of interconnected tasks, such as claim detection, claim span identification, claim check-worthiness estimation, etc., rather than viewing it as a straightforward process of retrieving evidence fragments. By automating these fact-checking subtasks, we aim to assist fact-checkers in identifying and evaluating claims more efficiently and accurately, ultimately improving the overall effectiveness of combating misinformation in digital spaces.

Ethical Considerations in Automated Fact-Checking. The deployment of automated fact-checking systems raises several important ethical considerations that must be acknowledged alongside technical challenges. These systems operate in complex socio-political environments, and their outputs can directly influence public opinion, trust, and freedom of expression. One primary concern is the presence of bias in training data and model outputs. Fact-checking datasets often reflect existing social, political, or geographic imbalances, which can lead to systems that disproportionately flag or ignore certain types of claims. To mitigate this, we ensure that the datasets used and constructed in this thesis are diverse and balanced across domains and topics. Another challenge is the risk of over-reliance on automation. Fact-checking is inherently nuanced, and fully automated systems can struggle with context, ambiguity, or evolving information. While automation can support scalability, it should not replace expert human judgment. In this thesis, we advocate for a modular, assistive approach: our system components, claim detection, simplification, check-worthiness prediction, and verification, are designed to integrate into

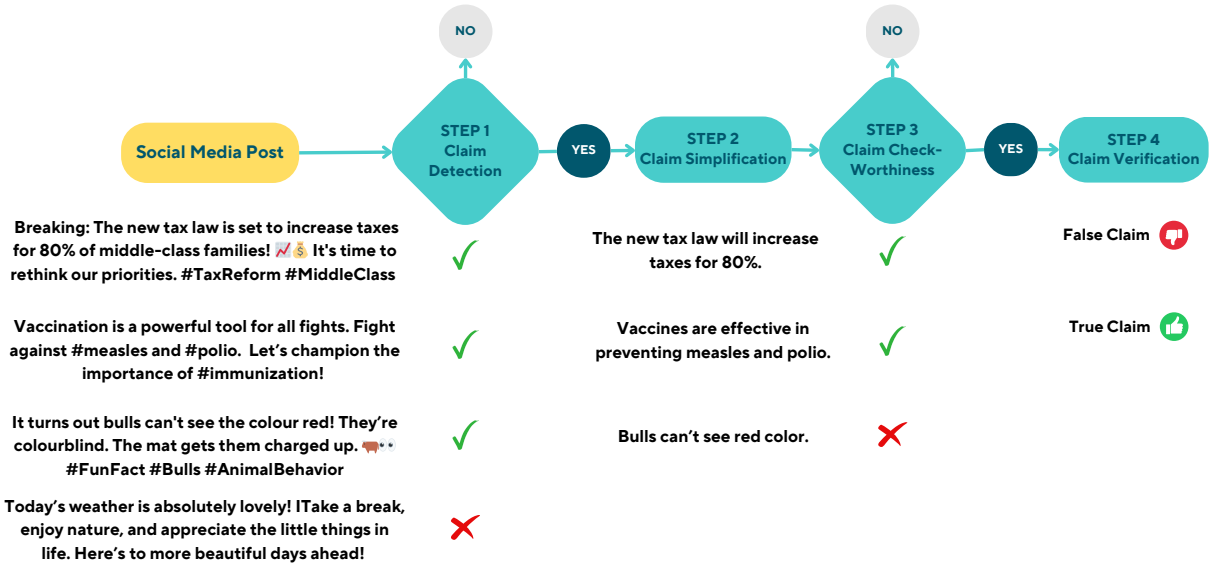


Figure 1.2: Overview of the fact-checking pipeline explored in this thesis: claim detection, claim simplification, claim check-worthiness, and claim verification.

human-in-the-loop pipelines. This allows fact-checkers to selectively engage with automation while retaining final control over decisions, thereby preserving interpretability and accountability. Additionally, we acknowledge the delicate balance between misinformation detection and freedom of expression. Fact-checking systems can unintentionally suppress legitimate dissent or opinion, particularly in politically sensitive contexts. To mitigate this issue, our definition of a 'claim' explicitly excludes opinionated or subjective expressions. We focus on factually verifiable statements, thereby helping to prevent the system's misuse for content moderation or censorship.

1.4 Thesis Scope

At the heart of the misinformation avalanche lies the concept of a *claim* – a statement that asserts something to be true without necessarily providing evidence or substantiation. Misinformation is, by definition, the spread of false or misleading information, often in the form of claims that lack verifiable facts. While the term misinformation has garnered significant attention, the claims drive the spread of this misinformation. The task of addressing misinformation, therefore, hinges on understanding and managing these claims at the core. The difficulty of analysing claims stems from their highly subjective and context-dependent nature. As early as the 1950s, philosopher Stephen Toulmin outlined a framework for argumentation, emphasizing the importance of claims as assertions that require proof (53). However, in social media, claims often lack context and proper grounding, making them harder to define and more prone to misinterpretation.

This thesis aims to address the challenges of misinformation by proposing a systematic framework to break down and assess claims in online content. The framework consists of four stages that are essential in ensuring that misinformation is efficiently understood and debunked – (a) claim detection, (b) claim simplification, (c) claim check-worthiness estimation and (d) claim verification. The structured approach illustrated in Figure 1.2 outlines the four critical stages followed in this thesis.

Step 1: Claim Detection. Claim Detection is the first step in automated fact-checking, where the system identifies statements that make assertions (54; 55; 56). Detecting claims in social media is even more challenging due to their highly subjective and context-dependent nature. Therefore, the first stage of the process focuses on identifying claims within the noisy social media environment. As illustrated in the figure, the first three posts are successfully detected as claims and proceed to the next steps for further evaluation. In contrast, the last example represents a personal opinion, which is not classified as a claim and is excluded at this stage.

Step 2: Claim Simplification. While existing pipelines focus on detecting and verifying claims, they often overlook the challenge posed by the linguistic complexity of online content. To address this, the second step focuses on isolating key assertions from the clutter of unstructured social media posts. Claims are refined to remove extraneous details, ensuring they are clear and concise for further evaluation.

Step 3: Claim Check-Worthiness. Not all claims warrant further scrutiny, and it is crucial to prioritize those that could have significant real-world consequences. Hence, the third stage of selecting check-worthy claims introduces a layer of intelligence into the fact-checking process. By filtering out less consequential claims, this step ensures the fact-checking process focuses on those needing verification (57; 58; 59). For instance, as shown in the figure, claims about vaccines and tax laws are flagged as check-worthy, while implausible or inconsequential claims, such as *'Bulls can't see red color,'* are filtered out.

Step 4: Claim Verification. The final phase of claim verification consists of analyzing relevant evidence to determine the truthfulness of each claim (60; 61; 62; 63; 64). Verification is vital to combating misinformation and ensuring false claims are not perpetuated within public discourse. For example, the claim that *'The new tax law will increase taxes for 80%.'* is verified as false, while the claim that *'Vaccines are effective in preventing measles and polio'* is verified as true. These examples underscore the importance of grounding the verification process in reliable evidence to maintain the integrity of public information.

These four steps – *detection, simplification, check-worthiness, and verification* – form a comprehensive approach for robust fact-checking. By focusing on the claim as the fundamental unit of misinformation, this thesis aims to create a more effective, scalable framework for automated fact-checking systems that can keep pace with the volume of claims circulating on social media. The ultimate goal is to improve the reliability of online information, enhance transparency in the fact-checking process, and reduce the spread of false claims in an increasingly digital world. Table 1.1 provides a conceptual breakdown of these steps, highlighting key challenges and our contributions.

1.5 Thesis Organization

The rest of the thesis is organized as follows, with each chapter contributing to the overarching goal of enhancing automated fact-checking in the context of social media.

- In Chapter 2, we review the literature relevant to automated fact-checking and other subtasks. This chapter critically examines prior research, highlighting critical methodologies, findings, and gaps that this thesis aims to address.
- In Chapter 3, we delve into the linguistic aspects of claim detection, focusing on the nuances of identifying claims within social media content. The chapter outlines a set of guidelines for claims, considering the complexity of language use and the challenges of distinguishing claims from other online discourse content.

Step	Core Challenges	Our Contributions
Claim Detection	1. Noisy, informal, and unstructured language on social media makes it difficult to identify assertive statements. 2. Distinguishing between opinions and factual claims is subtle and highly subjective. 3. The lack of universally agreed definitions for claims hinders consistent annotation and modeling.	1. We develop two claim detection frameworks – LESA and DESYR – specifically designed for noisy social media text. 2. We release a manually annotated dataset for claim detection in social media 3. We provide the first exhaustive set of claim annotation guidelines tailored for social media contexts. 4. We propose DARÉ embeddings and release a 100-dimensional pre-trained version for public use.
Claim Simplification	1. Social media posts often contain ambiguous and linguistically complex content that obscures the core assertion. 2. Generic summarization systems fail to capture claim-specific intent. 3. The absence of clear boundaries for what constitutes a claim adds further difficulty.	1. We introduce the novel task of Claim Simplification in the fact-checking pipeline. 2. We release two large-scale datasets for both abstractive and extractive claim simplification. 3. We propose robust claim simplification systems: DaBERTa for extractive simplification and CACN for abstractive simplification.
Claim Check-Worthiness	1. Current models lack explainability for why a claim is marked as check-worthy. 2. Fact-checkers need to manually interpret such claims due to the absence of justifications. 3. There is limited understanding of the contextual factors that make a claim check-worthy.	1. We introduce 'rationality labels' for check-worthiness estimation to improve explainability, focusing on Information, Social Impact, and Actionability. 2. We release a manually annotated dataset with check-worthiness and rationality labels. 3. We propose a unified framework for explainable claim check-worthiness that outputs both a check-worthiness label and accompanying rationality labels.
Claim Verification	1. Heavy reliance on external evidence retrieval leads to latency in verification. 2. There is limited capability to verify claims based on user knowledge or platform-specific context. 3. Existing systems often fail to capture subtle changes in quantitative claims.	1. We propose a novel framework that leverages user comments to assess the veracity of claims. 2. We design a system to verify quantitative claims by focusing on their statistical components. 3. We release two large-scale claim verification datasets annotated for veracity. One centers on user comments for contextual verification, and the other focuses on quantitative claims requiring statistical accuracy.

Table 1.1: Conceptual breakdown of the proposed framework, capturing core challenges and contributions of each step.

- In Chapter 4 and Chapter 5, we explore two distinct but complementary methods for claim simplification. Chapter 4 addresses the task of identifying the specific span of text that constitutes a claim. In contrast, Chapter 5 focuses on normalizing claims in an abstract manner.
- In Chapter 6, we discuss the criteria and methodologies used to assess which claims are worth fact-checking. This chapter outlines the process of determining check-worthiness, considering factors such as claim verifiability, social impact, etc.
- In Chapter 7 and Chapter 8 we examine two distinct approaches to claim verification. Chapter 7 investigates the verification of claims based on the stances and arguments presented by other users in online discussions. Meanwhile, Chapter 8 focuses on verifying claims by retrieving factual evidence from external sources.
- In Chapter 9, we summarize the key findings and contributions of this thesis. Further, we discuss the potential avenues for future research in the realm of automated fact-checking.

2. LITERATURE REVIEW

As the volume of user-generated content grows, the need for compelling content moderation and fact-checking has become increasingly urgent. In this chapter, we provide an overview of online misinformation and fact-checking research, emphasizing the computational approaches that leverage natural language processing and machine learning. While content moderation and fact-checking have been extensively studied within media studies and journalism, this thesis focuses on how advanced computational techniques can enhance these practices, addressing the challenges posed by the digital information landscape.

2.1 Misinformation as Phenomenon

As information flows rapidly across social media platforms, the lines between fact and fiction often blur, leading to widespread confusion and mistrust. In the context of social media, misinformation is pervasive, aided by algorithms that prioritize sensational content and by users who often engage with information that conforms to their biases. Recent studies highlight how misinformation spreads faster and reaches a wider audience than true information (65; 66). Cognitive biases such as *confirmation bias*, where individuals favour information that aligns with their pre-existing beliefs, also contribute significantly to the spread of misinformation (67; 68).

Misinformation is not a new phenomenon, but the digital age has intensified the speed, scale, and scope of its spread. Today, it has emerged as a pervasive and complex phenomenon that significantly impacts public discourse and decision-making. As a result, it has been a subject of discussion among researchers for several decades (69; 67; 66; 65; 68). At a broader level, misinformation is described as inaccurate or misleading information. However, the formal definition of misinformation varies across the literature, reflecting differing perspectives and nuances. Some scholars define it as false information shared *unintentionally*, emphasizing the absence of intent to deceive (70). Others consider it an umbrella term encompassing all types of false information, including fake news, rumours, and hoaxes, regardless of intent (71; 67; 72). These varying definitions, in turn, highlight the complexity of the concept and underscore the need for clarity when addressing the challenges posed by misinformation in different contexts. Thus, to clarify better, we formally define misinformation as follows and stick to it throughout this thesis:

> *Misinformation refers to false or inaccurate information that is spread, regardless of whether it is done intentionally or unintentionally. It can include errors or distortions of facts.*

Throughout this thesis, we refer to misinformation as an umbrella term to include all false or inaccurate information that is spread in online social media. Several related terms are often interchanged with misinformation. For instance, disinformation also refers to inaccurate information but is typically distinguished by its deliberate intent to deceive (6; 73). Fake news describes false information presented in the form of news, though it is not always disinformation, as innocent users may unknowingly share it (74). Rumours refer to unverified information, which can be either true or false, while spam involves irrelevant information broadcast to a large audience (75). However, in this work, we do not delve into these distinctions. Instead, we focus specifically on false information shared on social media and refer to it broadly as *misinformation*.

Claims are a crucial component of misinformation, acting as the foundation upon which false narratives

are built. Toulmin defines a *claim* as *'an assertion that deserves our attention,'* offering a starting point for understanding its role in discourse, though this definition remains somewhat vague. More formally, the Oxford Dictionaries describe a claim as *'a statement that something is true although it has not been proven, and other people may not agree with or believe it.'* This highlights the subjective nature of claims, which can vary widely across different domains, time periods, and contexts. Even for most individuals, constructing a set of persuasive and compelling claims that can effectively support an argument is complex and challenging. The difficulty lies in the subjective interpretation of claims and the constantly evolving landscape of social media language and jargon. As online discourse shifts and adapts, so too do the forms and strategies of misinformation, making it even harder to identify, evaluate, and debunk misleading claims in real time. The fluidity of social media platforms, with their rapid spread of information and emerging slang, adds another layer of complexity, allowing false claims to proliferate and evade scrutiny more easily than ever before.

2.2 Detection and Intervention Strategies

As social media platforms continue to grow, detection of misinformation has become a critical area of research. Traditional methods of content moderation, such as manual fact-checking, have proven insufficient due to the sheer scale of information generated daily. As a result, automated tools powered by machine learning and artificial intelligence have become central to misinformation detection.

In recent years, there have been notable breakthroughs in automated fact-checking. Researchers have sought to develop models that can reliably detect misinformation across various domains. Most of these efforts have focused on textual information, verified against structured or unstructured data sources. Graph-based models have also been used to facilitate the reasoning over multiple pieces of evidence (76; 77). Although such models achieve sizable performance gains, they lack explainability, and they rely on large amounts of training data. Recent studies indicate that LLMs can perform well and be dependable for verification tasks despite the possibility of hallucinations (78). Lee et al. (79) demonstrated that the inherent knowledge of LLMs can be leveraged for fact verification. Previous research suggests that incorporating external information improves performance on reasoning-intensive tasks (80; 81). Recent advancements aim to simplify complex claims into manageable sub-questions to enhance evidence retrieval, showing promise in improving fact-checking accuracy, particularly for claims involving implicit reasoning or multiple verification steps. A plethora of research has been conducted focusing on various aspects of identifying and mitigating misinformation, including fake news, rumours, etc (82; 83; 84; 85). Researchers have investigated several approaches for identifying misinformation, including methods that analyze the content, examine the network, and predict falsehoods in advance. Content-based approaches rely heavily on linguistic clues such as writing styles, lexical aspects, sentiment analysis, and subject relevance (83; 84; 85). For example, Castillo et al. (86) discovered that highly reputable social media messages have more URLs and longer text lengths than less credible ones. Similarly, Pawan et al. (87) introduced the WELFake model, which detects bogus news by combining word embedding vector and linguistic data. Anshika et al. (85) proposed a deep learning model that leverages news stories' syntactic, grammatical, semantic, and readability aspects to identify bogus news. Validated over several publicly accessible datasets, Jain et al. (84) created the Confake algorithm using a comprehensive set of content-based features and word vector attributes taken from news items. These content-based methods analyze the credibility of individual posts in isolation, often neglecting the high correlation between a post's veracity and the replies it garners from other users. Current research has concentrated on propagation-based approaches to overcome the limitations of content-based methods. These methods employ social context knowledge to identify false information by examining the dissemination of information on social networks, the people responsible for disseminating it, and the relationships between these propagators

(88; 89; 90; 91). Ma et al. (88) developed a neural network with a tree structure that uses false news cascades to detect disinformation. Dhawan et al. (89) introduced GAME-ON, a system that utilizes Graph Neural Networks to enable detailed interactions inside and between several modalities to detect multimodal fake news. Song et al. (90) created a temporal learning model based on graphs to capture the changing patterns of tweets organized as tree-structured data. Kang et al. (92) constructed a news detection graph that links bogus news with many third-party information sources. Sun et al. (93) proposed a hypergraph learning model that utilizes a unique hyperedge walking technique and hyperedge expansion method to create comprehensive representations for entire graphs. These studies indicate that models utilizing network information outperform those that merely rely on content. In addition to propagation features, user-specific attributes, including the number of followers and the stance of other users toward the post, have also been employed to detect misinformation. The significance of other users' reactions in verifying information has been emphasized in numerous studies (86; 94; 95). Li et al. (96) conducted a study on the semantic aspects of false information, analyzing their dissemination and user characteristics, using 421 false statements and 1.47 million related tweets. One of the most significant discoveries is that individuals are inclined to disseminate falsehoods without incorporating their personal opinions when uncertain about their veracity. According to Zubiaga et al. (96), users are inclined to endorse unverified information. In contrast, highly reputable users, such as news organizations, endeavour to publish well-reasoned statements supported by evidence and appear to be certain. These findings underscore that models using network information and user responses are highly efficient for misinformation detection on social media. With the quick dissemination of information on social media, false information can reach thousands of users in minutes, causing significant confusion, panic, and even physical harm. It is essential for real-time systems to identify misinformation at its inception before its rampant dissemination. Early detection algorithms have been the subject of numerous investigations (97; 98; 99). Kwon et al. (97) analyzed feature stability over time. They discovered that user and linguistic features are more effective than structured and propagation features in determining the veracity of information during the early stages. Using network embedding techniques on social network graphs, Liu and Wu (99) constructed user representations using network embedding approaches on the social network graph.

Datasets have played a pivotal role in standardizing automated fact-checking and encouraging the development of benchmark models. FEVER dataset is one of the most recognized, primarily focusing on textual claims sourced from Wikipedia (100). While FEVER has been instrumental in training fact-checking models, only about 10% of the claims in its FEVEROUS (101) extension involve numerical reasoning. This ongoing deficiency is a common limitation across many datasets, which often rely on synthetic claims or oversimplified representations that fail to reflect real-world complexities. Datasets like TabFact (102), which includes claims derived from Wikipedia tables, and SciTab (103), designed for scientific claims, do not adequately capture the nuances of verifying numerical content in broader contexts. Efforts to compile real-world claims have been more effective in the political sphere, with datasets like ClaimDecomp (104), LIAR (105), and MultiFC (106) featuring claims verified by professional fact-checkers. However, these datasets do not prioritize numerical claims specifically, and their treatment of statistical and temporal expressions is limited. In contrast, earlier works focusing on quantitative claims have had a narrow scope. Recently, Venktesh et al. (107) introduced the QuanTemp dataset that focuses on validating numbers across various domains. However, our proposed dataset goes beyond this by encompassing a broader range of quantitative aspects. While QuanTemp targets verifying explicit numbers only, our dataset also addresses the complexities of comparative, statistical, interval, and temporal elements. This comprehensive approach allows for a more nuanced understanding of claims, making our dataset better suited for developing advanced fact-checking models that tackle real-world quantitative assertions' intricacies.

Even though ample studies have handled automated fact-checking and misinformation detection, they still fall short of capturing the nuanced nature of claims, particularly in the social media context.

These systems often focus on retrieving evidence to support or refute information, overlooking the subtle linguistic and contextual complexities inherent in the claims. Additionally, the rapid spread of content across social networks allows misinformation to proliferate faster than these systems can extract and verify evidence. Social media's diverse and noisy nature further complicates post-level analysis, making it computationally expensive and prone to inaccuracies. To enhance effectiveness in combating misinformation, we advocate for a shift towards *claim-oriented approaches* that concentrate on extracting and analyzing specific claims.

2.3 Claim-Oriented Approaches

Traditional fact-checking systems begin with a statement, which is then analyzed to find supporting evidence. In contrast, claim-oriented approaches break this process into several simpler tasks targeting the claim itself. This typically involves a series of steps: identifying the claims made in the statement, prioritizing the most important claims, gathering relevant evidence associated with those claims, and concluding with a final verdict by evaluating the claims against the gathered evidence. In addition to these steps, we introduce a claim simplification step to decompose complex claims into shorter, more precise forms. This structured pipeline enhances clarity and efficiency in the fact-checking process.

2.3.1 Claim Detection

Claim detection refers to the process of identifying and extracting assertions that can be verified as true or false from a given text. In the context of fact-checking, this is a critical first step because not all posts on social media are verifiable claims. Some statements may be opinions, non-factual statements, or merely background information. In the past decade, the task of claim detection has become a popular research area in text processing with a pioneering attempt by Rosenthal et al. (108). They used a supervised technique to mine claims from Live Journals and discussion forums, utilizing features based on sentiment and word-gram analysis. Levy et al. (54) proposed a context-dependent claim detection (CDCD) approach. They described 'context-dependent claim' (CDC) as '*a general, concise statement that directly supports or contests the given Topic.*' Their approach was evaluated over Wikipedia articles; it detects sentences that include CDCs using context-based and context-free features. This is followed by ranking and detecting CDCs using logistic regression. Lippi et al. (55) proposed context-independent claim detection (CICD) using linguistic reasoning and encapsulated structural information to detect claims. They used constituency-parsed trees to extract structural information and predicted parts of the sentence holding a claim using SVM. Although their approach achieved promising results, they used the Wikipedia dataset, which was highly structured and domain-dependent. Daxenberger et al. (109) used six datasets and contrasted the performance of several supervised models. They performed two sets of experiments – in-domain CD (trained and tested on the same dataset) and cross-domain CD (trained on one and tested on another unseen dataset). They learned divergent conceptualisations of claims over cross-domain datasets. Levy et al. (110) proposed the first unsupervised approach for claim detection. They hypothesised a '*claim sentence query*' as an ordered triplet: $\langle \text{that} \rightarrow \text{MC} \rightarrow \text{CL} \rangle$; according to the authors, a claim begins with the word 'that' and is followed by the main concept (MC) which is further followed by words from a pre-defined claim lexicon (CL). This strategy would not work well for texts generated by social media platforms due to a lack of structure and the usage of *'that'* as an offset for statements. Transformer-based language models have been used for claim identification as deep neural networks have gained popularity. Chakrabarty et al. (111) used over 5 million self-labelled Reddit comments to fine-tune their model. However, they made no attempt to encapsulate the structure of a sentence. Recently, CLEF-2020 shared task (112) attracted multiple models which are tweaked specifically for claim detection. Williams et al.

(113) bagged the first position in the task using a fine-tuned RoBERTa (114) model with mean pooling and dropout. First runner-up of the challenge, Nikolov et al. (115) used logistic regression on various meta-data tweet features and a RoBERTa-based prediction. Cheema et al. (116), the second runner-up, incorporated pre-trained BERT embeddings and POS and dependency tags as features trained using PCA and SVM models.

2.3.2 Claim Simplification

Claim simplification is key in making complex, convoluted claims more accessible for analysis and verification. In many cases, claims may be difficult to verify directly because they are worded in a complicated or ambiguous manner. Simplification techniques help transform these claims into more straightforward statements that retain their original meaning but are easier to assess. Zaidan et al. (117) first introduced the concept of rationales, highlighting text segments that supported their label's judgment. They reported a significant boost in performance when they included these rationales in the training process for sentiment classification of movie reviews. Trautmann et al. (118) released the Argument Unit Recognition and Classification *AURC*-8 dataset with token-level span annotations for the argumentative stance components and their corresponding label. Mathew et al. (119) proposed a quality corpus for explainable hate identification with token-level annotations. The *SemEval* community has initiated fine-grained span identification concerning other domains of argument mining such as toxic comments (120) and propaganda techniques (121). These shared tasks amassed many solutions constituting transformers (122), convolutional neural networks (123), data augmentation techniques (124; 125), and ensemble frameworks (126; 127). Wuhrl et al. (56) resembled the closest study to ours, wherein they compiled a corpus of around $1.2k$ biomedical tweets with claim phrases. In summary, existing literature on claims concentrates entirely on sentence-level claim identification and does not investigate eliciting fine-grained claim spans.

The task of claim normalization is closely related to text summarization. In the latter, given a lengthy document, the goal is to summarize it into a much shorter summary. Previous work on text summarization has explored various approaches, including large pre-trained seq2seq models to generate high-quality summaries (128; 129; 130). One issue has been the faithfulness of the summary with respect to the source. To address this, Kryscinski et al. (131) introduced FactCC, a weakly supervised BERT-based entailment model, which augments the dataset with artificially introduced faithfulness errors. Similarly, Utama et al. (132) trained a model for detecting factual inconsistencies in data from controllable text generation that perturbs human-annotated summaries, introducing varying types of factual inconsistencies. Durmus et al. (133) proposed a question-answering framework that compares answers from the summary to those from the original text. All these approaches primarily focused on general-purpose summarization and did not provide means for models to generate summaries primarily focusing on specific needs. To address this limitation, controlled summarization was introduced by Fan et al. (134). One aspect of controlled summarization is length control, in which users can set their preferred summary length (135; 136). Recent research has discovered that, despite their fluency and coherence, state-of-the-art abstractive summarization systems produce summaries with contradictory information. While text summarization systems can assist in condensing social media posts into shorter summaries, their primary goal is not to ensure verifiability. It aims to capture the text's key points rather than emphasise the specific claims that must be fact-checked.

2.3.3 Claim Check-Worthiness

Once a claim has been identified and simplified, the next crucial task is determining whether the claim is worth verifying—this is known as check-worthiness assessment. Not all claims need to be fact-checked, as many may be trivial, irrelevant, or already widely accepted as true. The task of identifying check-worthy claims has drawn the attention of numerous researchers (137; 138; 139). With a pioneering attempt by Hassan et al. (140), claim check-worthiness detection has emerged as a major research area in recent years. They developed ClaimBuster, the first-of-its-kind system to target a claim's check-worthiness in a debate. It was trained on a massive manually annotated dataset of US election debates from 1960 to 2012, totalling 30 debates and $28,029$ transcribed sentences. Each statement made during a political discussion was categorized into one of three categories: non-factual, unimportant factual, or check-worthy factual. Focusing on the 2016 US Presidential debates, Gencheva et al. (137) obtained binary annotations for check-worthiness from various fact-checking organizations. Later, they developed ClaimRank, which was trained on additional data, including Arabic content. To determine whether a sentence would be selected for fact-checking, Patwari et al. (141) used a model that is similar to boosting. (139) used a multi-task learning neural network. The task was to predict whether a sentence would be selected for fact-checking by each of the nine different fact-checking organizations: CNN, The Washington Post, PolitiFact, FactCheck, ABC, NPR, NYT, Chicago Tribune, and The Guardian. The CheckThat! Lab has organized similar shared tasks since 2018, focusing on political debates and speeches (142; 143; 144). Research on estimating claim-worthiness in debates and political speeches has been thoroughly studied and made substantial progress. Due to the lack of structure and proper linguistic properties, these approaches do not work well for noisy social media texts.

The social media expansion has boosted the preponderance of false and misleading claims; as a result, current research has turned to identifying claim check-worthiness in social media. To address this, in their recent editions, CheckThat! Labs organized shared tasks on identifying check-worthy claims with a primary focus on social media texts (142; 143; 144). These shared tasks attracted multiple systems modelled to handle noisy text from social media platforms. The 2020 edition featured three central tasks: detecting previously fact-checked claims, evidence retrieval, and actual fact-checking of claims. Along with English, they offered the tasks in Arabic and Spanish. Several teams improved pre-trained models for Arabic, such as AraBERT and multilingual BERT (145; 113). In the case of the English task, multiple systems harnessed the strength of pre-trained Transformers, specifically BERT and RoBERTa (113; 146). Other methods extracted tweet embeddings from tweets using pre-trained models like GloVe and Word2Vec, which were then fed into a neural network or an SVM. The top-ranked system in the 2021 edition also leveraged transformer-based models (142). The first shared task in the 2022 edition was anticipating which Twitter posts should be fact-checked, emphasizing COVID-19 and politics. It was put forth in Arabic, Bulgarian, Dutch, English, Spanish, and Turkish (143). A total of 19 teams competed, with most submissions achieving significant improvements over baselines using Transformer-based models such as BERT and GPT-3. Alam et al. (147) created a multi-question annotation schema of COVID-19 tweets structured around seven claim check-worthiness questions. While better models for detecting claim check-worthiness are constantly being developed, there is a lack of literature on the explainability aspects of these binary decisions.

2.3.4 Claim Verification

Claim verification is the final step, where the goal is to determine whether the claim is true or false based on the available evidence. The widespread dissemination of false or misleading information on social media platforms has attracted much scrutiny recently. The process involves analyzing claims, retrieving relevant evidence, and determining their veracity based on the evidence. Over the years, numerous approaches

have been developed. A common framework in claim verification involves a two-stage pipeline: evidence retrieval and verification. In the first stage, systems retrieve relevant evidence from trusted sources, such as encyclopedias, news articles, or structured databases. Methods like BM25 (148) and neural models such as BERT-based retrieval (149) have been used to enhance retrieval performance. The second stage evaluates the retrieved evidence to determine whether it supports or refutes the claim. Recent approaches have emphasized semantic and linguistic analyses to improve claim verification accuracy. Studies have highlighted challenges in handling nuanced languages, such as hedging, metaphors, and implicit meaning (150). Transformer-based models have demonstrated state-of-the-art performance in capturing contextual nuances. For instance, neural entailment techniques (151) and semantic similarity models (152) align claims with supporting evidence more effectively. Graph-based reasoning methods have also emerged, leveraging knowledge graphs to link claims with evidence and contextual data (153; 154). However, many existing systems aim to handle entire social media posts at once, which introduces inefficiencies and overlooks the finer granularity of individual claims.

Question-answering has recently emerged as a potential strategy for claim verification. Yang et al. (155) developed a model that does not require a specific dataset of annotated question-answer pairings. Two recent datasets, Fan et al. (156) and Chen et al. (104), have framed claim verification as a question-answer task. However, the question-answer pairs used by Fan et al. (156) were primarily designed to give relevant context rather than to capture the entire fact-checking process, resulting in insufficient evidence. Ousidhoum et al. (157) emphasized context dependence as a major issue in question-answer generation, stating that many queries cannot be generated only from the claim since they refer to entities and events stated only in the original fact-checking articles. Chen et al. (104) sought to establish evidentiality sufficiency but did not demonstrate success. Furthermore, their proof was drawn straight from fact-checking publications published after the assertions, which raises the possibility of temporal leaking. Recently, Schlichtkrull et al. (64) illustrated that reasoning about evidence could be efficiently modelled using an approach based on questions and answers. They engaged human annotators to formulate questions and supply akin responses that included evidence supporting the claim.

2.4 Research Gap and Questions

The review of the literature reveals significant advancements in automated fact-checking; however, it also highlights critical gaps, particularly in handling the scale and complexity of content on social media. Existing systems that handle multiple social media posts simultaneously are often inefficient and struggle to cope with the diverse and noisy nature of user-generated content. While claim-oriented techniques provide a more concentrated alternative, they fall short of capturing the subtle syntax and contextual intricacies that exist in real-world claims. To address these limitations, this thesis aims to develop a claim-centric fact-checking framework guided by the following research questions:

RQ1: *Given that existing claim detection systems have largely overlooked the unique linguistic and structural characteristics of social media text, how can complex claims be effectively identified within such unstructured content?*

RQ2: *How can we effectively extract and refine key assertions from noisy, linguistically complex social media claims to support accurate downstream fact-checking?*

RQ3: *How can we enhance claim check-worthiness detection by moving beyond binary classification to provide more explainable and human-aligned reasons for prioritizing claims in fact-checking pipelines?*

RQ4: *Given the reliance on external evidence in existing claim verification systems, what potential do user responses hold for enhancing the speed and contextual accuracy of claim verification?*

Part 1

Understanding and Detecting Claims

3. CLAIM DETECTION

The formulation of a claim rests at the core of automated fact-checking. Demarcating between a claim and a non-claim is arduous for humans and machines, owing to latent linguistic variance between the two and the inadequacy of extensive definition-based formalization. Furthermore, the increase in online social media usage has resulted in an explosion of unsolicited information on the web presented as informal text. A pressing issue in this area of research is the unavailability of labelled datasets for claim detection. Thus, in this chapter, we develop a Twitter dataset that aims to provide ground for testing on unstructured claims. We also release comprehensive data annotation guidelines that emerged during the annotation phase (which was missing in the current literature). Additionally, we present two claim detection frameworks – LESA and DESYR. LESA aims to advance headfirst into expunging the former issue by assembling a source-independent generalized model that captures syntactic features through part-of-speech and dependency embeddings and contextual features through a fine-tuned language model. Experimental results elucidate that LESA improves upon the state-of-the-art performance across six benchmark claim datasets by an average of 3 claim-F1 points for in-domain experiments and by 2 claim-F1 points for combined-domain experiments. We propose another framework, DESYR, that intends to annul the issues for informal web-based text by leveraging a combination of hierarchical representation learning (dependency-inspired Poincaré embedding), definition-based alignment, and feature projection. We do away with fine-tuning compute-heavy language models in favour of fabricating a more domain-centric but lighter approach. We see an increase of 3 claim-F1 points on our Twitter dataset, an increase of 1 claim-F1 point and 9 macro-F1 points on the Online Comments (OC) dataset, an increase of 24 claim-F1 points and 17 macro-F1 points on the Web Discourse (WD) dataset, and an increase of 8 claim-F1 points and 5 macro-F1 points on the Micro Texts (MT) dataset. We also perform an extensive analysis of the results.

3.1 Introduction

The concept of claim is entrenched in the roots of the misinformation. Toulmin (53), in his argumentation theory, described the term 'claim' as '*an assertion that deserves our attention*'; albeit not very precise, it still serves as an initial insight. In recent years, Govier (158) described a 'claim' as '*a disputed statement that we try to support with reasons*.' The quandary of claim detection exists, given the disparity in conceptualization and lack of a proper claim definition. The task of claim detection across different domains has garnered tremendous attention so far owing to an uprise in social media consumption and, by extension, the existence of fake news, online debates, widely-read blogs, etc. As an elementary example, claim detection can be used as a precursor to fact-checking, wherein segregation of claims aids in restricting the corpus that needs a fact-check. A few examples are shown in Table 3.1. Most existing works are built upon two fundamental pillars – contextual (109; 111) and syntactic (54; 55). Firstly, they mainly focus on adapting to texts from similar distributions, topics, or both. Secondly, they often exercise against well-structured and laboriously pre-processed formal texts because they lack a labelled corpus comprising unstructured texts. As a result, claim detection from unstructured raw data remains under a relatively less explored umbrella.

Text	Claim?
Alcohol cures corona.	Yes
Wearing mask can prevent corona.	Yes
Lord, please protect my family & the Philippines from the corona virus.	No
If this corona scare doesn't end soon imma have to intervene	No

Table 3.1: A few examples of claim and non-claim.

Claims can be sourced from various references, e.g., online social media texts, microblogs, Wikipedia articles, etc. It is, however, crucial to pay special attention to claims impending from online social media sites. As a major social media platform, Twitter provides a perfect playground for different ideologies and perspectives. Over time, it has emerged as the hub for short, unstructured text that describes anything from news to personal life. Most individuals view and believe things that align with their compass and prior knowledge, *aka conformity bias* (159) – users tend to make bold claims that usually create a clash between users of varied opinions. At times, these claims incite a negative impact on individuals and society. As an example, a tweet that reads "*alcohol cures corona*" can lead to a mass frenzy and massive retweeting, especially in times of a pandemic, when people are more vulnerable to suggestion. In such cases, automated promotion of claims for immediate further checks has been paramount. An automated system is especially pivotal since OSM data is far too voluminous to allow for manual human checks, even if it was an expert. Nowadays, most claims can be seen over social media, and the highly unstructured nature of these platforms makes it more difficult to process and capture their semantics. Thereby, the need of the hour is to develop an efficient model for claim detection in social media texts. However, a major bottleneck is the unavailability of an annotated dataset. This chapter acknowledges this bottleneck and develops a qualitative annotated resource for claim detection in tweets.

Since the task of claim detection has a strong association with the structure of the input, as argued by Lippi et al. (55), we leverage two linguistic properties – part-of-speech and dependency tree, in LESA to capture the linguistic variations of each category. Subsequently, we amalgamate these features with BERT (160) for classification. We also propose DESYR, which achieves competence over the better segregation of feature space for classification and, moreover, learns to leverage the guidelines for identifying claims and non-claims to further the feature constitution. It attains this by employing an intelligible unity of feature projection, attention-based alignment, and pre-transformer Deep Learning. Also, in line with Lippi et al. (55), who argued that a claim could be grounded in linguistics, we propose DARÉ embedding, which is a dependency-inspired variant of Poincaré embedding (161). It helps assimilate enhanced representations of word vectors by capturing intrinsic hierarchies in dependency trees. We evaluate LESA and DESYR on different datasets (including our Twitter dataset). The comparative study suggests the superior performance of our proposed systems against the existing systems. We conduct a result analysis across several baselines and a general analysis for our predictions. The comparative investigation espouses the finer performance of DESYR compared to other state-of-the-art systems. We summarize the major contributions below:

- **Twitter dataset and comprehensive annotation guidelines.** To mitigate the unavailability of the annotated datasets for claim detection in Twitter, we develop a large COVID-19 Twitter dataset, the first of its kind, with $\sim 10,000$ labelled tweets, following a comprehensive set of claim annotation guidelines.

- **LESA, a generalized claim detection system.** We propose a generalized claim detection model, LESA, that identifies the presence of a claim in *any* online text without prior knowledge of the source. To our knowledge, this is the first attempt to define a model that handles claim detection from structured and unstructured data in conjunction.

- **DESYR, a claim detection system tailored towards informal web-based texts**. The system

determines claims in the online text by aligning the query to encoded definitions and projecting them into a purer space.

- **Comprehensive evaluation and state-of-the-art results.** We evaluate LESA and DESYR against the state-of-the-art claim detection systems. The contrast suggests the superiority of our proposed architectures, and our ablation study highlights the importance of each module.
- **Dependency-inspired Poincaré embedding**. We propose DARÉ embeddings that show promising results for linguistically grounded NLP tasks. We release a pre-trained 100-dimensional version of DARÉ.

3.2 Related Work

In the past decade, the task of claim detection has become a popular research area in text processing with a pioneering attempt by Rosenthal et al. (108). They worked on mining claims from Live-Journal and Wikipedia discussion forums and employed a supervised approach with features based on sentiment and word-grams. Their methodology was limited in the number of frequency-based features that could be hand-constructed. Most of the existing works proposed were limited to a specific domain. Following this, Levy et al. (54) proposed a context-dependent claim detection (CDCD) approach. They described 'context-dependent claim' (CDC) as '*a general, concise statement that directly supports or contests the given Topic.*' It is from here that topic-dependent claim detection picked up pace. The authors stressed that every argument contains a claim that the argument aims to prove. They asserted that to define an AM tool, it is vital to understand the context (topic). Their approach was evaluated over Wikipedia articles and predefined topics. Given a topic and a set of relevant articles, it detects sentences that include CDCs using context-based and context-free features. This is followed by ranking and detecting CDCs using maximum likelihood and logistic regression. Lippi et al. (55) proposed a less restrictive approach called context-independent claim detection (CICD). It used linguistic reasoning and encapsulated structural information to detect claims. They used constituency-parsed trees to extract structural information and predicted parts of the sentence holding a claim using SVM. Although their approach achieved promising results, they used the Wikipedia dataset, which was highly structured and domain-dependent. Recent years have seen a shift from context-dependent to cross-domain setups. Daxenberger et al. (109) used six datasets and contrasted the performance of several supervised models. They performed two sets of experiments – in-domain CD (trained and tested on the same dataset) and cross-domain CD (trained on one and tested on another unseen dataset). They learned divergent conceptualisations of claims over cross-domain datasets and performed weakly unsupervised learning in domain adaptation. Levy et al. (110) proposed the first unsupervised approach for claim detection. According to the authors, a claim begins with the word 'that' and is followed by the main concept which is further followed by words from a pre-defined claim lexicon.

With the growing precedence of deep neural networks, transformer-based language models have been employed for claim detection. Chakrabarty et al. (111) used over 5 million self-labelled Reddit comments that contained the abbreviations IMO (In My Opinion) or IMHO (In My Honest Opinion) to fine-tune their model. However, they made no attempt to encapsulate the structure of a sentence. Recently, CLEF-2020 shared task (112) attracted multiple models which are tweaked specifically for claim detection. Williams et al. (113) bagged the first position in the task using a fine-tuned RoBERTa (114) model with mean pooling and dropout. First runner-up of the challenge, Nikolov et al. (115) used logistic regression on various meta-data tweet features and a RoBERTa-based prediction. Cheema et al. (116), the second runner-up, incorporated pre-trained BERT embeddings and POS and dependency tags as features trained using PCA and SVM models. Traditional approaches focused primarily on the syntactic representations of

claims (55; 54; 110) and textual feature generation, while recent neural methods leverage the transformer model (108). With the advent of neural networks, the focus has shifted to semantic representations of input using transformer-based models (109; 111; 113).

3.3 Dataset

A few datasets exist for claim detection in the online text; however, most are formal and structured texts (162; 163). As we discussed earlier, OSM platforms are overwhelmed with various claim-ridden posts. Despite the abundance of tweets, the literature does not suggest any significant effort for claim detection on Twitter, and arguably, the prime reason is the lack of a large-scale dataset. Recently, a workshop on claim detection in Twitter was organized under CLEF-2020 (112). It had two subtasks related to the claim identification with separate datasets. The first dataset consists of $1,060$ COVID-19 tweets for claim detection, whereas the second one comprises another $1,000$ tweets for claim retrieval. There are $1,704$ claim tweets and 365 non-claim tweets. Another recent dataset on claim detection has only 305 claim and 199 non-claim tweets (1). Unfortunately, such a few instances are unarguably insufficient to develop an efficient deep neural model. To the best of our knowledge, no large-scale dataset exists for detecting claims in unstructured, noisy data. Therefore, we attempted to develop a new and relatively larger dataset for claim detection in OSM platforms. We collected $\sim 40,000$ tweets from various sources (164; 165; 166; 167; 168) and manually annotated them. We additionally included claim tweets of Alam et al. (1) and CLEF-2020 (112).

3.3.1 Preprocessing

Before annotating the dataset, we perform the preliminary task of data cleaning. Our pre-processing stage involves removing hashtags, URLs, user handles, and non-ASCII characters. All tweets with a character count of less than 20 and a word count of less than 4 are also removed, owing to the lack of context for their interpretation. Finally, we spell-check the words using symspellpy[1]. The final dataset contains $9,894$ tweets, which were split into 70% for training, 15% for validation, and the remaining 15% for testing.

3.3.2 Annotation

We extended the claim annotation guidelines of Alam et al. (1) to annotate the remaining tweets. The authors targeted and annotated only a subset of claims, i.e., factually verifiable and non-factually verifiable claims. They did not consider personal opinions, sarcastic comments, implicit claims, or claims existing in a sub-sentence or sub-clause level. This motivated us to extrapolate the existing guidelines to be more exclusive, nuanced, and applicable to diverse claims. Our official definition adopted for claims is to *state or assert that something is the case, with or without providing evidence or proof.* Following are our guidelines for what qualifies as a claim. Anything not qualifying as a claim is labelled as 'non-claim'. However, certain clarifying guidelines are also given for non-claims. The tweets are labelled 1 for claim, 0 for non-claim and x in obscure situations.

[1] https://pypi.org/project/symspellpy/

Guidelines for Claims

- Tweets mentioning statistics, dates or numbers.
  Example: *["just 1 case of corona virus in india and people are crazy for masks daily 400 people die in road crashes still no craze for helmetsthinking face safetysaves be it virus or road crashes"]*

- Tweets mentioning a personal experience.
  Example: *["i live in seattle i have all symptoms of covid19 and have a history of chronic bronchitis since i work in a physical therapy clinic with many 65 patients and those with chronic illnesses i decided to be responsible and go to get tested this is how that went"]*

- Tweets 'reporting' something to be true or an instance to have happened or will happen.
  Example: *["breaking boris johnson says he visited kettering hospital shook hands with corona patients but the hospital doesnt have corona cases shaking hands would be dangerous not sure how ill gloss over the fact the pm is a liar a complete fucking idiot but ill find a way x"]*

- Tweets containing verified facts also account for a claim, a veracious claim that is.
  (*Note*: a fact known by one, may not necessarily be known by another)
  Example: *["The Chinese CDC has started research and development of a vaccine for the #coronavirus."]* - known fact

- Tweets that negate a possibly false claim are also accounted as claims.
  Example: *["disinfectants are not a cure for coronavirus"]*

- Tweets that indirectly (subtly) imply that something is true.
  Example 1: *["b52questions 1 why is rudy not under arrest 2 why is harvey not in rikers 3 why is cuccinelli still working 4 has barr quit yet 5 when is flynn being sentenced 6 who trusts pence and mrs miller with messaging about coronavirus 7 are rs happy wtheir guy"]* - indirectly implies rudy is not under arrest and harvey is not in rikers
  Example 2: *["do rich people know theyll get the virus if poor people cant be tested and diagnoseddo rich people know theyll get the virus if poor people cant be tested and diagnosed"]* - indirectly implies rich people will get the virus if poor people can't be tested

- Claims made in sarcasm or humour.
  Example 1: *["@TheDailyShow Newsflash! #trumpfact If you paint your face #orange you will be #immune to #coronavirus"]*
  Example 2: *["RT @_saraellen: If you've ever used messers bathroom you're immune to corona virus."]*
  Example 3: *["corona virus minding its business by avoiding africa and going to other continents"]*
  Example 1 and 2 are both examples of claims, even if evidently sarcastic. Example 3 is a humour oriented claim

- All opinions are not claims. Opinions that have societal implications are considered as claims.
  Example 1: *["@derekgilbert I think the Chinese stole a bio weapon https://t.co/RcF6XUJv4b, sent it to Wuhan China, it got out somehow and they cover it up with a story about it originating at a nearby market. They know how bad the virus is and quarantine entire cities. https://t.co/ZwMqsiWGau"]*

  Example 2: *["I think Burger King fries are better than Mc'D's"]*
  Example 1 is a opinion that claims something to have happened, whose veracity will affect a certain section of the society. Example 2, on the other hand, is a personal belief that majorly impacts only the person making the tweet.

- Tweet that says something is true and provides an attachment as evidence or to support the statement.
  Example 1: *["RT @Jawn42: If you ate here growing up, you're immune to the Coronavirus. https://t.co/b9a0hm171b"]*
  Example 2: *["RT @TerminalLance: This kills Coronavirus in the system https://t.co/iOFNlkSrUj"]*

  * However, if a person says something is provided in the attachment, that will not be a claim.
  Example: *["im stunned by the depth of coronavirus information being released in singapore on this website you can see every known infection case where the person lives and works which hospital they got admitted to and the network topology of carriers all laid out on a timeseries link"]*

- A claim can be a sub-part of a question.
  Example: *["Does the pneumonia shot help protect from developing pneumonia caused by #covid19"]* - the claim here being that pneumonia is caused by COVID-19.

Guidelines for Non-claims:

- Hoping that something happens or feeling something is true is not claiming it.
  Example 1: *["World doesn't end if u don't give your opinions about corona virus. I'm drinking #nilavembu boiled in hot water and hope it prevents. #COVID-19 #coronapocolypse"]*
  Example 2: *["I feel like I'm immune to coronavirus."]*

- Inclusion of words that project doubt over the said statement.
  Example: *["politicalelle Political correctness infecting the #Coronavirus. Let's change words describing the virus maybe that will cure it."]*

  * Tweets containing doubt-casting words can still, however, contain claims.
  Example: *["Coronavirus may have originated in lab linked to China's biowarfare program #coronavirus https://t.co/2NSWidMkoa"]*

- Urging one to not claim something or to spread misinformation is not a claim.
  Example: *["#Covid19 - Dear all: Stop telling the public that plaquenil/Azithromycine is a cure!!!!! Plz some leadership is needed regarding this matter!"]*

- Questioning a possible claim is not a claim.
  Example: *["Do disinfectants really cure Corona?"]*

  * However, a tweet containing a question can still comprise a claim.
  Example: *["Would you like to promote a cure that can kill 90% #COVID-19 virus in the body in 3-15 min?"]* - the tweet claims that there exists a cure that can kill 90% of COVID-19 virus in the body

- Warning someone against a claim is not a claim.
  Example: *["If you think drinking disinfectants will cure #Covid_19 , you deserve death #trump"]* - this tweet may be hate speech but is not claiming something to be true or untrue.

Following the extended guidelines, we annotated the collected tweets, and to ensure coherence and conformity, we re-annotated the tweets of Alam et al. (1) and CLEF-2020 (112). It is intriguing to see the differences and similarities between the two guidelines; therefore, we compile a confusion matrix for the CLEF-2020 claim dataset, as presented in Table 3.2. Each tweet in our corpus of $9,894$ tweets has been annotated by at least two annotators, with an average Cohen's kappa inter-annotator agreement (169)

score of 0.62. In case of a disagreement, the third annotator was considered, and a majority vote was used for the final label.

	Our Annotation	
CLEF-2020	Non-claim	Claim
Non-claim	301	47
Claim	64	550

Table 3.2: Confusion matrix highlighting the differences and similarities between CLEF-2020 (1) and our annotation guidelines for CLEF-2020 claim dataset.

Other Datasets: Since we attempt to create a generalised model that can detect the presence of a claim in any online text, we accumulate, in addition to the Twitter dataset, six publicly available datasets: (i) Online Comments (OC) containing Blog threads of LiveJournal (170), (ii) Wiki Talk Pages (WTP) (171), (iii) German Micro-text (MT) (162), (iv) Persuasive Student Essay (PE) (163), (v) Various Genres (VG) containing newspaper editorials, parliamentary records and judicial summaries, and (vi) Wed Discourse (WD) containing blog posts or user comments (172).

Furthermore, considering the structure of the input texts in these datasets, we group them into three categories as follows: Noisy (Twitter), Semi-noisy (OC, WTP), and Non-noisy (MT, PE, VG, WD). This categorization considers factors such as grammaticality, consistency, spelling, vocabulary, and overall text structure. Noisy datasets contain highly informal, user-generated content with irregular grammar, abbreviations, slang, and typos. Semi-noisy datasets feature conversational or partially moderated text, which may still include informal expressions or inconsistencies. Non-noisy datasets consist of edited or professionally written content, typically exhibiting standard grammar and well-formed sentences. We list one example from each dataset in Table 3.3.

Dataset		**Text**
Noisy	TWR	@realDonaldTrump Does ingesting bleach and shining a bright light in the rectal area really cure #COVID19? Have you tried it? Is that what killed Kim Jong Un? #TrumpIsALaughingStock #TrumpIsALoser
Semi-noisy	OC	*smacks blonde wig on Axel* I think as far as DiZ is concerned, he is very smart but also in certain areas very dumb - - witness the fact that he didn't notice his apprentices were going to turn on him, when some of them (cough Vexen cough) aren't exactly subtle by nature.
	WTP	Not to mention one without any anonymous users TALKING IN CAPITAL LETTERS !!!!!!!!
Non-noisy	MT	Tax data that are not made available for free should not be acquired by the state.
	PE	I believe that education is the single most important factor in the development of a country.
	VG	When's the last time you slipped on the concept of truth?
	WD	The public schools are a bad place to send a kid for a good education anymore.

Table 3.3: One example from each dataset. The underlined text highlights noisy and semi-noisy phrases.

We also highlight the noisy and non-noisy phrases in Twitter, OC, and WTP datasets. Moreover, we present detailed statistics of all seven datasets in Table 3.4.

Dataset		Noisy	Semi-noisy		Non-noisy			
		Twitter (TWR)	Online Comments (OC)	Wiki Talk Pages (WTP)	Micro Text (MT)	Persuasive Essays (PE)	Various Genres (VG)	Web Discourse (WD)
Train	Claim	7354	623	1030	100	1885	495	190
	Non-claim	1055	7387	7174	301	4499	2012	3332
Test	Claim	1296	64	105	12	223	57	14
	Non-claim	189	730	759	36	509	221	221
Overall	Claim	8650	687	1,135	112	2,108	552	204
	Non-claim	1244	8117	7933	337	5008	2233	3553

Table 3.4: Dataset statistics of all seven claim detection datasets used for experiments.

3.4 Proposed Methodology

3.4.1 LESA: Linguistic Encapsulation and Semantic Amalgamation

Traditionally, the narrative on claim detection is built around either syntactic (110; 55) or contextual semantic (109; 111) properties of the text. However, given our purview on integrating both, we propose a combined model, LESA that incorporates exclusively linguistic features leveraged from part-of-speech (POS) tags and dependency tree (DEP) as well as semantic features leveraged from the BERT. Moreover, we employ three pre-trained models representing noisy, semi-noisy, and non-noisy texts for both POS and dependency-based features. The intuition is to leverage the structure-specific linguistic features in a joint framework. Under digital media platforms, we usually have to deal with texts from three kinds of environments: (a) a controlled platform where texts are pre-reviewed (e.g., news, essays, etc.); (b) a free platform where writers of the text have the freedom to express themselves without any restrictions on the length (e.g., blogs, online comments, etc.); and (c) a free platform with restrictions on the text length (e.g., tweets). The texts in the former case are usually free of grammatical and typographical mistakes; thus, they belong to the non-noisy category. On the other hand, the latter case exhibits a significant amount of noise in terms of spelling variations, hashtags, emojis, emoticons, abbreviations, etc., to express the desired information within the permissible limit. Thus, it belongs to the noisy class. The second case combines the two extreme cases, reflecting the semi-noisy category.

As evident from existing research, the domain adaptation from a structured environment to an unstructured one is non-trivial and requires specific processing. Therefore, to ensure generalization, we process each input text from three different viewpoints and intelligently select their contributing features through an attention mechanism. We adopt the above process to extract the POS and DEP-based linguistic features. Subsequently, we fuse the linguistic and semantic features using another attention layer before feeding a multilayer perceptron (MLP) based classifier. The idea is to amalgamate diverse features from different perspectives and leverage them for the final classification. Figure 3.1 depicts a high-level architectural diagram.

Part-of-speech (POS) Module. The POS module consists of an embedding layer followed by a BiLSTM and an attention layer to extract the syntactic formation of the input text. As mentioned earlier, we pre-train the POS module for each category separately and later fine-tune them during the entire model training process. At first, each sequence of tokens $\{x_1, x_2, ..., x_n\}$ is converted to a sequence of corresponding POS tags, such that we get $\{p_1, p_2, \ldots, p_n\}$. However, the foremost problem with modelling this way is the small vocabulary size of 19 owing to a specific number of POS tags. To tackle this, we resort to using k-grams of the sequence, i.e., the sequence of POS tags (with $k = 3$) now becomes $\{(p_0, p_1, p_2), (p_1, p_2, p_3), (p_2, p_3, p_4), ..., (p_{n-2}, p_{n-1}, p_n), (p_{n-1}, p_n, p_{n+1})\}$, where p_0 and p_{n+1} are dummy tags. Subsequently, a skip-gram model (173) is trained on each dataset's POS-transformed corpus,

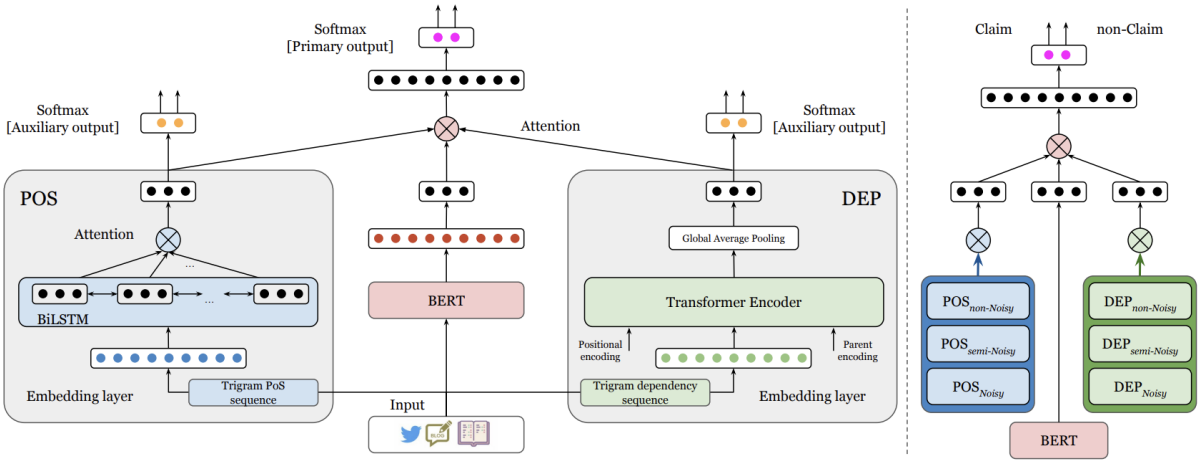


Figure 3.1: Schematic diagram of our proposed LESA model.

which translates to a POS embedding, E_P.

Dependency Tree (DEP) Module. Dependency parsing is the function of abstracting the grammatical assembly of a sequence of tokens $\{x_1, x_2, \ldots, x_n\}$ such that there exists a directed relation (dependency), $d(x_i, x_j)$, between any two tokens x_i and x_j, where x_i is the headword and x_j is modified by the headword. Initially, each sequence is rendered into a combination of the *dependency-tag* arrangement $\{d_1, d_2, \ldots d_n\}$ and a *parent-position* arrangement $\{pp_1, pp_2, \ldots pp_n\}$. Here, each d_j represents a dependency tag, where x_j is modified by x_i, and pp_j is the index of the modifier (headword) x_i.

We then leverage the transformer encoder (174), where traditionally, a position-based signal is added to each token's embedding to help encode the placement of tokens. In our modified version, the token sequence is the *dependency-tag* sequence $d_e = \{d_1, d_2, ..., d_n\}$, wherein a *parent-position* based signal is additionally added to encode the position of the modifier words.

$$d'_e = d_e + [(E_{p_1}, E_{pp_1}), ..., (E_{p_n}, E_{pp_n})] \tag{3.1}$$

where $d'_e \in \mathbf{R}^{d \times n}$ is the modified dependency embedding of a sequence of length n, and E_{p_i} and E_{pp_i} are the encodings for the *token-position* and the *parent-position* (position of token's modifier).

This helps us create a flat representation of a dependency graph. The transformer architecture we employ comprises 5 attention heads with an embedding size of 20. This, however, still poses the problem of a limited vocabulary size of 37, given that there are only a handful of dependency relations. Having accounted for the parent positions already, we decide to use again tri-gram sequences $\{(d_0, d_1, d_2), (d_1, d_2, d_3), (d_2, d_3, d_4), ..., (d_{n-2}, d_{n-1}, d_n), (d_{n-1}, d_n, d_{n+1})\}$ in place of uni-grams.

3.4.2 DESYR: Definition and Syntactic Representation

In LESA (2), we incorporated three modules, including two transformer-based models for claim detection, thus making the architecture heavy on computing. Additionally, within the literature, we observe an inferior performance of the minority class despite efforts to imbalance eradication. Given our outlook on the assimilation of syntax and context and driven by an attempt to prune the aforementioned drawbacks, we propose DESYR. It leverages representation learning by exercising a novel dependency-inspired variant

of the Poincaré embedding (161). Furthermore, we aim to amputate the class-invariant features and obtain superior class-representing features by incorporating the technique for the feature projection, highlighted by Qin et al. (175). DESYR's backbone comprises of two networks in parallel – the regulation-net (*r-net*) and the spotlight-net (*s-net*). The regulation net learns the class-invariant features through a gradient-reversal layer (176). On the other hand, the spotlight net draws representations of the input devoid of the class-invariant features, thereby allowing for better distinction between our binary labels. Both *r-net* and *s-net* incorporate a feature extractor module, called feature-net (*f-net*), in their respective modules – *f-net*$_s$ and *f-net*$_r$. In addition, we also incorporate a definition network (*d-net*) that aims at aligning input texts to definitions of a claim and non-claim to augment class variance further. We leverage *d-net* in our spotlight network module by enabling the feature network (*f-net*$_s$) to learn the segregation between claim and non-claim definitions. We hypothesize that learning such alignment would be exploited by DESYR in successive layers for claim detection. In particular, we fine-tune the hidden representation in *s-net* to extract the essence of segregation between the claim and non-claim definition through *d-net*.

Moreover, DESYR is specifically calibrated towards short informal web-based texts, and we leverage claim and non-claim definitions to engineer our *d-net*. Intricate details of the aforementioned modules are discussed in further sections. We present a high-level architecture for DESYR framework in Figure 3.2.

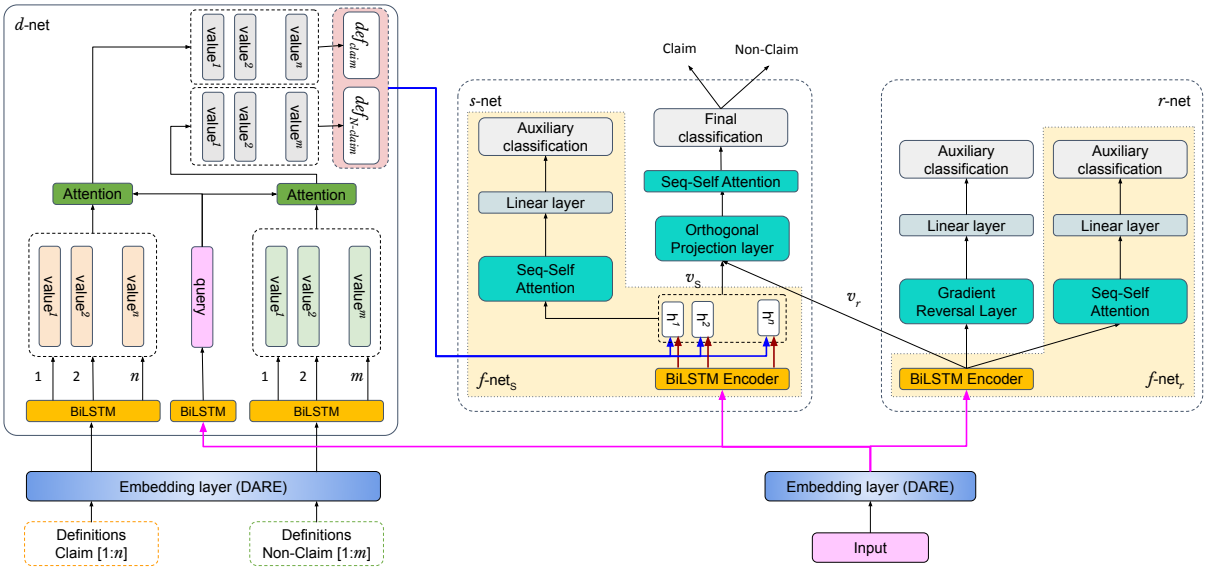


Figure 3.2: A schematic diagram of the DESYR framework for the claim detection.

Dependency-Poincaré Embedding (DARÉ). The Depedency-poincARÉ (DARÉ) embedding is DESYR's variant of the Poincaré embedding (161). Employing DARÉ, we attempt to capture latent linguistic properties in textual dependency. Below, we briefly introduce dependency parsing, followed by a dossier on how the former was imbued into the Poincaré ball. Dependency parsing is a function that maps a sequence of tokens $\{n_1, n_2, ...\}$ to a dependency tree. A dependency parse tree is a directed graph with N nodes and E edges, where each node represents an individual token n_i, and each edge represents the syntactic dependency between n_i and n_j. An edge $n_i \xrightarrow{d_k} n_j$ comprises a parent node directed at the child node with the dependency d_k, where d_k is the nature of the dependency. We employ spaCy[2] for the dependency parsing. An example is shown in Figure 3.3.

As highlighted by Nickel et al. (161), embedding hierarchical graph-based information in Euclidean

[2] https://spacy.io/

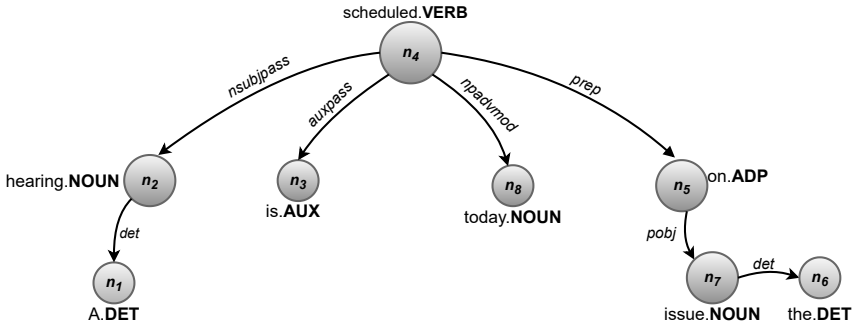


Figure 3.3: Hierarchy (graph) formulation for DARÉ training: dependency-based tree for the sentence '*A hearing is scheduled on the issue today*'.

space can be difficult, owing to the exponential growth of nodes, which brings leaf nodes in different branches close to one another, thereby distorting hierarchies. With the Poincaré ball, the extent from the centre grows exponentially, allowing one to fit an erratic amount of levels in the hyperbolic space. We then optimize our word vectors on this space and optimize the following loss function with negative sampling.

$$L(\theta) = \sum_{(x,y)\in\ \Delta} \log \frac{\exp^{-dist(x,y)}}{\sum_{y'\in\ R(x)} \exp^{-dist(x,y')}} \tag{3.2}$$

where $dist(x, y)$ is the Poincaré distance such that,

$$dist(x, y) = \text{arccosh}(1 + 2\frac{||x - y||^2}{(1 - ||x||^2)(1 - ||y||^2)}) \tag{3.3}$$

where θ is our set of vectors, Δ is the set of all embedded hierarchies-cum-dependencies (in this case, $\forall\ n_i \rightarrow n_j$), and $R(x)$ is the set of random tokens that are not associated with x. Additionally, we attempt the disambiguation of part-of-speech (POS) by formulating our θ as word vectors of the tokens augmented with their POS, such that $x = n_i.\text{POS}_i$. Our loss function L is trained similar to how it would be in Euclidean space. However, the only difference is that we employ Riemann Gradient Descent (177) for the optimization. We utilise Gensim[3] to train DARÉ.

Feature Net (*f-net*). Before discussing *r-net* and *s-net*, we discuss the component *f-net*, which is expected to both the aforementioned networks and serves as the feature extractor. To put this into context, an *f-net* comprises stacked BiLSTMs, whose hidden units are then processed by a sequential self-attention mechanism, as suggested by Zheng et al. (178). Inspired by the Inception (179) architecture, we optimize *f-net* through an auxiliary softmax layer. The intuition behind having these auxiliary outputs is that they would act as implicit assistance against the vanishing gradient problem and make low-level features of the network more accurate. As mentioned before, the transitional outputs (all hidden units) from the BiLSTM are used as inputs for our *r-net* and *s-net*. To emphasize again, each has an individual *f-net* with no shared parameters.

Definition Net (*d-net*). As mentioned earlier, the only distinguishing feature between *f-net*$_s$ and *f-net*$_r$ is that the former comprises of *d-net*. The *d-net* module helps align the inputs to predefined guidelines/definitions that elucidate the characteristics of claims and non-claims. Intuitively, we find its

[3] https://radimrehurek.com/gensim/models/poincare.html

alignment against the aforementioned sets of definitions using an attention-based mechanism for any given input text. This helps us draw divergent associations from the input with respect to claims and non-claims. To put it formally, suppose we have two sets of definitions, $C = [c_1, c_2, \cdots, c_n]$ for claims and $NC = [nc_1, nc_2, \cdots, nc_m]$ for non-claims. The input text t forms our query; this query is then processed (discussed below) against each of the C and NC definitions to get the claim-definition-map and the non-claim-definition-map, respectively.

The *d-net* comprises of *d-net*$_C$ and *d-net*$_{NC}$. These are carbon copies where one is initialized with the definitions of claims and the other with those of non-claims. We process these definitions and input text through a BiLSTM encoder. For each definition, we obtain a *value* vector, and for the input text, we take the last time-step representation as the query vector. Subsequently, for each definition encoding (value^i), we calculate its attention score (180) against the query. Finally, the query-value-attention-score is pooled (global average) over the sequence axis to obtain a 1-dimensional representation. We repeat the process for each pair of query and definition – $\langle$*query*, $c_i\rangle \forall i = \{1, 2, \cdots, n\}$ and $\langle$*query*, $nc_j\rangle \forall j = \{1, 2, \cdots, m\}$. The concatenation of all query-value-attention-scores forms the definition-based representation for t against the respective *definition-set*. We append the representations from *d-net*$_C$ and *d-net*$_{NC}$ behind the BiLSTM output at each time step to enhance the feature learning in *f-net*$_s$ and thereby, explicitly helping the transient BiLSTM features to become more archetypal. The *d-net* is only part of *f-net*$_s$ owing to the fact that *s-net* is our primary classifier whereas *r-net* acts analogously to a feature-selector.

Regulation Net (*r-net*). The regulation network acts to collect class-invariant (common, shared amongst the classes) vector representations from *f-net*$_r$. It is a network trained parallel to the *s-net*. As highlighted in Qin et al. (175) and Ganin et al. (181), we employ a Gradient Reversal Layer (*GRL*) to capture the class-invariant features. In a nutshell, a gradient-reversal layer can be thought of as a pseudo-functional mapping where the forward and backward propagation are respectively defined by two opposed equations as follows:

$$GRL(x) = x \qquad \frac{\delta GRL(x)}{\delta x} = -I \tag{3.4}$$

The transient BiLSTM's output from *f-net*$_r$, that serves as the input to *r-net*, learns the invariant features and is then drifted to *s-net* for computing the orthogonal projection.

Dataset	Twitter		Online Comments (OC)		Web Discourse (WD)		Micro Text (MT)	
	Claim	Non-claim	Claim	Non-claim	Claim	Non-claim	Claim	Non-claim
Train set	7354	1055	623	7387	190	3332	100	301
Test set	1296	189	64	730	14	221	12	36
Overall	8650	1244	687	8117	204	3553	112	337

Table 3.5: Statistics of four datasets used in our experiments.

Spotlight Net (*s-net*). The spotlight network is the prime module of DESYR. We amalgamate the class-invariant features of *r-net*s through an attention orthogonal project layer (*a-OPL*). The orthogonal projection layer aims at drawing the choicest feature representations (175). For convenience, we refer to the feature representation from *f-net*$_s$ and *f-net*$_r$ as $v_s \in \mathbb{R}^{l \times d}$ and $v_r \in \mathbb{R}^{l \times d}$, respectively. We design *s-net* to extract the semantic representation for our input text t and project the same into a non-homogeneous domain space. To accomplish this, we project v_s onto the orthogonal direction of v_r. In theory, the space orthogonal to v_r should be rid of class homogeneity, and projecting v_s onto it should lead to discriminative information being stripped of class-invariant knowledge. Further, we describe the mathematical details of

a-OPL. The projection between two vectors u and v is defined as,

$$proj(u, v) = \frac{u.v}{|v|}\frac{v}{|v|} \tag{3.5}$$

Utilizing the above equation, we project v_s onto v_r while tending to each time-step from their respective LSTM using a TimeDistributed Layer (TDL) such that,

$$v_{(s,i),(r,i)} = \textit{proj}(v_{(s,i)}, v_{(r,i)})$$
$$v_{s,r} = \textit{TDL}_{\forall i}(v_{(s,i),(r,i)})$$

We then find the projection of v_s in the orthogonal direction of $v_{s,r}$ while again tending to each step using a TDL such that,

$$\widetilde{v}_{(s,i)} = \textit{proj}(v_{(s,i)}, (v_{(s,i)} - v_{(s,i),(r,i)}))$$
$$\widetilde{v}_s = \textit{TDL}_{\forall i}(\widetilde{v}_{(s,i)})$$

To further refine this feature representation, we then attend to each time-step in $\widetilde{v}_s$ using sequential self-attention (178) such that,

$$attentive_{\widetilde{v}_s} = \widetilde{v_s}^a = softmax(\sigma(W_a\widetilde{v}_{s,i} + b_a)) \tag{3.6}$$

The aforementioned forms the basis for our *a-OPL*. The attentive vector $\widetilde{v_s}^a$ is then utilized for classification. We train *s-net* and *r-net* parallel to each other and employ sparse categorical focal loss. In a classification setting, with labels y, the loss is defined as,

$$L(y, \hat{p}) = -(1 - \hat{p}_y)^\gamma log(\hat{p}_y) \tag{3.7}$$

where p is a vector that represents the approximate probability distribution across our two classes, and γ is the *focusing* parameter, which, in essence, acts to down-weigh easy-to-classify examples. Higher γ implies high discounting of the easy-to-classify examples.

3.5 Experiments and Results

3.5.1 Evaluating LESA

Experimental Setup. We compute POS embeddings by learning word2vec skip-gram model (173) on the tri-gram. The choice of tri-gram sequence is empirical. In Table 3.6, we report results on bi-gram and four-gram sequences as well. POS sequence. For the skip-gram model, we set *context window* $= 6$, *embedding dimension* $= 20$, and discard the POS sequence with *frequency* ≤ 2. Similarly, we compute dependency embeddings with *dimension* $= 20$ using a transformer encoder with 5 *attention heads*. The outputs of the POS and dependency embedding layers are subsequently fed to a BiLSTM and *GlobalAveragePooling* layers, respectively. Their respective outputs are projected to a 32-dimensional representation for the fusion.

We employ the HuggingFace implementation of BERT for computing the tweet representation. Subsequently, the 768-dimensional BERT embeddings are downsampled to a 32-dimensional representation to maintain consistency among the POS and dependency representations. We employ sparse categorical cross-entropy loss with Adam optimizer and use softmax for the final classification. For evaluation, we adopt macro-F1 (m-$F1$) and claim-F1 (c-$F1$) scores adopted by the existing methods (109; 111).

Models	Noisy		Semi-Noisy				Non-Noisy								Wt Avg	
	Twitter		OC		WTP		MT		PE		VG		WD			
	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$		
POS-only [2-gram]	0.41	0.59	**0.52**	0.14	**0.54**	0.23	0.42	0.00	0.43	0.06	**0.55**	0.30	**0.50**	0.09	0.47	0.47
POS-only [3-gram]	**0.49**	**0.74**	0.51	**0.17**	0.49	**0.24**	0.54	0.32	**0.51**	**0.32**	**0.55**	**0.39**	0.43	**0.10**	**0.50**	**0.62**
POS-only [4-gram]	0.23	0.23	0.50	0.05	0.53	0.14	**0.61**	**0.35**	0.43	0.06	0.53	0.18	0.47	0.00	0.41	0.19

Table 3.6: Macro-F1 and claim-F1 for POS n-gram experiments using LESA.

We perform our experiments in two setups. In the **first in-domain setup**, we train, validate, and test on the same dataset and repeat it for all seven datasets independently. In the **second combined-domain setup**, we combine all datasets and train a unified generic model. Subsequently, we evaluate the trained model on all seven datasets. Furthermore, for each experiment, we ensure a balanced training set by down-sampling the dominant class at $1:1$ ratio. However, note that we use the original test set for a fair comparison against the existing baselines and state-of-the-art models.

Baseline Systems. We employ the following baselines (some of them are state-of-the-art systems for claim detection): ▷ **BERT** (160): We fine-tune BERT for the claim detection task. It resembles our proposed system without PoS and dependency modules. ▷ **XLNet** (182): It is similar to the BERT model, where we fine-tune XLNet for the claim detection; ▷ **Accenture** (113): A RoBERTa-based system that ranked first in the CLEF-2020 claim detection task (112); ▷ **Team Alex** (115): The second-ranked system at CLEF-2020 task that fused tweet meta-data into RoBERTa for the final prediction; ▷ **CheckSquare** (116): An SVM-based system on top of pre-trained BERT embeddings in addition to incorporating POS and dependency tags as external features. ▷ **CrossDomain** (109): Among several variations reported in the paper, their best model incorporates CNN (random initialization) for the detection.

Experimental Results. We report our comparative analysis in Table 3.7. In the non-noisy category, LESA obtains better m-$F1$ scores than three of the four state-of-the-art systems, i.e., it reports 0.77, 0.74, and 0.68 m-$F1$ scores compared to 0.71, 0.71, and 0.61 m-$F1$ scores of the comparative systems on MT, PE, and VG test sets, respectively. On WD, we observe similar m-$F1$ and c-$F1$ scores for both the best baseline and LESA. On the datasets in other categories, we observe comparative m-$F1$ scores; however, none of the baselines are consistent across all dataset – e.g., CrossDomain (109) reports the best m-$F1$ scores on Twitter and OC, but yields (joint) fourth-best performance on WTP. Moreover, LESA yields the best m-$F1$ score across the seven datasets on average with $\geq 1\%$ improvements. On the other hand, we obtain best c-$F1$ scores for five out of seven datasets. In addition, LESA reports overall c-$F1$ of 0.75 with a significant improvement of $\geq 3\%$. Since our work culminates in developing a model that is able to detect claims irrespective of the source and origin of the text, we also analyze the weighted-average scores for each category in Table 3.8. We observe that LESA obtains the best c-$F1$ scores in each category, in addition to the best m-$F1$ score in the non-noisy category as well. For the other two categories, LESA yields comparative performances.

Table 3.11 shows m-$F1$ and c-$F1$ for different variants of LESA. We begin with conventional fine-tuned BERT model and observe the performances on test sets of all seven datasets. On Twitter dataset, BERT architecture yields m-$F1$ score of 0.60 and c-$F1$ score of 0.83. We also report the weighted-average score for the combined test set as 0.58 m-$F1$ and 0.73 c-$F1$, in the last two columns of Table 3.11. Since we hypothesize that the claim identification has a strong association with the structure of the text, we amalgamate POS and dependency (DEP) information in the BERT architecture in step-wise manner. The BERT+POS model reports an increase of 1% m-$F1$ and c-$F1$ scores on the Twitter dataset. We observe similar trends in other datasets and the overall weighted-average score as well. We also perform

Model	Noisy		Semi-Noisy				Non-Noisy								Wt Avg	
	Twitter		OC		WTP		MT		PE		VG		WD			
	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$		
BERT	0.60	0.83	0.52	0.24	0.53	**0.32**	0.70	0.63	0.69	0.64	0.58	0.43	0.48	0.22	0.58	0.73
XLNet	0.59	0.81	0.56	**0.28**	**0.57**	0.29	0.68	**0.69**	0.71	0.64	0.61	**0.44**	**0.52**	**0.25**	0.59	0.72
Accenture	0.49	0.43	0.31	0.12	0.40	0.18	0.36	0.13	0.51	0.36	0.38	0.17	0.37	0.04	0.43	0.38
Team Alex	0.54	0.75	0.54	0.25	0.54	0.30	0.71	0.65	0.71	0.63	0.61	0.43	0.48	0.19	0.57	0.67
Check Square	0.58	0.82	0.51	0.23	0.48	0.28	0.56	0.53	0.68	0.59	0.56	0.38	0.47	0.21	0.56	0.72
CrossDomain	**0.65**	0.82	**0.57**	0.27	0.53	0.28	0.71	0.63	0.66	0.57	0.61	0.43	**0.52**	**0.25**	0.60	0.71
LESA	0.62	**0.85**	0.53	0.24	0.55	**0.32**	**0.77**	**0.69**	**0.74**	**0.66**	**0.68**	0.41	**0.52**	**0.25**	**0.61**	**0.75**

Table 3.7: Macro-F1 and claim-F1 on combined datasets.

Models	Noisy		Semi-Noisy		Non-Noisy	
	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$
BERT	0.60	0.83	0.52	**0.29**	0.63	0.58
XLNet	0.59	0.81	**0.57**	**0.29**	0.65	0.59
Accenture	0.49	0.43	0.36	0.16	0.45	0.30
Team Alex	0.54	0.75	0.54	0.28	0.65	0.57
CheckSquare	0.58	0.82	0.49	0.26	0.61	0.53
CrossDomain	**0.65**	0.82	0.55	0.28	0.63	0.53
LESA	0.62	**0.85**	0.54	**0.29**	**0.69**	**0.60**

Table 3.8: Category-wise weighted-F1 scores.

experiments on other permutations, and their results are listed in Table 3.11. Finally, we combine both POS and DEP modules in the BERT architecture (*aka.* LESA). It obtains improved results for most of the cases, as shown in the last row of Table 3.11. The best result on average stands at 0.61 m-$F1$ and 0.75 c-$F1$ for the proposed LESA model. This supplements our hypothesis to show the significance of combining syntactic and semantic representations for better detection of claims.

In all aforementioned experiments, we use our pre-defined concept of three viewpoints, i.e., noisy, semi-noisy and non-noisy. Therefore, for completeness, we also construct a combined viewpoint for all three components. However, we observe that the obtained results are inferior to the variant with separate viewpoints for each component (c.f. second last row of Table 3.11). Thus, providing disparate attention to datasets based on the noise in their content is demonstrated by a significant increase of $\sim 2\%$ m-$F1$ from combined viewpoint to multiple viewpoints experiment.

3.5.2 Evaluating DESYR

Experimental Setup. In this section, we lay out the backdrop for our experiments and highlight the critical conditions and practices. To compute the DARÉ embedding of size 100, we use the open source Sentiment140 corpus, comprising 1.6 million tweets[4]. We use stacked BiLSTMs with 256 hidden units for both *f-nets*. To emphasize again, there are no shared parameters between *f-net*$_s$ and *f-net*$_r$. Additionally, to encode our query and definitions in the *d-net*, we use a BiLSTM with 64 hidden units. We use pre-furnished definitions proposed in our previous work (2) to harbour our *d-net*; we encipher *d-net*$_C$ with 10 definitions, and *d-net*$_{NC}$ with 8 definitions.

To train our model, we proceed with a vocabulary size of $30k$, and a maximum document length of 50. We use the Adam (183) optimizer and the sparse categorical focal loss (184). We train the model for 100 epochs with a batch size of 32 and exercise early stopping. Since most of the datasets are

[4] https://www.kaggle.com/kazanova/sentiment140

Models	Noisy		Semi-Noisy				Non-Noisy								Wt Avg	
	Twitter		OC		WTP		MT		PE		VG		WD			
	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$		
BERT	0.60	0.83	0.52	**0.24**	0.53	**0.32**	0.70	0.63	0.69	0.64	0.58	**0.43**	0.48	0.22	0.58	0.73
BERT + POS	0.61	0.84	**0.53**	**0.24**	0.54	0.31	0.75	0.69	0.72	0.64	0.59	**0.43**	0.51	0.24	0.60	0.74
BERT + Dependency	0.59	0.82	0.51	0.23	0.52	0.30	**0.79**	**0.73**	0.69	0.62	0.56	0.41	0.48	0.22	0.57	0.72
POS + Dependency	0.45	0.70	0.48	0.19	0.47	0.25	0.57	0.46	0.50	0.45	0.56	0.41	0.44	0.17	0.48	0.61
LESA (Combined-view)	0.61	0.85	0.51	0.23	0.53	0.31	0.77	0.71	0.71	0.64	0.57	0.40	0.48	0.22	0.59	**0.75**
LESA	**0.62**	**0.85**	**0.53**	**0.24**	**0.55**	**0.32**	0.77	0.69	**0.74**	**0.66**	**0.68**	0.41	**0.52**	**0.25**	**0.61**	**0.75**

Table 3.9: Macro F1 (m-$F1$) and claim-F1 (c-$F1$) for ablation studies.

unbalanced towards one class, we adopt the sampling technique to alleviate the issue. We experiment with multiple sampling ratios (c.f. Table 3.11). As a result of fine-tuning, we select a sampling ratio 5:2 (claims: non-claims) for the Twitter dataset based on foraging through the sampling ratio search space. For consistency, we select a sampling ratio of 5:2 (non-claims: claims) for the OC, WD and MT datasets as well. Note that the sampling technique aligns with the previous neural attempts that procured best results on a 1:1 ratio (109; 2). Additionally, to ensure the riddance of seed stochasticity and to incorporate maximal data, we train for 5 random splits and average them using voting. Also, to incorporate a more holistic weighting mechanism, we vote on predictions across three different values (1, 2, 3) of the γ parameter in the focal loss. For evaluating the claim detection systems, we compute **claim-F1** (c-$F1$) and **macro-F1** (c-$F1$) scores.

Another fact of importance is that with DESYR, we design an architecture which is lighter in comparison to its previous state-of-the-art systems, especially LESA. LESA churns approximately $111M$ model parameters, while our proposed model has only $7M$ model parameters. Moreover, the standard XLNet and BERT-based models require a few $100M$ parameters for the same. As is evident, constructing models in adherence with the task statement at hand can be equally as effective if not more, and at times can be accomplished with a fraction of the compute.

Baseline Systems. Due to highly subjective nature of claims, claim detection can more often than not prove to be a demanding task even for humans let alone machines. As with automated, neural (machine-based) claim detection, the problem becomes even more acute in case of web-based short texts, that usually lack soundness in their linguistic edifice. Most of the existing claim detection models (including state-of-the-art systems) struggle with the accurate identification of claims. To assess and contrast the performance of DESYR, we consider the following systems as baselines: ▷ **BERT** (160): It is a bidirectional transformer-inspired auto-encoder LM that we fine-tune for classification. ▷ **XLNet** (182): Similar to BERT, this too is a bidirectional transformer-inspired LM, the only difference being that this is an auto-regressive LM. We fine-tune it for classification. ▷ **Accenture** (113): The authors employed a fine-tuned RoBERTa-based system and nabbed the first position in the CLEF-2020 claim detection shared task (112). ▷ **Team Alex** (115): The system ranked second at CLEF-2020 shared task. The authors proposed the fusion of RoBERTa-based features and Twitter meta-data to detect claims. ▷ **LESA** (2): This is the state-of-the-art claim detection system wherein the authors proposed a system that leverages part-of-speech and dependency-based linguistic encoders in sync with a BERT-based encoder to detect claims.

In addition, we also perform a simple **K-means** clustering-based evaluation. We assign dataset points to one of the two clusters – claim and non-claim, considering their BERT and Poincaré representations separately.

Model	Twitter		Online Comments (OC)		Web Discourse (WD)		Micro Text (MT)		Average	
	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$	m-$F1$	c-$F1$
K-means – BERT	0.46	0.70	0.39	0.09	0.41	0.04	0.49	0.35	0.44	0.30
K-means – DARÉ	0.52	0.83	0.39	0.16	0.47	0.11	0.50	0.21	0.47	0.33
BERT	0.50	0.67	0.50	0.24	0.48	0.23	0.75	0.69	0.56	0.46
XLNet	0.52	0.70	0.45	0.24	0.51	0.12	0.49	0.43	0.49	0.37
Accenture	0.48	0.15	0.44	0.16	0.34	0.11	0.48	0.28	0.44	0.18
Team Alex	**0.70**	0.88	0.46	0.23	0.60	0.34	0.75	0.64	0.63	0.52
LESA	0.67	0.89	0.51	0.26	0.61	0.35	0.80	0.71	0.65	0.55
DESYR	0.67	**0.92**	**0.60**	**0.27**	**0.78**	**0.59**	**0.85**	**0.79**	**0.73**	**0.64**

Table 3.10: Macro-F1 (m-$F1$) and claim-F1 (c-$F1$) of the competing models on different web-based datasets. The first two rows determine the effectiveness of DARÉ embeddings over the standard BERT representation.

Experimental Results. We present our collated results in Table 3.10. To compute the base efficacy of the DARÉ embedding, we compare it against the BERT embeddings (185) without proffering any external supervision. Please note that the Sentence-Transformers package(185) facilitates out-of-box computation of BERT-based dense vector representations for sentences [5]. We employ K-means clustering to segregate the claim with non-claim clusters. We evaluate the test data points in both clusters and report the results in the first two rows of Table 3.10. We observe that DARÉ performs better than BERT on the Twitter dataset by a considerable margin. We also perform K-means clustering on the remaining three datasets as well and observe improvements in most of the cases. The improvements could possibly be attributed to our trained distribution being closer to web-based informal texts despite the training corpus being significantly smaller than BERT's (Wikipedia: $2,500$ million words, Book Corpus: 800 million words).

Furthermore, we evidently observe that DESYR outperforms all the existing baseline systems, including the current state-of-the-art LESA (2). On the Twitter dataset, DESYR obtains the foremost c-$F1$ score in contrast to all other baseline systems – it accounts for a $+3.3\%$ improvement over LESA in c-$F1$. With the OC dataset, we find that all the baseline systems, including DESYR, report low scores for claims. However, DESYR does yield m-$F1$ of 0.60 (with $+9$ points improvement over LESA's performance) – suggesting it performs well for the non-claim class. Out of the four datasets, we observe the highest relative improvement on the WD dataset with 0.59 c-$F1$ and 0.78 m-$F1$, translating to a climb of 68.5% and 27.8% over LESA, respectively. On the MT dataset as well, we observe an increment of 11.26% in c-$F1$ and 6.25% in m-$F1$.

On average, DESYR improves the state-of-the-art performance by 16.36% in c-$F1$ and by 13.3% in m-$F1$. As a general observation, we see how systems grounded in linguistics, such as DESYR and LESA outperform large LMs like BERT and XLNet, which in turn goes to indicate the importance of task-specificity in model adaptation.

Ablation Study. Given that we hinge our model around Twitter, we perform the ablation study by hacking off individual components from DESYR one at a time, and thereafter, we evaluate the same on the Twitter dataset (2). We present the ablation study results in Table 3.11. We report c-$F1$ along with m-$F1$ and weighted-F1 (w-$F1$).

We draw the simplest variant of DESYR by dropping *d-net* and employing sparse categorical cross-entropy instead of focal loss to train our model. Additionally, along with the mentioned withdrawals, we train on a 1:1 sampling instead of the 5:2 sampling. As can be seen in Table 3.11 (rows 4-5), we

[5] https://www.sbert.net/

Ablation	Sampling ratio [Claim:Non-Claim]	c-$F1$	nc-$F1$	m-$F1$	w-$F1$
DESYR	Org	**0.92**	0.38	0.65	0.85
	[1 : 1]	0.89	0.33	0.61	0.81
	[5 : 2]	**0.92**	**0.41**	**0.67**	**0.86**
− {*focal loss, d-net*}	[1 : 1]	0.85	0.36	0.60	0.78
	[5 : 2]	0.90	0.39	0.65	0.83
− {*focal loss*}	[1 : 1]	0.86	0.37	0.61	0.80
	[5 : 2]	0.91	**0.41**	0.66	0.84
− {*DARÉ*} + {*GloVe*}	[1 : 1]	0.90	0.38	0.64	0.83
	[5 : 2]	0.91	0.36	0.63	0.84

Table 3.11: Ablation result for DESYR on the Twitter dataset. The symbol (−) signifies the absence of the respective module. Claim-F1 (c-$F1$), non claim-F1 (nc-$F1$), macro-F1 (m-$F1$) and weighted-F1 (w-$F1$) are reported. The sampling ratio shows our attempt to alleviate the label skewness.

observe an increase of 5 c-$F1$ points and 5 m-$F1$ points simply by reverting back to the 5:2 sampling. We conduct another similar ablation, while retaining *d-net* with categorical cross-entropy. We again observe an increase of 5 c-$F1$ points and 5 m-$F1$ points on a reversion to the 5:2 sampling. It is worth mentioning that the previous benchmark on the Twitter dataset applied a 1:1 sampling across all their experiments (2). We, however, espy worse results on using the same. We additionally discern that in two distinct cases, the addition of *d-net* results in a boost – we see a boost of 1 c-$F1$ point and 1 m-$F1$ point when comparing ablations that differ in their residence of *d-net*.

Within DESYR we experiment with different sampling ratios, as is clearly evident we outperform the original sampling and the 1:1 sampling. We observe that with the original sampling, we see competitive results on c-$F1$. However, the same doesn't hold true for nc-$F1$; thereby, the 5:2 sampling is a better fit, given that in addition to better results, it also helps keep the skew subservient.

Furthermore, we detect that the variant of DESYR that comes without focal loss performs worse compared to DESYR; the c-$F1$ and the m-$F1$ values drop by 1 point each. Another interesting ablation is where we choose to initialize DESYR with GloVe (186) in place of DARÉ. We see that DESYR in its native state outperforms GloVe initialization by 1 c-$F1$ point and 4 m-$F1$ points (rows 8-9).

3.6 Error Analysis

To qualitatively appraise the performance of DESYR and LESA, we perform error analysis in this section. Table 3.12 highlights a few randomly sampled instances from the Twitter, OC, WD, and MT datasets, along with their gold and output labels as predicted by DESYR and LESA. In some cases, both DESYR and LESA fail to identify claims; however, in most of the cases, DESYR performed well. We additionally report instances misclassified by DESYR and/or LESA.

As underlined previously, on the Twitter dataset, we observe that DESYR obtains better classification results in comparison to all other baseline systems. In examples t_1 and t_3, we see that LESA misclassifies both the examples. In comparison, DESYR identifies the assertions and classifies them correctly. With mistakes being inevitable, we observe that most of the misclassified samples are non-claim ($t_2, t_4, t_5, t_9, t_{12}$). A potential reason could be the skewed nature of the Twitter dataset (2) – the dataset is imbalanced with a hulking predisposition towards claims (the number of claims is greater even after the 5:2 sampling

Dataset		Example	Gold	Prediction	
				DESYR	LESA
TWR	t_1	*RT @PirateAtLaw: No no no. Corona beer is the cure not the disease https://t.co/fnba2fr2m2*	claim	claim	non-claim
	t_2	*@zlj517 Vaccine development is urgent. Please let me know when an effective vaccine is completed in the world. I go get it and bring it to China. Wait for early development. #pichperfact*	non-claim	claim	claim
	t_3	*#China has just discovered #vaccine against #coronavirus. Thank you china*	claim	claim	non-claim
OC	t_4	*I usually walk around and attempt to read all the names on the old gravestones.*	non-claim	claim	non-claim
	t_5	*Casey Siemezko is Charlie in Young Guns...which bothers me because I keep wantin to see him without his beard/mustache and this is the closest thing to it, but he has his 3D glasses...cant win.*	non-claim	claim	claim
	t_6	*My only big smile moment was Kurt's dad knowing Kurt was g*y since he was a t*t, lovely little scene, playing against the dad's macho image.*	claim	claim	non-claim
WD	t_7	*I don't know about the rest of Virginia, but in Northern Virginia we have excellent public schools.*	claim	non-claim	non-claim
	t_8	*I am a 12-year-old 6th grader turning 13 in May, so I know what it's like.*	non-claim	non-claim	claim
	t_9	*Here is my experience and I hope it will help you with your decision: In preschool, I had some issues, just like your son.*	non-claim	claim	claim
MT	t_{10}	*That's why Germany should not introduce capital punishment!*	claim	non-claim	claim
	t_{11}	*Alternative treatments should be subsidized in the same way as conventional treatments, since both methods can lead to the prevention, mitigation or cure of an illness.*	claim	claim	non-claim
	t_{12}	*Besides it should be in the interest of the health insurers to recognize alternative medicine as treatment, since there is a chance of recovery.*	non-claim	claim	claim

Table 3.12: Error analysis of the outputs on various datasets. For comparison, we also show the predictions of LESA (2).

ratio). The presence of the phrase, *'effective vaccination is completed'* in example t_2, drives the system to assert it and incorrectly predict it as a claim. Example t_7 has a dearth of context and could possibly have been a part of a bigger phrase, and on top of that, it is not a statement that would severely affect public opinion, which is presumably why DESYR misclassifies it. The latter argument was based on the guidelines proposed in our previous work (2). Clearly, the gold label for t_9 indicates that it is not a claim; however, within the realm of possibility, it does seem that the person indirectly claims about having issues. It is possible that DESYR misclassifies in this case owing to this dormant pattern. Sentence t_{12} emphasizes a chance of recovery with an alternative medicine, i.e., they report a medical fact to be true, albeit with chance. There exists a possibility that DESYR and LESA interpret a chance of recovery using medication as a claim and, therefore, misclassify it.

Human Evaluation. Over time, OSM sites have emerged as the hub for short, unstructured pieces of informal text, where the amount of slang and incoherence in writing is generally more significant than other online platforms. Considering the prime focus of DESYR, we tend to evaluate our DESYR's performance against real-world data to detect claims on the web. We collect 50 random samples from Twitter and predict their labels (claim or non-claim) using DESYR. Note that we collect these examples

in the wild. Subsequently, we present the predictions to three human evaluators[6] and ask them to verify the labels following our claim annotation guidelines. Finally, a majority vote was used to get the final gold-label for these 50 tweets. Among the three evaluators, we obtain an inter-annotator score (*Fleiss kappa*) of 0.76.

We present some of the instances in Table 3.13. As expected, most of the claims were correctly labelled by DESYR. Out of 50 samples, DESYR classifies 32 samples correctly. We observe m-$F1$ score of 0.60, while c-$F1$ score of 0.73. Our model also marked some false positives at the expense of its precision. This is not ideal but a better scenario than biasing towards false negatives where claims are wrongly classified as non-claims. Examples t_1 and t_3 are claims exhibiting some statistics. Our model DESYR rightly captured these values and classified both tweets as claims. However, DESYR was unable to learn the importance of numerical features in t_2, especially when they occurred several times within the text, where it failed to interpret the significance of the numerical features and ended up mislabelling the tweet. Following the claim guidelines, negating a false claim also accounts for a claim. In example t_6, the user tries to negate a claim and further claims to document the evidence. DESYR comprehends the assertions and rightly labels them as claims. Through example t_8, the user imparts his/her personal beliefs that might or might not affect the public; thus, DESYR possibly interpreted it as a personal opinion and marked it as non-claim. Examples t_7 and t_{10} encompass strong claim phrases *'has died in India'* and *'paying them nothing'*, respectively. Clearly, these two examples make strong social assertions that would interest a larger audience, which is possibly why it is labelled as a claim. Finally, in example t_9, the user commends doctors' determination during the global pandemic and expresses their experiences. This example would not fall under the claim category; however, DESYR mis-classifies it as a non-claim, possibly due to the presence of the phrase '*curing COVID-19*'.

Our observations from the in-the-wild evaluation suggest that DESYR can efficiently and accurately assign labels to unseen tweets. Moreover, we do not follow any unexpected behaviour of DESYR. Thus furnishing us with empirical shreds of evidence that DESYR can be used for claim detection in informal texts.

3.7 Summary

Through this detailed and systematic study, we tend to make notable contributions which could significantly advance the field of claim detection. We applied Poincaré embeddings to an NLP task, which showed promising results for the claim detection task. The proposed model, DESYR, determined the existence of claims in the online text by aligning the query to encoded definitions and projecting them into a purer space. We evaluated DESYR across four web-based datasets, which comprise short informal texts, and observed up-to-par results. The comparative investigation highlighted the more nuanced performance of our model compared to various existing systems. Experiments demonstrated the superiority of our model with $\geq$3% claim-F1 improvements over the existing state-of-the-art claim detection system, LESA. Additionally, we exhibited every individual component's performance and significance in our model through an exhaustive ablation study. Finally, we showed the robustness of DESYR through a qualitative human evaluation in the wild on 50 random samples.

[6]They are linguistics by profession, and their age ranges between 24-45 years.

	Example	DESYR	Human
t_1	*#CovidVaccine | Assam vaccinated only 15% residents, young struggle to book slots. NDTV's Ratnadip Choudhury reports*	claim	claim
t_2	*We went from May 1 to May 17 in 2 day*	claim	non-claim
t_3	*Nearly 51 lakh COVID-19 vaccine doses will be received by states/UTs within next 3 days: Health ministry.*	claim	claim
t_4	*I STAND WITH HUMANITY #IndiaStandwithPalestine*	non-claim	non-claim
t_5	*Par for the course. As if we'd trust an internal review. We are asking for #COVIDPublicEnquiryNow*	non-claim	claim
t_6	*They claim they are no longer asking for Aadhaar/mobile to give food at Indira canteens. But our Youth Congress team found otherwise. And documented it. We will continue to expose this government's lies.*	claim	claim
t_7	*A third Australian has died in India from COVID-19. The family members of 11,000 stranded Aussies are pleading for the government to bring them home. #9News*	claim	claim
t_8	*My heartfelt gratitude to the men in uniform who did not deter from putting their lives in danger to save the lives of our citizens under extreme conditions.*	non-claim	non-claim
t_9	*Interacted with doctors across India. They shared insightful inputs based on their own experiences of curing COVID-19. The determination of our doctors during these times is remarkable!*	claim	non-claim
t_{10}	*Abolish unpaid internships. There is absolutely no valid reason that justifies why you're having students work 40 hours/week and paying them nothing.*	claim	claim

Table 3.13: Human evaluation on random tweets (non-dataset examples). DESYR reports macro-F1 of 0.60 and claim-F1 of 0.73 on 50 random samples.

Part 2

Simplifying Complex Claims

4. EXTRACTIVE CLAIM SIMPLIFICATION

The widespread diffusion of medical and political claims in the wake of COVID-19 has led to a voluminous rise in misinformation and fake news. The current vogue is to employ manual fact-checkers to efficiently classify and verify such data to combat this avalanche of claim-ridden misinformation. However, the rate of information dissemination is such that it vastly outpaces the fact-checkers' strength. Therefore, to aid manual fact-checkers in eliminating the extraneous content, it becomes imperative to automatically identify and extract the snippets of claim-worthy (*mis*)*information* present in a post. In this work, we introduce the novel task of *Claim Span Identification* (*CSI*). We propose CURT, a large-scale Twitter corpus with token-level claim spans on more than $7.5k$ tweets. Furthermore, along with the standard token classification baselines, we benchmark our dataset with DABERTa, an adapter-based variation of RoBERTa. The experimental results attest that DABERTa outperforms the baseline systems across several evaluation metrics, improving by about 1.5 points. We also report detailed error analyses to validate the model's performance, along with the ablation studies. Lastly, we release our comprehensive span annotation guidelines for public use.

4.1 Introduction

The swift acceleration of Online Social Media (OSM) platforms has led to tremendous democratized content creation and information exchange. Consequently, these platforms serve as ideal breeding grounds for malicious rumourmongers and talebearers, leading to a significant rise in misinformation. Such misinformation manifests in many ways, including bogus claims, fabricated information, and rumours. The massive COVID-19 *'Infodemic'* is one such malignant byproduct that led to the rampant spread of political and social calumny (36; 187; 188; 189), accompanied by counterfeit pharmaceutical claims (190). Therefore, finding such claim-ridden posts on OSM platforms, investigating their plausibility, and differentiating the credible claims from the apocryphal ones has risen to be a pertinent research problem in Argument Mining.

A claim is the key component of any argument (53; 191). Consider the second tweet, *'We don't have evidence...'*, as given in Figure 4.1. For the task of claim identification at the coarse level, the entire tweet will be marked as a claim. However, on closer inspection, we find that the text fragments of *'our wine keeps you from getting #COVID19'* and *'Better alternative to #DisinfectantInjection'* represent the finer argumentative units of claim and form the set of evidence, based on which this tweet is considered a claim. Segregating such argumentative units of misinformed claims from their benign counterparts fosters many benefits. To begin with, it partitions the otherwise independent claims in a single post, enabling us to retrieve a larger number of claims. Secondly, it acts as a precursor to the downstream tasks of claim check-worthiness and claim verification. Thirdly, it will also bring in the angle of *explainability* in

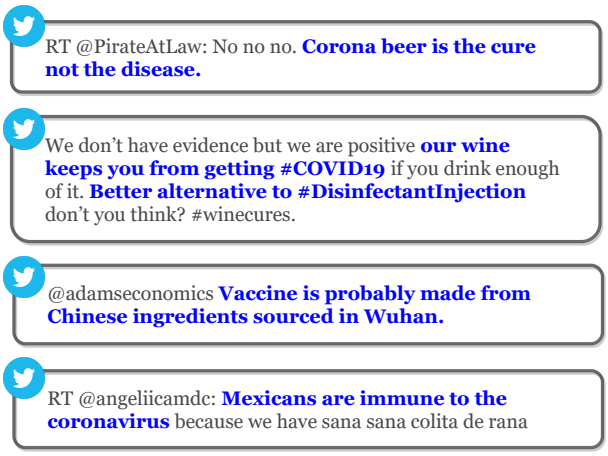


Figure 4.1: Examples of claim tweets and their ground truth claim spans highlighted in boldface text (blue).

coarse-grained claim identification. Finally, it will serve the manual fact-checkers and hoax-debunkers[1,2] to conveniently strain out the unnecessary shreds of text from further processing. Though the recent literature reflects extensive work done in claim detection (191; 111; 2), limited forays have been made in claim span identification i.e., recognizing the argumentative components of a claim (56). In the recent past, commendable work has been done on span-level argument unit recognition pertaining to other computational counterparts under the umbrella of AM, such as hate speech (119), toxic language (120) and propaganda techniques (121). Such a study, however, has eluded the realm of claims, owing to the lack of quality annotated datasets. This heralds a specialized corpus creation on claim span identification.

Task Motivation. As stated, we hypothesise that claim span identification would aid fact-checkers to quickly segregate claim-ridden content from the rest of the post. Moreover, we suppose that it will be a promising precursor for claim verification and fact-checking, facilitating better retrieval of relevant evidences. We back our hypothesis with a small experiment of evidence-based document retrieval. We collect 50 random samples from CURT, along with their corresponding ground-truth claim spans. Further, for both the tweets and the claim spans, we extract top-k relevant articles from a knowledge-base leveraging the traditional retrieval system, BM25 (192). We use the recently released publicly available CORD19 corpus (193) to retrieve factual documents. Finally, we present retrieved documents to three evaluators and ask them to mark whether or not the retrieved shards of evidence are relevant to the given input tweet/span from our dataset. All three annotators label each text-evidence pair independently. Eventually, to obtain the final relevancy score, majority voting is employed. We obtain a high inter-annotator score (*Fleiss Kappa*) of 0.63 and 0.67 for tweets and spans, respectively. We compare the performance of tweet-based and span-based retrievals in terms of precision (P) and normalized Discounted Cumulative Gain (nDCG) scores and report them in Table 4.1. For comparison, we consider two different top-k settings (k=3 and 5). We begin by examining the retrieval performance using P@k, which measures the fraction of relevant documents extracted in the top-k set. Span-based document retrieval consistently improves precision scores when compared to tweets. For nDCG@5, we discover that span-based retrieval outperforms tweet-based retrieval by more than 3%. When we limit the retrieval depth to 3, we see a similar pattern. This, in turn, demonstrates that entire posts contain much extraneous information, frequently impeding the performance of evidence retrieval systems that are a prerequisite for both automated and manual fact-checking. In summary, we reinforce that our hypothesis positively stands true, as span-based document

[1] https://www.snopes.com/
[2] https://www.politifact.com/

retrieval results in a better score for precision as well as nDCG. This attests to the task's feasibility and importance in the realm of claims.

Input	P@5	P@3	*n*DCG@5	*n*DCG@3
Tweets	0.3922	0.2745	0.2733	0.2280
Spans	**0.4407**	**0.3390**	**0.3038**	**0.2521**

Table 4.1: nDCG@k and P@k scores for tweet and spans using BM25 retrieval system and CORD19 dataset.

To this end, we propose **CURT** (**C**laim **U**nit **R**ecognition in **T**weets), a large-scale, claim span annotated Twitter corpus. We also present several baseline models for solving claim span identification as a token classification task and evaluate them on CURT. Furthermore, we introduce *claim descriptions*, which are generic prompts aimed to assist the model in focusing on the most significant regions of the input text using explicit instructions on what to designate as a '*claim*'. They are elucidated later in detail. Finally, we benchmark our dataset with **DABERTa** (**D**escription **A**ware **R**oBERTa), a plug-and-play adapter-based variant of RoBERTa (114), endeavoured to infuse the Pre-trained Language Model (PLM) with the description information. Empirical results attest that DABERTa outperforms the conventional baselines and generic PLMs for our task consistently across various metrics. Through this work, we make the following tangible contributions:

- **Formulation of a novel problem statement**: We propose the novel task of *Claim Span Identification* that aims to identify argument units of claims in the given text.
- **Claim span identification dataset and extensive annotation guidelines**: We posit a large-scale Twitter dataset, the first of its kind, with $7.5k$ claim span annotated tweets, to placate the absence of the annotated dataset for claim span identification. Additionally, we develop comprehensive annotation guidelines for the same.
- **Claim span identification system**: We propose a robust claim span identification framework based on *Compositional De-Attention* (*CoDA*) and *Interactive Gating Mechanism* (*IGM*).
- **Extensive evaluation and analysis**: We evaluate our model against different baselines to confirm sizable improvements over them. We also report thorough qualitative and quantitative analysis along with the ablation studies.

4.2 Related Work

Claims on Social Media. The prevailing research on claims could be cleft into three categories – claim detection (54; 111; 2), claim check-worthiness (138; 59), and claim verification (60; 194; 195). (196) pioneered the efforts in claim detection by introducing the *AAWD* corpus. Subsequent studies largely relied on using linguistically motivated features such as sentiment, syntax, context-free grammars, and parse-trees (108; 54; 55). Recent works in claim detection have engendered the use of large language models (LMs). Chakrabarty et al. (111) re-enforced the power of fine-tuning, as their ULMFiT LM, fine-tuned on a large Reddit corpus of about 5M opinionated claims, showed notable improvements in claim detection benchmark. Gupta et al. (2) proposed a generalized claim detection model for detecting claims independent of their source. They handled structured and unstructured data in conjunction by training a blend of linguistic encoders (POS and dependency trees) and a contextual encoder (BERT) to exploit the input text's semantics and syntax. As LMs account for significant computational overheads, Sundriyal et al. (197) addressed this quandary and proposed a lighter framework that attempted to fabricate

discernible feature spaces. The *CheckThat! Lab's CLEF*-2020 shared task has garnered the attention of several researchers (198). Williams et al. (113) won the task by fine-tuning the RoBERTa (114) accentuated by mean pooling and dropout. Nikolov et al. (115) ranked second with their out-of-the-box RoBERTa vectors supplemented with Twitter meta-data.

Span Identification. Though coarse-grained claim identification reflects a substantial pool of relevant work, fine-grained claim span identification is humbly explored. Zaidan et al. (117) introduced the concept of rationales, highlighting text segments that supported their label's judgment. They reported a significant boost in performance when they included these rationales in the training process for sentiment classification of movie reviews. Trautmann et al. (118) released *AURC*-8 dataset with token-level span annotations for the argumentative components of stance along with their corresponding label. Mathew et al. (119) proposed a quality corpus for explainable hate identification with token-level annotations. The *SemEval* community has initiated fine-grained span identification concerning other domains of argument mining such as toxic comments (120) and propaganda techniques (121). These shared tasks amassed many solutions constituting transformers (122), convolutional neural networks (123), data augmentation techniques (124; 125), and ensemble frameworks (126; 127). (56) resembled the closest study to ours, wherein they compiled a corpus of around $1.2k$ biomedical tweets with claim phrases.

In summary, existing literature on claims concentrates entirely on sentence-level claim identification and does not investigate eliciting fine-grained claim spans. In this work, we endeavour to move from coarse-grained claim detection to fine-grained claim span identification. We consolidate a large manually annotated Twitter dataset for the claim span identification task and benchmark it with various baselines and a dedicated description-based model.

4.3 Dataset

Over the past few years, several claim detection datasets have been released (108; 111). However, none of these corpora come with claim-based rationales that quantify a post as a claim. To bridge this gap, we propose **CURT** (**C**laim **U**nit **R**ecognition in **T**weets), a large-scale Twitter corpus with token-level claim span annotations.

4.3.1 Collection

We annotate our claim detection Twitter dataset for the claim spans. However, the guidelines presented certain reservations that do not explicitly account for benedictions, proverbs, warnings, advice, predictions, and indirect questions. As a result, tweets such as *'Dear God, Please put an end to the Coronavirus. Amen'* and *'@FLOTUS Melania, do you approve of ingesting bleach and shining a bright light in the rectal area as a quick cure for #COVID19? #BeBest'* have been mislabeled claims. This prompted us to extend the existing guidelines and introduce a more exclusive and nuanced set of definitions based on claim span identification.

4.3.2 Annotation

Annotation Guidelines. While different frameworks and models of argumentation range in intricacy and claim conceptualization, the claim element is colloquially perceived as a principal component of an

argument. Following Stab and Gurevych (199), we define the claim as '*the argumentative component in which the speaker or writer conveys the central, contentious conclusion of their argument*.' Aharoni et al. (200) proposed a framework in which an argument is often divided into claims and premises. The premise, another crucial argument component, encompasses all shreds of evidence obliged to either corroborate or refute the claim. We confine our corpus to claim components only. However, claims and premises are usually indistinguishable and frequently blend together. As a result, distinguishing them can be challenging, especially when authors use claim-like statements as a premise.

Due to the highly subjective nature of claims, it is imperative to devise structured annotation guidelines to annotate a new dataset for the claim span identification task. Therefore, we established an initial set of annotation guidelines after rigorous analysis and discussion. To acclimate better with the dataset, we progressed through iterations of improvements. In every iteration, 100 random tweets were annotated by three annotators[3] following the initial set of annotation guidelines. The annotators resolved the ambiguous cases mutually. We addressed the unsettled tweets that necessitated clarifications in the annotation guidelines in successive iterations. We reconsidered all prior annotations for every change in the guideline to ensure that the annotations emulated the most advanced version of the annotation guidelines. The final sprint of pilot annotation included annotating another set of 100 randomly chosen tweets with the final guidelines. Following Trautmann et al. (118), we calculated the inter-annotator agreement using the Krippendorff agreement measure (201). We computed the mean pairwise value per post, where each token can be classified into a claim or a non-claim. We obtained a more than satisfactory agreement score of 0.87. Finally, the entire Twitter dataset was annotated by the same annotators who carried out the prefatory pilot annotations. We present the general guidelines as follows:

- A claim is a statement that says you strongly believe something is true. The action of showing, using, or stating something strongly.
- We use tweets annotated with a binary label using LESA guidelines (2), which indicates whether a tweet is a claim.
- The claim span is that part of a sentence that contains the semantic representation of the claim. Example: @realDonaldTrump A lot of people are saying cocaine cures COVID-19. Claim span: cocaine cures COVID-19.
- Since our primary goal is to tackle misinformation in OSM, we focus on claims with some social impact.

Guidelines and Examples.

- In the case of facts, we annotate the fact/span that may not be known by everyone, such as scientific facts or legal (law) facts, and doesn't involve commonsense. However, we do not include universal facts in the claim span.
  Example 1: "Water is colorless" is a universally known fact and should not be marked as a claim span.
  Example 2: "Virus always mutate" is a scientific fact that may not be known by everyone. Hence we will annotate this fact as a claim.
- An assertion about future eventualities/predictions will not be included in the claim span. Prediction is an extrapolation based on an assertion associated with a confidence level that can never be greater than or equal to 100% Example: @realDonaldTrump Uh no, actually, The virus will never go away. Scientists will develop a vaccine for it that should be ready by next June, which will allow nearly everyone to be immune to #CoronaVirus This really isn't hard to understand even for a very stable genius.

[3]They are linguistic experts, and their age ranges between 20-35 years.

- A proverb is a simple, concrete, traditional saying that expresses a perceived truth based on common sense or experience which contains wisdom, truth, morals, and traditional views in a metaphorical, fixed, and memorizable form. The proverbs are not facts. The elements of proverbs should be annotated as claim span.
  Example: "Prevention is better than cure" is not a claim.

- If a claim contains statistics or dates, they should be included in the span. But not all numbers are important.
  Example 1: "@FernandoSVZLA @AP So far 50 people outside China have it with no deaths. If China was hiding information and it was more lethal, we would see that fairly quickly". Here the claim span is [50 people outside China have it with no deaths]
  Example 2: "57 round trip to LA thanks coronavirus". The number is not important here.

- In case there are multiple conclusive independent claims in one tweet, we annotate each one of them separately.
  Example: "5 million left Wuhan before the lockdown. If they were really interested in knowing, they'd be testing at least 1 in 100 cases of all viral pneumonia. They're limiting who's being tested so they aren't accused of lying. Oh, and they might be asked to actually do something." Claim span would consist of: "5 million left Wuhan before the lockdown" and "They're limiting who's being tested so they aren't accused of lying"

- Tweets that negate a possibly false claim are also considered to be claims.
  Example: "disinfectants are not a cure for coronavirus".

- Tweets 'reporting' something to be true or an instance to have happened or will happen are claims.

- In cases of claims made in the form of a conditional sentence, the premise/context would be included in the span.
  Examples: if you've been in the McDonald's play place you're immune to the coronavirus.

- For claims containing humor/sarcasm, only the humorous phrase will be considered as a claim span if it has some social impact. For satire, the complete sentence will be considered.
  Example: @TheRickWilson Drinking bleach and/or injecting Disinfectant will cure COVID19. And cancer, heart disease, OCD, schizophrenia and AIDS. And life. #Covid19 #COVID Claim: Drinking bleach and/or injecting Disinfectant will cure COVID19. And cancer, heart disease, OCD, schizophrenia and AIDS. And life.

- Personal experience will only be part of the claim phrase if they are opinions with societal impacts/implications.
  Example: Story about how #HydroxyChloroquine likely help people recover from #Coronavirus. IMO, it was never touted as the cure but as option for treatment doctors should consider and it appears to work in some cases....39 in one place. https://t.co/2hhi6aSVrY
  Claim: [it was never touted as the cure but as option for treatment doctors should consider and it appears to work in some cases....39 in one place.]

- A claim can be a sub-part of a question, only if it is not a direct question.
  Example: @FLOTUS Melania, do you approve of ingesting bleach and shining a bright light in the rectal area as a quick cure for #COVID19 ? #BeBest"
  Claim: [ingesting bleach and shining a bright light in the rectal area as a quick cure for #COVID19]

- Ground/Reasoning to justify a claim will not be a part of the claim phrase.
  Example: Covid-19 vaccine development and deployment in China, when available, will be made a global public good, which will be China's contribution to ensuring vaccine accessibility and affordability

in developing countries
Claim phrase: [Covid-19 vaccine development and deployment in China, when available, will be made a global public good]

- Mocking/attacking a group or individual is not a part of the claim phrase.
  Example: Because #coronavirus has tremendous chances of getting cured but your anti-national agenda is worse than death
  Claim: [coronavirus has tremendous chances of getting cured]
- Claim phrases do not include the predicate part that does not contribute to it being a claim.
  Example: I firmly believe that [if they found a way to bottle the @andersoncooper giggle, it would cure the coronavirus.]

4.3.3 Preprocessing

We employ *NLTK*[4] to tokenize the tweets. Each token in the tweet is *BIO (Begin-Inside-Outside)* encoded to generate the labels (202). Tag *'B'* indicates that the token is at the start of a span, tag *'I'* indicates that the token is within the span, while tag *'O'* denotes that the token is outside the span. As RoBERTa tokenizes each word into subwords (114), each subword is given the BIO tag as per its parent word. We eliminate tokens made of non-ASCII and special characters and remove the URLs provided in the tweets. Finally, we split hashtag terms by underscore delimiter and over non-consecutive uppercase characters. For instance, *#WuhanLab* splits into *'Wuhan'* and *'Lab'*.

4.3.4 Statistics and Analysis

We segment CURT into three partitions – training set, validation set, and test set, in the split of 80:10:10. Dataset-related statistics are given in Table 4.2. One important point to note here is that while a claim tweet is typically 27 tokens long, a claim span is only around 10 tokens long. This implies that the claim-ridden tweets have a lot of extraneous information. Arguments can also perhaps comprise several claims that may or may not be related to each other. Around 19% of the claim tweets in our dataset contain multiple claim spans. As a result, we obtain 9458 claim spans from 7555 tweets. We observe that the majority of the tweets contain single claims. Out of 7555 tweets, 6039 include a single claim, demonstrating that most tweets contemplate a single assertion at a time.

Dataset	Train	Test	Validation
Total no. of claims	6044	755	756
Avg. length of tweets	27.40	26.93	27.29
Avg. length of spans	10.90	10.97	10.71
No. of span per tweet	1.25	1.20	1.27
No. of single span tweets	4817	629	593
No. of multiple span tweets	1201	121	161

Table 4.2: Dataset statistics of CURT. All the lengths are in tokens.

4.4 Methodology

In this section, we outline DABERTa and its intricacies. The main aim is to seamlessly coalesce critical domain-specific information into Pre-trained Language Models (PLM). To this end, we introduce *Descrip-*

[4] https://www.nltk.org/

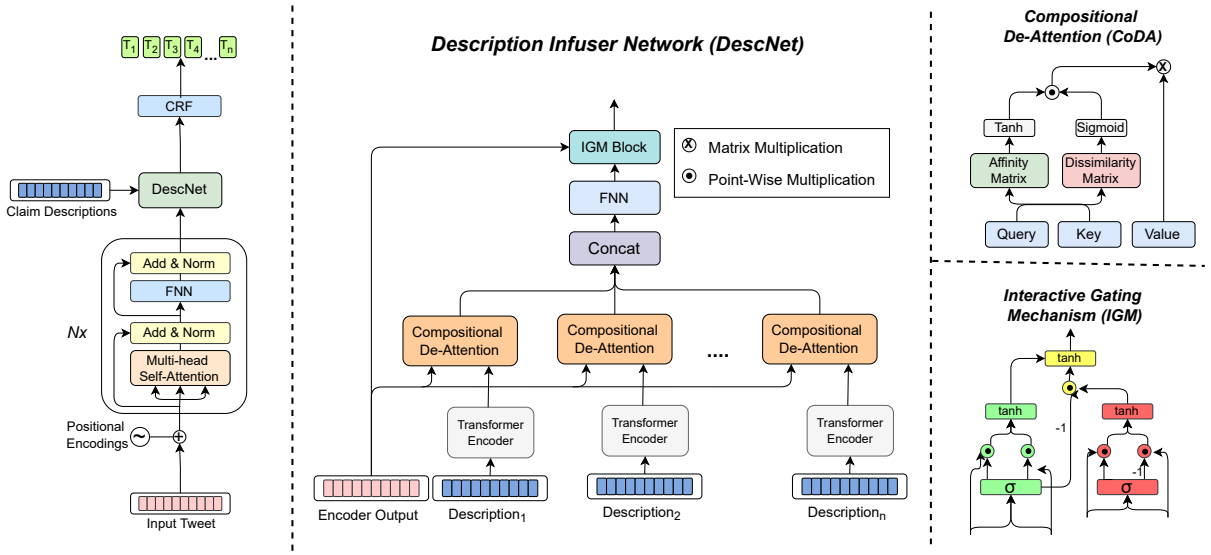


Figure 4.2: A schematic diagram of DABERTa for the claim span identification. $\odot$ represents point-wise multiplication, and $\otimes$ represents matrix multiplication.

tion Infuser Network (*DescNet*), a plug-and-play adapter module that conditions the LM representations with respect to the handcrafted descriptions. The underlying principle behind this theoretical formalization is to link a claim span to a claim description to guide the model on what to focus on explicitly. As shown in Figure 4.2, DescNet houses two sub-components, namely, *Compositional De-Attention block* (*CoDA*) and *Interactive Gating Mechanism* (*IGM*). The particulars of each component are delineated in the following sections.

Claim Description	**Example**
Texts in the tweet mentioning statistics, dates or numbers	Another case for more testing for #coronavirus! *Blood tests show 14% of people are now immune to covid-19 in one town in Germany* https://t.co/MVOq3nc4hn
Texts in the tweet that negate a possibly false claim	No! *#Bleach won't cure #COVID19. Disinfectants can't kill the #coronavirus in your body*. In fact, they will hurt you. If you or someone you know has been exposed to bleach, call Poison Control for help (1-800-222-1***). https://t.co/DtIfi77vLz https://t.co/9MxSFoVM0L
Texts in the tweet made in sarcasm or humour	@username I think the *cure to coronavirus is a 6 pack of corona* only. yeah
Texts in the tweet containing opinions that have societal implications	@username @username I think *it's a bio weapon made by China* so I'm not surprised it has a lot of carriers.
Texts in the tweet in the form of conditional statement	*if you smoke weed you are immune to coronavirus*
Texts in the tweet containing a quote from someone	The president said *injecting disinfectant into the body can cure the virus*. What in the holy hell? And @Lysol issued a statement that people should not ingest Lysol. WTF? #Covid_19 #lysol #DontDrinkLysol

Table 4.3: Examples of handcrafted claim descriptions, along with some aligning examples. Claim spans are highlighted in italics.

4.4.1 Claim Descriptions

Before delving into *CoDA* and *IGM*, we first examine Claim Descriptions, the cornerstone of the proposed model. Claim Descriptions are handcrafted templates that guide the model on where to concentrate its focus. Including a claim description encourages the model to focus on the essential phrases in the input tweet, which may be thought of as guided attention that leads to increased performance. We judiciously

curated our claim descriptions in accordance with the annotation guidelines for claims and non-claims. In Table 4.3, we list some of the claim descriptions and the most aligned claims. It is noteworthy that a claim can align with more than one claim description as well.

4.4.2 PLMs for Token Classification

To begin with the details of the proposed framework, DABERTa, we present the working of PLMs for the token classification task. PLMs such as BERT (160), DistilBERT (203), and RoBERTa (114) are widely used for various downstream NLP tasks owning to their strong contextual language representation capabilities and fine-tuning ease. As the input to these PLMs, each i^{th} input text is first tokenized into a sequence of sub-word embeddings $X_i \in \mathbb{R}^{N \times d}$, where N is the maximum sequence length and d is the feature dimension. Then a positional embedding vector $PE_{pos} \in \mathbb{R}^{N \times d}$ is added to the token embeddings in a pointwise fashion to retain the positional information (174).

The vector $Z_i \in \mathbb{R}^{N \times d}$, hence obtained, is fed to a stack of transformer encoder blocks. Each encoder block is a modular unit consisting of two sub-layers: (i) *Multi-Headed Self-Attention*, and (ii) *Feed-Forward Network*. Furthermore, each sub-layer contains a residual connection, followed by dropout and layer normalization. For token classification, the output of the last encoder layer is passed to a CRF layer (204). This modularity of PLMs enables easy integration of adapter modules in their architecture to make these PLMs task-specific and domain-dependent. We choose RoBERTa (114) as our backbone network as it is the best-performing baseline (see Table 4.4).

4.4.3 Description Infuser Network

DescNet is designed to facilitate deep semantic interaction among the input text and claim descriptions and help underline the key fragments of claims. It consists of precisely engineered components of CoDA and IGM, each devised to augment the process of claim span identification.

To put formally, consider $D = \{d_1, d_2, ..., d_m\}$ as the set of m claim descriptions and $T = \{t_1, t_2, ..., t_n\}$ as the corpus of n input texts. The description representations are extracted from pre-trained RoBERTa (114) and passed through a transformer encoder layer. To begin with, each i^{th} PLM generated vector $Z_i \in \mathbb{R}^{N \times d}$ of input text t_i interacts with each j^{th} description vector $D_j \in \mathbb{R}^{M \times d}$ via the CoDA block. Here, the vector Z_i forms the query, which is processed against the vector D_j acting as the key and value (Equation 4.1).

$$Z_{ij}^{C} = CoDA(Z_i, D_j)D_j \tag{4.1}$$

All such compositionally manipulated vectors Z_{ij}^{C}, after interacting with each j^{th} description vectors, are concatenated and passed through a dropout layer before going through a non-linear transformation for dimensionality reduction (Equation 4.2). The resultant vector Z_i' along with the vector Z_i is passed to the *IGM* module to extract the semantically appropriate features pertinent for fine-grained claim span identification (Equation 4.3).

$$Z_i' = Concat(Z_{i1}^{C}, ..., Z_{im}^{C}) \tag{4.2}$$

$$\hat{Z}_i = IGM(Z_i'W, Z_i) \tag{4.3}$$

The vector $\hat{Z}_i$ is then passed to a CRF layer.

4.4.4 Compositional De-Attention Block

The traditional narrative on attention mechanism (205; 206; 207; 174) is heavily biased on the use of *Softmax* operator where the attention weights are always bounded between $[0, 1]$. Such a convex weighted addition scheme allows the vectors to only contribute in an additive manner. To counter this bottleneck, Tay et al. (208) devised a quasi-attention technique that enables learning of additive as well as subtractive attention weights, allowing the input vectors to add to $(+1)$, not contribute to (0), and even subtract from (-1) the output vector. They decomposed the original Softmax-based self-attention as pointwise multiplication between two matrices as shown in Equation 4.4, where $G(.)$ is the negative pointwise L_1 distance between query Q and key K.

$$A_{quasi} = \left(tanh(\frac{QK^T}{\sqrt{d_k}}) \odot \sigma(\frac{G(Q,K)}{\sqrt{d_k}}) \right) V \tag{4.4}$$

We adopt this quasi-attention strategy to promote more meaningful interaction between the input text and claim descriptions and generate more precise claim-relevant representations.

4.4.5 Interactive Gating Mechanism

To further distinguish salient tokens inclusive in claim spans, we posit *Interactive Gating Mechanism*. To begin with, the vectors Z_i and Z_i' are max pooled to obtain Z_{ip}, $Z_{ip}' \in \mathbb{R}^d$. These vectors are passed through a series of gates, the first of them being the *conflict gate* C, aimed at capturing the semantically conflicting features in Z_i and Z_i' (Equation 4.6).

$$\mu_c = \sigma(Z_{ip}W_{c1} + Z_{ip}'W_{c2} + b_{c1}) \tag{4.5}$$

$$C = \tanh(Z_{ip} \odot \mu_c W_{c3} + Z_{ip}' \odot (1 - \mu_c)W_{c4} + b_{c2}) \tag{4.6}$$

The *refine gate* R, on the other hand, endeavors to capture the semantically similar features between Z_{ip} and Z_{ip}' (Equation 4.8).

$$\mu_r = \sigma(Z_{ip}W_{r1} + Z_{ip}'W_{r2} + b_{r1}) \tag{4.7}$$

$$R = \tanh(Z_{ip} \odot \mu_r W_{r3} + Z_{ip}' \odot \mu_r W_{r4} + b_{r2}) \tag{4.8}$$

We employ an adaptive gating scheme to retain maximum differential information from each gate to congregate the conflicting and similar semantic representations spawned by the gates C and R. It is given by Equation 4.10.

$$A = R + (1 - \mu_r) \odot C \tag{4.9}$$

$$\hat{Z}_i = tanh(AW_a + b_a) \odot Z_i \tag{4.10}$$

Finally, this vector $\hat{Z}_i$ is passed to a CRF layer for token classification.

4.5 Experiments and Results

Baseline Systems. We employ the following baseline systems. ▷ **CNN+CRF**: A Convolutional Neural Network (CNN) trained with GloVe (209) and a CRF head on top. ▷ **BiLSTM+CRF** (210): A sequence labeling model comprising Bidirectional Long Short-Term Memory (BiLSTM) and CRF layer. ▷ **BERT** (211): A bidirectional transformer-inspired auto-encoder language model fine-tune for our span identification task. ▷ **DistilBERT** (203): A smaller, faster, and lighter version of BERT fine-tune on our dataset for the task at hand. ▷ **SpanBERT** (212): An enhanced version of the BERT trained on span prediction objective. ▷ **RoBERTa** (114): A robustly optimized BERT approach, RoBERTa, is a variant of BERT with improved training methodology. We fine-tune it on our dataset. ▷ **NLRG** (122): A system proposed at SemEval-2021 Task 5 on toxic span detection (120). It is a combination of SpanBERT and RoBERTa where the former model is used for predicting the span start and end, while the latter is used for token classification. ▷ **HITSZ-HLT** (213): The system topped the SemEval-2021 task on toxic span detection. They approached the task as a combination of sequence labeling and span extraction and proposed an ensemble of three BERT-based models.

Evaluation Metrics. In concordance with Pavlopoulos et al. (120), we evaluate the performance of all the systems, based on token-level precision (P), recall (R), and F1 scores. To further put a lens over how the models fare for different token types, we calculate the micro-level precision, recall, and F1 score for each of the *'B'*, *'I'*, and *'O'* tokens.[5] Lastly, to quantify the number of tokens included in the spans, we also report the Dice Similarity Coefficient (DSC) (214).

Experimental Results. We summarize our collated results in Table 4.4. Evidently, `DABERTa` outperforms all the baseline systems against most evaluation metrics. We analyze all the systems based on the following research questions.

R1. How accurately do the models predict? To gauge how well each model performs for the token classification task, we monitor precision, recall, and F1 scores. As it can be inferred from Table 4.4, the traditional word embedding-based deep learning models of CNN and BiLSTM give the poorest token classification performance. An appreciable improvement of about 10-14% across all three metrics is observed when we move from the classical deep learning architectures to the transformer-based models of DistilBERT, BERT, SpanBERT, and RoBERTa. This underlines the importance of contextual word embeddings and transformer-based architectures for the task at hand. The addition of the CRF layer further amplifies the performance of these models. SpanBERT also fares better than BERT as it is trained using span prediction objectives. We also notice that employing the CRF layer results in a somewhat better balance of precision and recall when compared to using a basic linear layer. The ensemble-based models of NLRG and HITSZ-HLT also give admissible results for our task. Our proposed model, `DABERTa`, surpasses all the models in terms of precision, recall, and F1 scores. An improvement of about 1.5% is observed between RoBERTa and `DABERTa` regarding these metrics. This justifies the inclusion of *claim descriptions* that amalgamate domain-specific semantic information into RoBERTa architecture via the deftly crafted adapter module. In summary, we see that all the models show a good trade-off between precision and recall.

R2. Are the models aggressive or defensive? Observing the precision, recall, and F1 scores for each of the *'B'*, *'I'*, and *'O'* tags, as shown in Table 4.4, we get an idea of how aggressive or defensive the models are at predicting claim spans. CNN and BiLSTM show considerable resistance in predicting the claim spans, as evidenced by high precision, recall, and F1 scores for the token *'O'* and less for the tokens

[5]Each token in the tweet is *BIO (Begin-Inside-Outside)* encoded to mark the claim spans.

Figure 4.3: A comparative study among DABERTa and baselines. The horizontal bar signifies the ration of the number of predicted spans and the number of gold spans.

'B' and *'I'*. The BERT-based models show a sizable improvement of about 22% and 15% for predicting tokens *'B'* and *'I'*, respectively, over the traditional deep learning models. The addition of CRF layers further bolsters the predictive power for the token *'B'*. DABERTa offers an improvement of about 4-5% over its traditional counterpart for predicting the token *'B'*. Upon close inspection, we observe that the ranges of precision, recall, and F1 scores for predicting the tokens *'I'* and *'O'* vary by not more than 3%. However, the predictive power for the token *'B'* varies vastly by about 25%. Hence, we hypothesize that the inclusion of *descriptions* makes our model cognizant of the syntactic and semantic constructs of claims.

R3. How the models behave for multiple spans?

Figure 4.3 illustrates how well the models identify multiple spans. It is observed that the models of CNN and BiLSTM find it challenging to identify multiple spans. The transformer-based models with a linear head tend to predict more claim spans in the tweet than required. This issue is mitigated when the linear head is replaced with a CRF layer. Still, these models can identify roughly only 80% of the time the occurrence of multiple spans. On the other hand, our model, DABERTa, correctly predicts multiple spans almost more than 85% of the time. Moreover, it does not predict more claim spans than required. Thus, the addition of domain-specific *claim descriptions* appropriately guides DABERTa in identifying the correct occurrence of spans.

Ablation Study. Table 4.4 also reports the ablation studies. Replacing *CoDA* with a naïve *Dot-Product Attention* (*DPA*), we observe a drop in the performance across almost all the metrics. Amongst all, the performance drop in predicting the token *'B'* is the most prominent ($\sim$ 1.5% across precision, recall, and F1). Thus, we conjecture that the quasi-attention mechanism is better able to spot the starting of a claim fragment than *DPA*. When *IGM* is removed, the performance for predicting token *'B'* slightly improves. However, it leads to a decrease in the predictive power for *'O'* token ($\sim$ 2.5% in F1). Therefore, the combination of *CoDA* and *IGM* obtains the most balanced performance. Figure 4.4 reflects the effect of integrating the adapter *DescNet* at different layers of RoBERTa. It is observed that the performance consistently increases as the integration is done at a higher level of RoBERTa layers. This is admissible as studies on probing the PLM layers suggest that different layers encode distinct linguistic properties (215). Furthermore, evidence by Peters et al. (216) suggests that the lower layers of a language model encode the syntactic information, whereas the higher layers capture the complex semantics. Our results are consistent with their findings as we strive to employ deep semantic interaction between the PLM representations and the *claim descriptions*.

Model Name	F1	P	R	F1			Precision			Recall			DSC
				B	I	O	B	I	O	B	I	0	
CNN+CRF	0.6635	0.6709	0.6947	0.3877	0.6766	0.9263	0.3952	0.6725	**0.9450**	0.3953	0.7718	0.9171	0.6964
BiLSTM+CRF	0.6825	0.6928	0.7048	0.4401	0.6717	**0.9356**	0.4653	0.6703	0.9428	0.4302	0.7459	**0.9382**	0.6884
DistilBERT	0.7645	0.7811	0.8068	0.6677	0.8164	0.7510	0.6560	0.7979	0.8310	0.7227	0.8989	0.7402	0.8277
BERT	0.7807	0.7996	0.8154	0.6900	0.8266	0.7699	0.6863	0.8163	0.8403	0.7302	0.8971	0.7634	0.8356
SpanBERT	0.7914	0.8093	0.8182	0.6971	0.8299	0.7901	0.7047	0.8384	0.8271	0.7203	0.8724	0.8048	0.8377
RoBERTa	0.8020	0.8163	0.8337	0.7221	0.8297	0.7942	0.7165	0.8371	0.8351	0.7624	0.8764	0.8022	0.8399
DistilBERT + CRF	0.8288	0.8581	0.8526	0.8722	0.8148	0.7431	0.8914	0.7852	0.8400	0.8621	0.9181	0.7219	0.8222
BERT + CRF	0.8368	0.8631	0.8556	0.8531	0.8284	0.7666	0.8781	0.8101	0.8375	0.8408	0.9042	0.7597	0.8343
SpanBERT + CRF	0.8390	0.8625	0.8562	0.8507	0.8253	0.7806	0.8742	0.8221	0.8302	0.8394	0.8827	0.7867	0.8316
RoBERTa + CRF	0.8457	0.8706	0.8635	0.8613	0.8340	0.7805	0.8874	0.8301	0.8321	0.8485	0.8972	0.7841	0.8402
NLRG	0.7494	0.7750	0.7832	0.6584	0.7892	0.7398	0.6631	0.7891	0.8119	0.6805	0.8600	0.7486	0.8087
HITSZ-HLT	0.7758	0.7966	0.8037	0.6754	0.8201	0.7780	0.6834	0.8230	0.8291	0.6978	0.8743	0.7850	0.8314
DABERTa	**0.8604**	**0.8814**	**0.8789**	0.9035	**0.8354**	0.7816	0.9205	**0.8242**	0.8379	0.8950	**0.9044**	0.7771	**0.8433**
$-\{IGM\}$	0.8539	0.8795	0.8768	**0.9121**	0.8310	0.7563	**0.9258**	0.8017	0.8473	**0.9051**	0.9277	0.7358	0.8401
$-\{CoDA\}+\{DPA\}$	0.8558	0.8788	0.8738	0.8886	0.8319	0.7821	0.9086	0.8184	0.8439	0.8785	0.9032	0.7752	0.8416

Table 4.4: Experimental results of DABERTa, its variants (last two rows), and baselines. DSC, P, and R denote Dice Similarity Coefficient, Precision, and Recall, respectively.

Implementation Details. We utilize the *base* version of RoBERTa (114) to propose DABERTa. The model is trained end-to-end using the Adam optimizer (183), learning rate of $4e-5$, and batch size of 32 for 20 epochs with early stopping if the dice score does not improve after 5 epochs. We used the Nvidia Tesla v100 32 GB GPU. The hyper-parameter tuning is done with respect to the validation dataset.

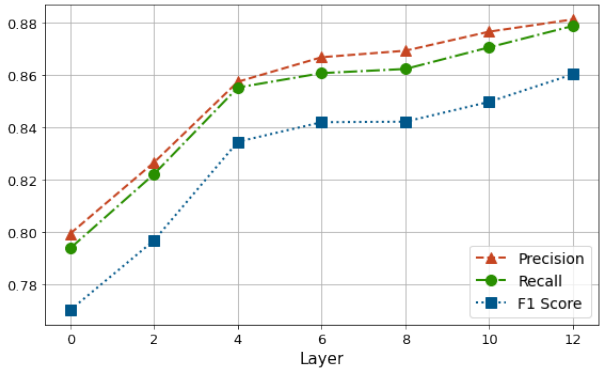


Figure 4.4: Performance of DABERTa when the adapter module is inserted at different layers of RoBERTa.

4.6 Error Analysis

In this section, we manually analyze the errors the models are prone to make. Table 4.5 highlights randomly sampled tweets from our dataset, CURT, along with their gold spans and predictions from DABERTa. In addition, we also consider the predictions from the best-performing baseline, RoBERTa, for a fair comparison. We analyze the errors committed by both the systems and divide them into three different categories: (i) tweets with a single-claim span, (ii) tweets with claim-like premises, and (iii) tweets with claims that can be inferred from the underlying undertone of the tweet but no explicit span can be marked to highlight the claim-specific connotation, e.g., figurative sentences, satire, indirect questions etc. (Note: For simplicity, we refer to such claims as *implicit* claims.) In the most straightforward situation where the tweets only contain a single claim DABERTa makes more precise predictions than the baseline system as shown in the first example of Table 4.5. We observe that both the models identify the claim-span correctly; however, RoBERTa identifies some unnecessary spans, which trespasses our objective of equipping the fact-checkers with only relevant information. The second type of error related to spans is the presence

of claim-like premises. Claims and premises[6] are closely related components of argument mining, and differentiating them is strenuous, even for humans. Example 2 in Table 4.5, exhibits a post containing claim-premise pair. There are two conclusive claims in the tweet – *'#coronavirus was used by the #CCP as a bio weapon'* and *'CCP is kicking out black people from hotels even if they don't have covid'*. Even though *'not only to kill people but to encourage racism among their citizens against foreigners'* appears to be a claim at first glance, it serves as the premise to support the conclusive part of the arguments brought forward in the tweet. In most cases, we discern that both the systems identify the claim-spans correctly, but they are easily fooled by the premises, hence leaving room for significant improvement in this regard.

Another prominent class of errors is implicit claims. Extracting the claim-spans in implicit claims is arduous. We observe that both the systems strive to understand the linguistic structure of the implicit claims. For instance, in sample 3, the user intends to assert that honey, ginger, garlic, or turmeric do not cure COVID19; however, DABERTa fails to understand the user's intention and yields the wrong span. We perceive similar behavior from the best-performing baseline, RoBERTa, as well. A plausible reason is the skewed nature of the dataset, which is lopsided with a significant bias toward explicit claims. According to our observations, DABERTa outperforms the best-performing baseline system significantly ($\sim 4\%; p < 0.0004$).[7] Hence, furnishing us with empirical shreds of evidence that DABERTa can be efficiently used for claim span identification.

	Model	Tweet
1	***Gold***	Gold Truly sobering analysis: **US more vulnerable than many countries to #coronavirus** owing to combination of high numbers of uninsured, many w/o paid sick leave, and a leadership that has downplayed the challenge while not preparing the country for it.
	RoBERTa	Gold Truly sobering analysis: **US more vulnerable than many countries to #coronavirus** owing to combination of *high numbers of uninsured*, many w/o *paid sick leave*, and a leadership that has downplayed the challenge while not preparing the country for it.
	DABERTa	Gold Truly sobering analysis: **US more vulnerable than many countries to #coronavirus** owing to combination of high numbers of uninsured, many w/o paid sick leave, and a leadership that has downplayed the challenge while not preparing the country for it.
2	***Gold***	Whether made on purpose or not **#coronavirus was used by the #CCP as a bio weapon**, not only to kill people but to encourage racism among their citizens against foreigners. Especially black people, **CCP is kicking out black people from hotels even if they dont have covid.**
	RoBERTa	Whether made on purpose or not **#coronavirus was used by the #CCP as a bio weapon**, *not only to kill people but to encourage racism among their citizens against foreigners*. Especially black people, CCP is kicking out black people from hotels even if they dont have covid.
	DABERTa	Whether made on purpose or not **#coronavirus was used by the #CCP as a bio weapon**, *not only to kill people but to encourage racism among their citizens against foreigners*. Especially black people, **CCP is kicking out black people from hotels even if they dont have covid.**
3	***Gold***	RT @HealtheNews: Can honey, ginger, garlic or turmeric or any other home remedies cure #Covid19? No, here's why.
	RoBERTa	RT @HealtheNews: Can *honey, ginger, garlic or turmeric or any other home remedies cure #Covid19?* No, here's why.
	DABERTa	RT @HealtheNews: Can *honey, ginger, garlic or turmeric or any other home remedies cure #Covid19?* No, here's why.

Table 4.5: Error analysis of the outputs. Bold text (green) highlights the correct claim span whereas the text in italics (red) represents the mistakes committed by our model, DABERTa, and vanilla RoBERTa as the baseline.

4.7 Summary

Through this systematic research, we introduced the novel task of *Claim Span Identification*, which is valuable on various fronts. We conducted an evidence-based document retrieval experiment, demonstrating that employing claim spans retrieves more relevant evidence than using the entire tweet. Furthermore,

[6](sub)sentences used to support the concluding claim

[7]We also perform significance t-test on F1 scores per input tweet comparing DABERTa and RoBERTa, as best-performing baseline.

as there exists no specialized corpus for claim span identification, we compiled CURT, a large-scale Twitter corpus consisting of around $7.5k$ tweets annotated with token-level claim spans. We showed convincing results using various token classification baselines on our dataset. Moreover, we benchmarked CURT with DABERTa, an adapter-based variant of RoBERTa, that encapsulates critical domain-specific information into the pre-trained model via *claim descriptions*. Through extensive qualitative, quantitative, and empirical results, we illustrated how DABERTa outperforms the other models on different fronts. Lastly, we also developed an extensive set of annotation guidelines and released them for further research. Though DABERTa yields the state-of-the-art performance in claim span identification; there are a few cases where it falls short. Even for humans, recognizing claim spans in figurative or metaphorical sentences is arduous; consequently, our suggested model also struggles with them. As a result, our future study will focus on boosting the claim span identification performance, especially for such sentences. Our analysis also showed that the high resemblance between claims and premises confuses the model, making it difficult to distinguish between the two. DABERTa shares the said limitation with other baseline systems as well. As a result, this could be another alluring open challenge to work on.

5. ABSTRACTIVE CLAIM SIMPLIFICATION

With the rise of social media, users are exposed to many misleading claims. However, the pervasive noise inherent in these posts presents a challenge in identifying precise and prominent claims that require verification. Extracting the important claims from such posts is arduous and time-consuming, yet it is an underexplored problem. Here, we aim to bridge this gap. We introduce a novel task, *Claim Normalization* (*aka ClaimNorm*), which aims to decompose complex and noisy social media posts into more straightforward and understandable forms, termed *normalized claims*. We propose CACN, a pioneering approach that leverages chain-of-thought and claim check-worthiness estimation, mimicking human reasoning processes, to comprehend intricate claims. Moreover, we capitalize on the in-context learning capabilities of large language models to provide guidance and to improve claim normalization. To evaluate the effectiveness of our proposed model, we meticulously compile a comprehensive real-world dataset, CLAN, comprising more than $6k$ instances of social media posts alongside their respective normalized claims. Our experiments demonstrate that CACN outperforms several baselines across various evaluation measures. Finally, our rigorous error analysis validates CACN's capabilities and pitfalls.

5.1 Introduction

Social media have enabled a new way of communication, breaking down geographical barriers and bringing unprecedented opportunities for knowledge exchange. However, this has also presented a growing threat to society, e.g., during the 2016 US Presidential Election (217), the COVID-19 pandemic (218; 219; 220), the Ukraine-Russia conflict (221), etc. False claims are an intrinsic aspect of fabricated news, rumors, propaganda, and misinformation. Journalists and fact-checkers work tirelessly to assess the factuality of such claims in spoken and/or written form, sifting through an avalanche of claims and pieces of evidence to determine the truth. To further address this pressing issue, several independent fact-checking organizations have emerged in recent years, such as Snopes,[1] FullFact,[2] and PolitiFact,[3] which play a crucial role in verifying the accuracy of online content. However, the rate at which online information is being disseminated far outpaces the capacity of fact-checkers, making it difficult to verify every single claim. This, in turn, leaves numerous unverified claims circulating online, potentially reaching millions before they can be verified. While the complete automation of the fact-checking pipeline may pose hazards to accountability and reliability, several recent studies have targeted identifying downstream tasks suitable for automation, such as detecting claims (191; 111; 222), evaluating their worthiness for fact-checking (223; 138; 59), making sure they were not fact-checked before (224; 225; 226; 227), and validating them by retrieving relevant shreds of evidence (60; 194; 195; 228).

In light of the growing challenges faced by fact-checkers in verifying the factuality of social media

[1] https://www.snopes.com
[2] https://fullfact.org
[3] https://www.politifact.com

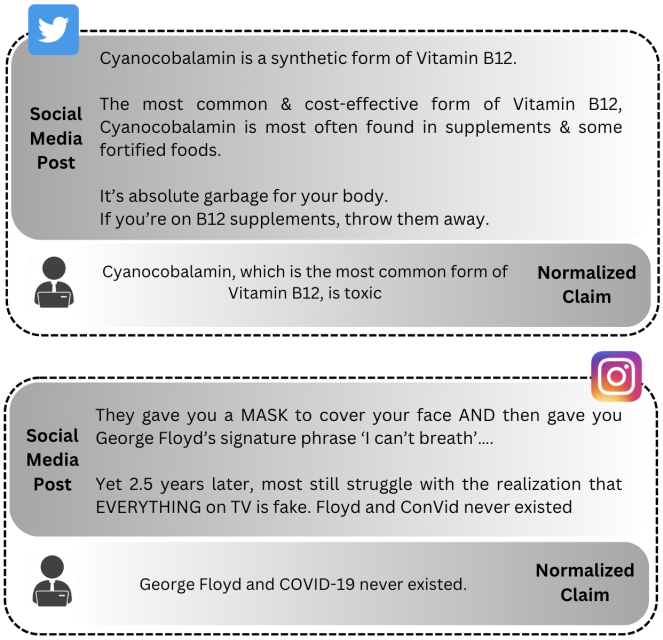


Figure 5.1: Illustration of our proposed *Claim Normalization* task, highlighting the normalized claims authored by fact-checkers for social media posts from distinct social media platforms.

claims, we propose the novel task of *claim normalization*. This task aims to extract and to simplify the central assertion made in a long, noisy social media post. This can improve the efficacy and curtail the workload of fact-checkers while maintaining high precision and conscientiousness. To better understand, we illustrate the task in Figure 5.1. The first social media post reads, *'Cyanocobalamin is a synthetic form of Vitamin B12...If you're on B12 supplements, throw them away.'* This post contains some extraneous information that has no relevance for fact-checkers. As a result, they distil the information and summarize it as, *'Cyanocobalamin, the most common form of Vitamin B12, is toxic.'* Fact-checkers tasked with verifying the accuracy of such noisy posts need to read through them and condense their content to obtain a concise claim that can be easily fact-checked. Unfortunately, this process can be exceedingly time-consuming. By automating the claim normalization process, fact-checkers can work more efficiently. Another aspect is that fact-checkers often choose what to fact-check based on the virality of a claim, for which they need to be able to recognize when the same claim appears in a slightly different form, and claim normalization is essential for this.

Task Motivation. Claim normalization holds significant promise for combating the spread of misinformation by streamlining fact-checking processes and enhancing the reliability of retrieved evidence. To substantiate our hypothesis regarding the effectiveness of claim normalization, we conducted a well-structured retrieval experiment using the Google API. The objective was to demonstrate the practical benefits of claim normalization in assisting fact-checkers. We randomly selected a sample of 35 instances from our dataset, encompassing social media posts and their normalized claims. Leveraging the capabilities of the Google API, we sought the top-5 most relevant articles for each post and its normalized claim. In a meticulous evaluation process, three annotators individually assessed the relevance (0 or 1) of each retrieved article to the input (post or normalized claim). We then used majority voting to determine the final relevance score for each retrieved article. As depicted in Table 5.1, the results of our experiment consistently demonstrated the advantage of normalized claims in evidence retrieval. In top-k precision evaluations for various values of k (1, 3, and 5), normalized claims consistently outperformed their

corresponding source posts. This observation indicates that claim normalization is not merely a theoretical concept, but significantly enhances the efficiency of evidence retrieval, resulting in more concise and effective tools for aiding the fact-checking process.

	P@1	P@3	P@5
Original Post	0.82	0.64	0.58
Normalized Claim	**0.88**	**0.73**	**0.69**

Table 5.1: Comparative top-k precision evaluations of normalized claim vs. original posts in evidence retrieval.

Our contributions are as follows:

- We introduce the novel task of *claim normalization*, which seeks to detect the core claim in a given piece of text.
- We present a meticulously curated high-quality dataset specifically tailored for claim normalization of noisy social media posts.
- We propose a robust framework for claim normalization, incorporating chain-of-thought, in-context learning, and claim check-worthiness estimation to comprehend intricate claims.
- We conduct a thorough error analysis, which can inform future research.

5.2 Related Work

Previous work has focused on distinct aspects of claims, including claim detection (191; 2; 197; 229; 222), claim check-worthiness estimation (140; 223; 230; 138; 58; 198; 231), claim span identification (232), etc. By curating the AAWD corpus, Bender et al. (196) pioneered the efforts in claim detection, the foremost step in the fact-checking tasks. Following this, linguistically motivated features, including sentiment, syntax, context-free grammar, and parse trees, were frequently used (191; 55; 110; 197). Recently, large language models (LLMs) have also been used for claim detection (111; 198; 2; 229; 222). Most previous claim detection and extraction work primarily concentrated on adapting to text from similar distributions or topics. Moreover, it often relied on well-structured formal writing. In contrast, our objective is to develop a system that specifically addresses the challenges posed by posts in social media and aims to extract the central claim in a more simplified manner, which goes beyond extracting a text sub span in a social media post and aims at abstractive claim extraction that *mimics what professional fact-checkers do*. To the best of our knowledge, we are the first to address the task of claim extraction in this very practical formulation.

Text Summarization. The task of claim normalization is closely related to text summarization. In the latter, given a lengthy document, the goal is to summarize it into a much shorter summary. Previous work on text summarization has explored various approaches, including large pre-trained seq2seq models to generate high-quality summaries (128; 129; 130). One issue has been the faithfulness of the summary with respect to the source. To address this, Kryscinski et al. (131) introduced FactCC, a weakly supervised BERT-based entailment model, which augments the dataset with artificially introduced faithfulness errors. Similarly, Utama et al. (132) trained a model for detecting factual inconsistencies in data from controllable text generation that perturbs human-annotated summaries, introducing varying types of factual inconsistencies. Durmus et al. (133) proposed a question-answering framework that compares answers from the summary to those from the original text. All these approaches primarily focused on

general-purpose summarization and did not provide means for models to generate summaries primarily focusing on specific needs. To address this limitation, controlled summarization was introduced by Fan et al. (134). One aspect of controlled summarization is length control, in which users can set their preferred summary length (135; 136). Recent research has discovered that, despite their fluency and coherence, state-of-the-art abstractive summarization systems produce summaries with contradictory information.

While text summarization systems can assist in condensing social media posts into shorter summaries, their primary goal is not to ensure verifiability. It aims to capture the text's key points rather than emphasizing the specific claims that need to be fact-checked. Our task of claim normalization, on the other hand, works at an entirely different level. It needs a thorough understanding of the claims made in the social media post and strives to ensure that the normalized claims are not only consistent with the original post, but are also self-contained and verifiable. Despite the progress in text summarization, the task of claim normalization remains underexplored. In this work, we aim to tackle this challenging problem by developing a robust approach specifically tailored to the unique aspects of this task.

5.3 Dataset

Existing text summarization datasets have not specifically addressed the need for claim-oriented summaries. To address this gap, we propose a novel dataset `CLAN` (**Cla**im **N**ormalization), consisting of fact-checked social media posts paired with concise claim-oriented summaries (known as *normalized claims*), created by fact-checkers as part of the verification process. As a result, our dataset is not subjected to external annotation, thus averting potential biases and ensuring its high quality.

5.3.1 Collection

We gathered our fact-checked post and claim pairs from two sources: (*i*) Google Fact-Check Explorer[4] and (*ii*) ClaimReview Schema.[5]

Google Fact-Check Explorer. We acquired a list of fact-checked claims from multiple reputed fact-check sources via Google Fact-Check Explorer's API (GFC). This data collection pipeline followed a three-step process. First, we extracted the *title*, which is usually a single-sentence short summary of the information being fact-checked, and the *fact-checking site's URL*. This step yielded a total of 22,405 unique fact-checks. We then proceeded to retrieve the social media post and the associated *claim review* if they were available on the fact-checking site. Due to the collected posts having already undergone fact-checking and containing misleading claims, a significant number of them were unavailable for inclusion in our dataset. Moreover, a significant number of the posts only contained images or videos, which were unsuitable for our task at hand. As a result, we were left with a considerably smaller number of relevant instances. We also noted that in certain instances, the *title* in the Google Fact-Check Explorer and the *claim review* were identical; consequently, we included only one in the final dataset.

ClaimReview Schema. We targeted the ClaimReview Schema elements with an entry for *reviewed items* as they were relevant to our requirements. Out of 44,478 entries, only 22,428 had this particular

[4] https://toolbox.google.com/factcheck/explorer
[5] https://schema.org/ClaimReview

	Social Media Post	Normalized Claim
1	Research into the dangers of cooking with aluminum foil has found that some of the toxic metal can contaminate food. Increased levels of aluminum in the body have been linked to osteoporosis, and Alzheimer's disease.	Cooking in Aluminum foil causes Alzheimer's Disease.
2	Did you know when ur child turns 6. U can add them as authorized user to one of ur credit cards. Never give them card, & all payments u make from 6 to 18 goes to ur child credit too..ur kid will have a unbelievable credit score from years of payment history.	6-year-old kids can be added as authorized users on all credit cards.
3	As if it couldn't get any worse. #Hope4Cancer says #RootCanal causes #CANCER Solution... Rip Cancer patients teeth out. Monsters #FalseHope4Cancer #ProtectCancerPatients	Having a root canal can cause cancer. Root canal treatment causes cancer

Table 5.2: Examples of social media posts and their corresponding normalized claims from CLAN. The first two examples come from the training set and each has one reference normalized claim, while the last one comes from the test set, and thus it has two reference normalized claims.

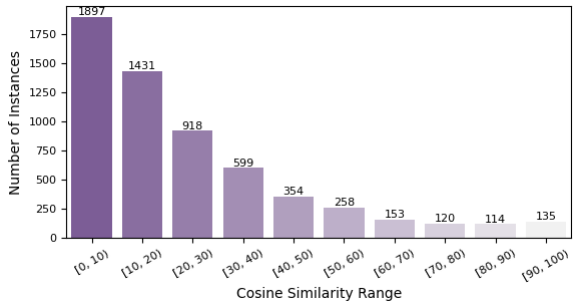


Figure 5.2: Histogram of the cosine similarity between the social media posts and the corresponding normalized claims from our CLAN dataset.

field. Therefore, we had to filter out the remaining entries. Next, we extracted all the links to social media posts and their corresponding claim reviews provided by the fact-checkers.

As mentioned above, we only processed textual claims and excluded other modalities, such as audio or video. Further, we ensured that all the entries were in English.

5.3.2 Statistics and Analysis

By using both data collection methods and exercising careful consideration, we curated a total of 6,388 instances. To ensure the creation of a diverse and high-quality test set, we chose posts that comprised not one but two reference normalized claims (*c.f.* Sec 5.3.1). This enabled us to capture different aspects and perspectives of the normalized claims by including multiple references, thereby increasing the test set's robustness and reliability. Representative examples from our dataset are shown in Table 5.2, and the final dataset statistics are shown in Table 5.3.

Dataset	Train	Val	Test	Overall
Total number of pairs	5,341	594	453	6,388
Avg. length of posts	39.52	37.12	57.97	44.87
Avg. length of claims	16.47	17.24	15.41	16.37

Table 5.3: Statistics about our CLAN dataset.

Figure 5.2 shows an analysis of the cosine similarities between the social media posts and the corresponding normalized claims. We can see that the cosine similarities are consistently low for most

examples, demonstrating that claim normalization involves more than just summarizing the social media post. This highlights the need for a specialized effort to accurately identify, extract, and normalize the claims within social media posts.

5.4 Methodology

In this section, we explain our proposed approach, **C**heck-worthiness **A**ware **C**laim **N**ormalization (CACN). We aim to integrate task-specific information with large language models (LLMs), as shown Figure 5.3. We focus our experiments on GPT-3 (`text-davinci-003`) (233). Our approach amalgamates two key ideas: (*i*) chain-of-thought prompting and (*ii*) reverse check-worthiness.

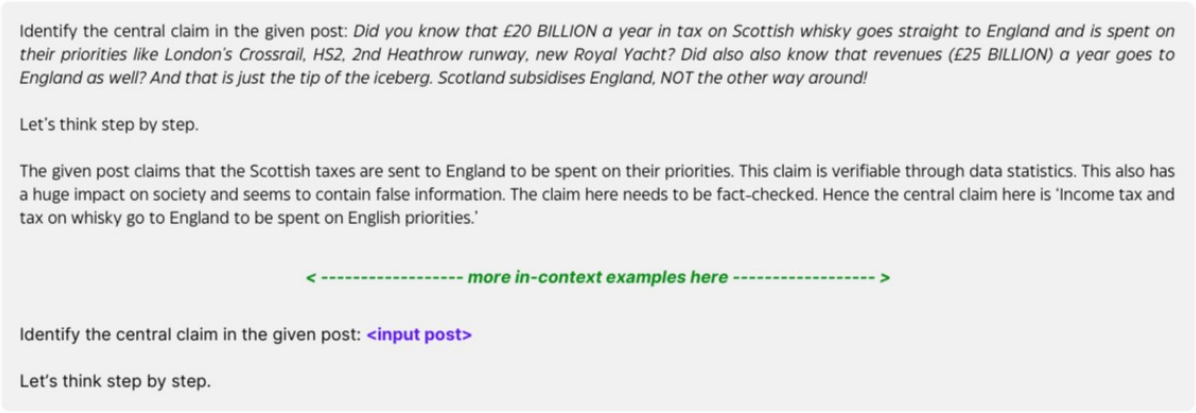


Figure 5.3: Illustration of our proposed approach, CACN. To generate a normalized claim, we use the CACN prompt template, which encompasses explicit task instruction and relevant in-context examples, as well as chain-of-thought reasoning.

5.4.1 Chain-of-Thought Prompting

The realm of chain-of-thought (CoT) prompting has emerged as a veritable tour de force within LLMs (234). Instead of undergoing the laborious process of fine-tuning individual model checkpoints for every new task, we use CoT to navigate the complexity of claim normalization by using step-by-step reasoning. To accomplish this, we use claim check-worthiness, as described in the following subsection. This enables the model to iteratively enhance its comprehension and effectively generate precise normalized claims while eliminating the need for extensive fine-tuning.

5.4.2 Reverse Check-Worthiness

The idea about reverse check-worthiness originates from the task of check-worthiness estimation, which in turn is an integral part of the manual fact-checking process (235). We leverage check-worthiness to steer the model's attention toward salient and pertinent information. By giving the model the ability to produce rationales in natural language that clearly explain the sequence of reasoning stages leading to the solution, we strengthen its capacity for cognitive reasoning with unwavering efficacy. Based on prior research on claim check-worthiness (230; 236; 237), we direct our model to prioritize claims that meet specific criteria within the given social media post. These criteria include identifying claims within social

media posts that (*i*) contain verifiable factual claims, (*ii*) have a higher likelihood of being false, (*iii*) are of general public interest, (*iv*) are likely to be harmful, and (*v*) are worth fact-checking. For instance, in Figure 5.3, the claim normalization process begins by identifying the central claim within the input social media post. Subsequently, we reckon the claim's verifiability, i.e., whether it is self-contained and verifiable (e.g., as opposed to not containing a claim or expressing an opinion, etc.). We further evaluate the likelihood of the claim being false and its overall check-worthiness. This step-by-step process ensures a comprehensive analysis of the central claim's characteristics, allowing for effective claim normalization. By incorporating these aspects into our approach, we aim to improve the model's ability to identify and prioritize claims that require scrutiny and verification.

5.5 Experiments and Results

Baseline Models. For comparison, we use several state-of-the-art generative systems and categorize them into two groups: (*i*) *Pre-trained Large Language Models (PLMs)*: T5 (130), BART (129), FLAN-T5 (238), and PEGASUS (239). For T5, BART, and FLAN-T5, we use `base` and `large` model sizes. For PEGASUS, we use the `reddit` model. (*ii*) *In-context Learning Model:* GPT-3 (`text-davinci-003`) (233).

Evaluation Measures. To evaluate lexical overlap, we use ROUGE $(1, 2, L)$ and BLEU-4 (240). We further use METEOR (241) and BERTScore (242) to assess the similarity between the gold and the generated normalized claims.

Experimental Setup. We perform four set of experiments: ▷ **Zero-Shot Learning:** It aims to apply the previously acquired capabilities of PLMs to similar tasks in a low domain. We hereby assess its suitability for the claim normalization task. ▷ **Few-Shot Learning:** We adopt few-shot learning with 10, 20, 50, and 100 training examples. This gradual exposure to additional labeled data aims to enhance the models' ability to generate accurate and contextually appropriate normalized claims. ▷ **Prompt Tuning:** Prompt-tuning entails adding a specific prefix to the model's input customized to the downstream tasks (243). We investigate the impact of affixing different prompts to the given posts on the performance of T5-based and GPT-3 models. To control the generated normalized claims, we use five control aspects: tokens, abstractness, number of sentences, claim-centricity, and entity-centricity. ▷ **In-Context Learning:** LLMs can tackle diverse tasks with a minimal amount of examples given in-context learning prompts (233). We use GPT-3 (`text-davinci-003`) with three different prompts: (*i*) direct prompt (DIRECT), (*ii*) question-guided prompt (Q-GUIDED), and (*iii*) zero-shot chain-of-thought (ZS-CoT). Detailed prompt templates are given in Table 5.4.

Implementation Details. We performed basic data cleaning, e.g., removing non-alphanumeric characters, removing links and hashtags, etc. on our dataset `CLAN`, using *nltk*. For a standardized evaluation, we relied on widely recognized evaluation libraries such as *py-rouge*,[6] *nltk-bleu*,[7] *nltk-meteor*,[8] and *hugging-face bert-score*.[9] We trained all models for 50 epochs, with early stopping based on validation loss. We set the patience value at 5, and we optimized the models using the Adam optimizer. We set the

[6] https://pypi.org/project/py-rouge/
[7] https://www.nltk.org/_modules/nltk/translate/bleu_score.html
[8] https://www.nltk.org/api/nltk.translate.meteor_score.html
[9] https://huggingface.co/spaces/evaluate-metric/bertscore

DIRECT

For the given post get the normalized claim:

Post: "Cough CPR is a form of self CPR. Coughing increases intrathoracic pressure and squeezes blood out of the heart into the aorta. Theoretically, one can remain conscious if one continues to cough during a cardiac arrest (not heart attack without cardiac arrest – all heart attacks do not cause cardiac arrest).
Normalized claim: Cough CPR can prevent cardiac arrest.

< -------- more in-context examples here -------- >

For the given post get the normalized claim:

Post: **<input_post>**
Normalized claim: **<generated_response>**

Q-GUIDED

What is being claimed in the given post?

Cough CPR is a form of self CPR. Coughing increases intrathoracic pressure and squeezes blood out of the heart into the aorta. Theoretically, one can remain conscious if one continues to cough during a cardiac arrest (not heart attack without cardiac arrest – all heart attacks do not cause cardiac arrest).
The answer is: Cough CPR can prevent cardiac arrest.

< -------- more in-context examples here -------- >

What is being claimed in the given post?

<input_post>
The answer is: **<generated_response>**

ZS-CoT

Identify the central claim in the given post. Let's think step-by-step.

Post: **<input_post>**
Central claim: **<generated_response>**

Figure 5.4: Our templates for in-context learning prompts used for GPT-3 (`text-davinci-003`).

weight decay to 0.01. For our proposed approach CACN, we used GPT-3 (`text-davinci-0003`) as the base model. Finally, we set the maximum length of the generated response to 120 with a temperature of 0.6.

Experimental Results. Our experiments reveal that CACN outperforms all baselines across most evaluation measures. We further examine all systems aiming to answer the research questions listed below.

R1. Do meticulously crafted prompts enhance the performance of generative models? The findings exhibit a significant performance improvement when using prompt-tuning, specifically with in-context examples. Table 5.4 shows the effectiveness of various prompts across all evaluation measures. However, a notable enhancement of approximately 2-3 points absolute is observed for all semantic measures when transitioning from conventional prompts to our proposed approach when using the same in-context

	Model	ROUGE-1			ROUGE-2			ROUGE-L			BLEU-4	METEOR	BERTScore
		P	R	F1	P	R	F1	P	R	F1			
Finetune	$T5_{BASE}$	23.65	43.60	28.99	11.38	20.70	13.90	20.45	37.91	25.15	4.57	28.86	85.13
	$T5_{LARGE}$	23.37	44.81	28.99	10.98	21.11	13.66	20.07	38.46	24.97	4.43	29.33	85.09
	$BART_{BASE}$	33.41	41.64	34.11	17.57	21.05	17.55	29.70	36.61	30.25	6.69	29.57	86.32
	$BART_{LARGE}$	35.83	42.88	36.12	**19.25**	21.73	18.97	31.64	37.65	31.93	7.71	31.07	86.92
	FLAN-$T5_{BASE}$	22.50	47.38	28.79	10.65	22.28	13.61	19.15	40.14	24.50	4.44	29.07	84.66
	FLAN-$T5_{LARGE}$	28.70	45.98	31.35	15.32	22.63	16.24	25.46	39.89	27.63	6.62	29.36	84.65
In-context	DIRECT	32.19	46.43	35.52	14.19	20.60	15.65	27.32	39.75	30.31	6.52	33.25	88.87
	Q-GUIDED	33.48	44.50	35.40	15.34	20.42	16.15	29.10	38.88	30.90	6.56	32.13	88.81
	ZS-CoT	27.77	**48.39**	32.42	13.18	22.33	15.16	23.84	**41.51**	27.79	6.37	32.93	88.44
	CACN (ours)	**37.54**	46.10	**38.64**	18.85	**23.08**	**19.32**	**33.14**	40.92	**34.30**	**9.66**	**35.10**	**89.00**
	$\Delta_{\text{CACN}-BEST(\%)}$	↑ 4.77	↓ 4.73	↑ 6.98	↓ 2.08	↑ 1.99	↑ 1.85	↑ 4.74	↓ 1.42	↑ 7.42	↑ 25.29	↑ 5.56	↑ 1.15

Table 5.4: Experimental results of CACN and baseline systems on CLAN. We report ROUGE $(1, 2, L)$, BLEU-4, METEOR, and BERTScore. The best scores are shown in **bold**, while the second-best scores are underlined, across each metric. The last row gives the percentage increase in performance between CACN and the best baseline.

N-shot	Model	ROUGE-1			ROUGE-2			ROUGE-L			BLEU-4	METEOR	BERTScore
		P	R	F1	P	R	F1	P	R	F1			
0	$T5_{BASE}$	23.38	42.25	27.79	11.16	19.54	13.14	20.29	36.52	24.10	4.30	28.45	85.36
	$T5_{LARGE}$	24.80	43.71	29.08	12.44	20.93	**14.31**	21.58	37.82	25.30	**5.01**	29.36	**85.65**
	$BART_{BASE}$	22.57	47.76	27.96	11.24	22.60	13.60	19.58	40.91	24.14	4.64	**30.44**	85.11
	$BART_{LARGE}$	20.80	39.68	24.41	9.86	17.20	11.16	18.23	33.85	21.20	3.92	24.67	84.78
	FLAN-$T5_{BASE}$	30.78	32.33	28.73	15.03	15.17	13.78	28.08	28.99	26.02	4.76	23.27	83.80
	FLAN-$T5_{LARGE}$	**31.20**	34.06	**30.22**	**15.67**	16.83	15.07	**28.41**	30.57	**27.40**	5.76	25.82	84.89
	PEGASUS	23.96	36.07	26.16	11.69	16.76	12.48	21.14	31.50	22.99	4.87	24.87	83.07
10	$T5_{BASE}$	21.55	44.22	27.44	9.87	20.36	12.57	18.31	37.77	23.40	4.04	28.64	84.97
	$T5_{LARGE}$	21.92	45.73	28.04	10.17	21.04	12.98	18.64	38.81	23.85	4.19	29.52	85.11
	$BART_{BASE}$	20.09	53.08	25.65	10.14	24.94	12.60	17.34	44.77	22.00	4.11	29.67	84.72
	$BART_{LARGE}$	19.71	**53.22**	25.30	9.83	24.87	12.27	17.01	**45.22**	21.70	4.00	29.51	84.65
	FLAN-$T5_{BASE}$	23.05	42.60	27.97	10.58	19.48	12.84	19.83	36.58	24.14	4.39	28.04	85.34
	FLAN-$T5_{LARGE}$	22.41	29.66	19.99	10.68	13.18	9.49	20.50	25.91	17.84	3.04	18.87	78.04
	PEGASUS	15.26	40.32	20.86	7.15	18.29	9.58	13.39	35.87	18.35	3.44	24.73	80.02
20	$T5_{BASE}$	21.56	44.22	27.46	9.89	20.36	12.59	18.33	37.78	23.42	4.05	28.64	84.97
	$T5_{LARGE}$	21.97	45.80	28.07	10.24	21.11	13.03	18.69	38.77	23.86	4.21	29.46	85.12
	$BART_{BASE}$	20.08	53.16	25.63	10.14	**25.03**	12.60	17.34	44.89	21.99	4.11	29.63	84.72
	$BART_{LARGE}$	19.66	53.11	25.25	9.80	24.84	12.24	16.96	45.08	21.64	3.97	29.46	82.69
	FLAN-$T5_{BASE}$	23.15	42.80	28.11	10.63	19.58	12.91	19.93	36.76	24.27	4.44	28.20	85.34
	FLAN-$T5_{LARGE}$	22.38	29.79	20.00	10.69	13.25	9.51	20.48	26.00	17.85	3.04	18.96	78.01
	PEGASUS	15.25	40.30	20.84	7.15	18.29	9.58	13.38	35.87	18.35	3.44	24.72	80.01

Table 5.5: Zero-shot and few-shot performance on our dataset CLAN. We report ROUGE $(1, 2, L)$, BLEU-4, METEOR, and BERTScore.

examples. This emphasizes the importance of our framework tailored for the specific task. Moreover, an upsurge in ROUGE-F1 scores (1, 2, and L) emphasizes the resemblance between the generated normalized claims and such created by humans. This, in turn, validates the incorporation of the "*reverse check-worthiness*" chain-of-thought process, which effectively integrates task-specific information into the generative system. We also attempt prompt-tuning in a zero-shot setup; the results are shown in Table 5.6. To summarize, the deliberate design of prompts and in-context learning substantially enhance generative models' performance.

R2. Is training models on a specific task less effective than in-context learning with a few examples? We observe substantial disparities in the performance of models trained on task-specific data compared

	Model	ROUGE-1			ROUGE-2			ROUGE-L			BLEU-4	METEOR	BERTScore
		P	R	F1	P	R	F1	P	R	F1			
$T5_{BASE}$	UNCONTROLLED	23.38	42.25	27.79	11.16	19.54	13.14	20.29	36.52	24.10	4.30	28.45	85.36
	TOKEN-LIMIT	22.87	42.71	27.59	11.42	20.61	13.60	20.21	37.47	24.34	4.60	28.67	85.31
	ABSTRACTNESS	23.02	42.21	27.58	11.40	20.20	13.47	20.17	36.79	24.16	4.53	28.46	85.41
	SINGLE SENTENCE	21.06	31.19	23.07	9.61	14.02	10.46	18.30	26.75	19.96	3.57	21.77	84.41
	CLAIM-CENTRIC	19.76	38.89	24.20	8.90	17.94	10.96	17.37	34.33	21.38	3.51	25.77	84.58
	ENTITY-CENTRIC	19.08	32.88	22.63	8.37	14.29	9.91	16.66	28.70	19.81	3.26	20.11	83.20
$T5_{LARGE}$	UNCONTROLLED	24.80	43.71	**29.08**	**12.44**	20.93	14.31	21.58	37.82	**25.30**	**5.01**	29.36	85.65
	TOKEN-LIMIT	14.51	36.24	19.53	6.24	16.01	8.45	12.79	32.47	17.33	2.56	22.13	82.08
	ABSTRACTNESS	23.53	44.44	28.62	11.69	20.98	13.96	20.66	38.64	25.10	5.02	29.74	85.58
	SINGLE SENTENCE	23.92	44.73	28.92	11.81	21.04	**14.04**	20.82	38.60	25.16	4.90	29.94	85.65
	CLAIM-CENTRIC	11.61	33.36	16.44	4.95	14.77	7.08	10.50	30.70	14.97	2.38	19.64	80.99
	ENTITY-CENTRIC	12.35	30.61	16.55	5.13	13.42	7.01	10.91	27.22	14.65	2.20	18.07	82.11
GPT-3	UNCONTROLLED	19.17	48.22	24.77	7.17	19.69	9.51	16.09	40.49	20.76	3.09	27.60	86.77
	TOKEN-LIMIT	**26.29**	40.66	28.92	9.75	15.03	10.59	**22.24**	34.07	24.34	3.75	26.25	87.11
	ABSTRACTNESS	15.74	54.44	23.27	6.14	22.29	9.17	12.52	44.37	18.63	2.84	29.50	86.91
	SINGLE SENTENCE	22.81	44.71	28.46	8.78	17.72	11.07	18.60	37.37	23.43	4.27	29.02	**87.85**
	CLAIM-CENTRIC	14.03	**56.41**	21.72	5.53	**23.93**	8.68	11.11	**46.18**	17.33	2.59	25.88	86.50
	ENTITY-CENTRIC	18.92	53.52	26.34	7.74	22.41	10.82	15.26	43.90	21.34	3.67	**30.98**	87.16

Table 5.6: Zero-shot prompt-tuning results for T5 and GPT-3 on our dataset CLAN.

to using in-context learning with a limited number of examples, as shown in Table 5.4. We can see that the models exposed to in-context examples showcase superior performance, highlighting their efficacy in capturing task-specific patterns. While the trained models exhibit excellence in lexical metrics, their performance in semantic metrics is noticeably lower. Notably, $BART_{LARGE}$, trained on our dataset, outperforms other trained models by sizable margins. These results strongly underline that, within the realm of LLMs, incorporating prompt tuning with in-context learning holds more promise, leading to enhanced generalization capabilities.

R3. Do models demonstrate inherent proficiency in generating normalized claims with minimal or no prior training? We examine the potential benefits of zero-shot and few-shot learning to investigate the inherent proficiency in generating normalized claims. Table 5.5 shows the zero-shot and the few-shot results. Zero-shot learning, which relies solely on the pre-trained language model without any task-specific fine-tuning, performs quite well. On the other hand, few-shot learning does not result in significant improvements. Surprisingly, the models trained using few-shot learning perform slightly worse than zero-shot learning, where the models have no exposure to task-specific data. After training on ten examples, the performance of FLAN-$T5_{LARGE}$ drops by 6 BERTScore points absolute, and it continues to decline as more examples are provided. See Table 5.7 for 50-shot and for 100-shot results. This unexpected result suggests that few-shot learning may be unsuitable for this intricate and complex task. The limited number of examples provided during few-shot learning may have been insufficient for the models to generalize and capture the underlying patterns of normalized claims effectively. Moreover, introducing task-specific data might have introduced conflicting information as these models were never trained on this task, leading to a degradation in performance.

5.6 Error Analysis

To comprehend the performance of CACN, we strive to qualitatively analyze the errors committed by our model in this section. Table 5.8 shows some randomly selected instances from our test dataset, along with

Training Samples	Model	ROUGE-1			ROUGE-2			ROUGE-L			BLEU-4	METEOR	BERTScore
		P	R	F1	P	R	F1	P	R	F1			
50	$T5_{BASE}$	21.57	44.24	27.46	9.88	20.36	12.58	18.32	37.78	23.41	4.05	28.64	84.97
	$T5_{LARGE}$	22.00	45.86	28.11	10.24	21.11	13.03	18.71	38.83	23.90	4.22	29.49	85.11
	$BART_{BASE}$	20.10	53.16	25.65	10.15	25.02	12.61	17.35	44.93	22.00	4.11	29.66	84.71
	$BART_{LARGE}$	19.63	52.96	25.20	9.78	24.73	12.21	16.94	44.88	21.61	3.97	29.46	84.65
	FLAN-$T5_{BASE}$	23.16	42.87	28.14	10.62	19.60	12.91	19.92	36.79	24.26	4.43	28.30	85.35
	FLAN-$T5_{LARGE}$	22.38	29.63	19.97	10.69	13.25	9.51	20.49	25.92	17.84	3.03	18.92	77.84
	PEGASUS	15.24	40.28	20.83	7.15	18.29	9.58	13.38	35.84	18.34	3.44	24.71	80.02
100	$T5_{BASE}$	21.55	44.24	27.45	9.87	20.36	12.57	18.31	37.78	23.40	4.04	28.65	84.97
	$T5_{LARGE}$	21.97	45.72	28.06	10.24	21.06	13.02	18.69	38.78	23.86	4.22	29.47	85.11
	$BART_{BASE}$	20.09	53.28	25.65	10.15	24.97	12.60	17.35	44.91	22.00	4.11	29.61	84.72
	$BART_{LARGE}$	19.68	53.00	25.26	9.80	24.72	12.22	16.99	44.94	21.67	3.97	29.43	84.66
	FLAN-$T5_{BASE}$	23.19	42.91	28.18	10.65	19.63	12.94	19.96	36.85	24.32	4.43	28.29	85.36
	FLAN-$T5_{LARGE}$	22.38	29.63	19.97	10.69	13.25	9.51	20.49	25.92	17.84	3.03	18.92	77.85
	PEGASUS	15.30	40.30	20.88	7.16	18.27	9.58	13.41	35.89	18.37	3.43	24.74	80.02

Table 5.7: Few-shot results on our dataset CLAN.

	Social Media Post	Normalized Claim	BS
Sample 1	They gave you a MASK to cover your face AND then gave you George Floyd's signature phrase 'I can't breath' .... Yet 2.5 years later, most still struggle with the realization that EVERYTHING on TV is fake. Floyd and ConVid never existed	**GOLD:** George Floyd and COVID-19 'never existed'.	–
		BART: George Floyd died of COVID-19	90.33
		DIRECT: George Floyd's death highlighted the lack of understanding of the reality of events portrayed on television.	86.84
		CACN: George Floyd and Covid never existed.	90.98
Sample 2	Good News. Finally, CAJY VAZ, an Indian EX Student from St. Xavier's High School from Mumbai, presently based in Goa, found a home remedy for *Covid 19*, which was approved by the WOH for the first time. He proved that *one teaspoon of pepper powder, two teaspoons of honey, a little ginger juice taken for 5 consecutive days can eliminate the effect of corona up to 100%*. The whole world is starting to take this precautionary treatment, finally a happy experience of 2021.	**GOLD:** Honey, Ginger, Black Pepper can cure Coronavirus.	–
		BART: One teaspoon of pepper powder, two teaspoons of honey, a little ginger juice taken for 5 consecutive days can eliminate the effect of Covid up to 100%.	84.89
		DIRECT: Indian ex-student Cajy Vaz found a home remedy of pepper powder, honey and ginger juice that can reduce the effects of Covid-19 up to 100%.	84.65
		CACN: A home remedy of pepper powder, honey, and ginger juice can eliminate the effect of COVID-19 up to 100%.	78.47
Sample 3	Zelensky sold 17 million hectares of agricultural land to Monsanto, Dupont, and Cargill. Yes, you read it well, 17 million hectares to GM0/chemical companies. This is very bad for the entire world since Ukraine is the largest exporter of wheat and other grains. Zelensky sold 28% of the entire Ukrainian arable land. Australian National Review reports three major US cross-border consortiums have bought 17 million hectares of Ukrainian farmland. To compare: In all of Italy, there are 16.7 million hectares of agricultural land. It turns out that three American companies in Ukraine bought more valuable agricultural land than in all of Italy. The entire area of Ukraine - 600,000 sqm, 170,000 sqm built.	**GOLD:** Ukraine President Volodymyr Zelenskyy sold 17 million hectares of farmland to US corporations.	–
		BART: Three American companies have bought 17 million hectares of farmland in Ukraine.	88.03
		DIRECT: Three major US companies have bought 17 million hectares of Ukrainian agricultural land, which is more than in all of Italy. These companies are Cargill, Dupont and Monsanto, with their main shareholders being American venture capitalists Blackrock, Vanguard and Blackstone.	83.57
		CACN: Ukrainian President Zelensky sold 17 million hectares of agricultural land to Monsanto, Dupont, and Cargill.	90.48

Table 5.8: Examples of generated normalized claims along with the gold reference. BART refers to $BART_{LARGE}$.

gold normalized claims and predictions from CACN. For comparison, we also show predictions from two best-performing baselines, $BART_{LARGE}$ and DIRECT.

Naturally, the predictions in the fine-grained analysis are much more intricate than in the coarse-grained quantitative setup. During our manual qualitative analysis, we unveiled several interesting patterns and errors in the generated responses. For example, although $BART_{LARGE}$ generated responses with a high BERTScore in example 1, we noticed that the factual alignment is incorrect, making this model untrustworthy for downstream tasks such as claim check-worthiness and claim verification. In contrast, our proposed model produced a response that is both correct and precise. The response generated by DIRECT is also accurate, but it is excessively long, which contradicts the objective of the normalized claims

Model	Fluency	Coherence	Relevance	Consistency	SC
BART	3.44	3.74	3.66	3.82	3.77
DIRECT	4.48	4.58	4.03	4.26	4.38
CACN	**4.59**	**4.63**	**4.17**	**4.34**	**4.39**

Table 5.9: Human evaluation on the generated normalized claims. SC denotes self-contextualization, while BART refers to BART$_{LARGE}$.

being concise and straightforward. This problem is also evident in example 3, where DIRECT produces a factually correct claim but is overly long. In example 2, we observe that the BART$_{LARGE}$ model demonstrates the lowest number of hallucinations and adheres closely to the input social media post. In contrast, our model's BERTScore performed the worst for this example. However, upon closer inspection, we noticed that the normalized claim that our model generated was indeed correct and most relevant for fact-checking. These findings highlight the complexity and the trade-offs involved in generating normalized claims. While certain models may excel in certain cases, there is often a compromise in other aspects, such as factual accuracy and conciseness.

Human Evaluation. We conducted an extensive human evaluation to assess the linguistic proficiency of the generated normalized claims. Building upon the measures proposed by van der Lee et al. (244), we evaluated the generated claims based on four aspects: fluency, coherence, relevance, and factual consistency. We further introduced the parameter of self-contextualization to measure the extent to which the normalized claims include the necessary context for fact-checking within themselves. Each of these measures played a unique and vital role in evaluating the quality of the generated claims. We formally define the five human evaluation measures as follows:

- *Fluency*: It measures the linguistic proficiency exhibited by the generated responses.
- *Coherence*: It evaluates the intrinsic structure and the organization of the generated normalized claims.
- *Relevance*: It appraises the discerning selection of contextually appropriate content within the generated response.
- *Factual consistency*: It examines the intricate alignment between the factual accuracy of the generated response and the source text.
- *Self-contextualization*: It measures the extent to which the normalized claims includes the necessary context.

To conduct the evaluation, we randomly selected 50 instances from our test set and assigned five human evaluators to rate every normalized claim on a scale of 1 to 5 for each of these five aspects. All evaluators were fluent English speakers with a Bachelor's or Master's degree. To ensure reliability, each example was evaluated by all five evaluators independently, and then we averaged their scores. The average scores of human evaluation are presented in Table 5.9. For comparison, we also included the results from the best-performing baseline systems, namely BART$_{LARGE}$ and DIRECT. Our analysis reveals that the outputs generated by CACN exhibit qualitative superiority compared to the baseline systems across all dimensions.

5.7 Summary

We introduced the novel task of *claim normalization*, which holds substantial value on multiple fronts. For human fact-checkers, claim normalization is a useful tool that can assist them in effectively removing unnecessary texts from subsequent processing. This also benefits downstream tasks such as identifying previously fact-checked claims, estimating claim check-worthiness, etc. We further compiled a dataset of social media posts comprising over 6k posts and their normalized claims. We further benchmarked this dataset with a novel approach, CACN, and showed its superior performance compared to different state-of-the-art generative models across multiple assessment measures. We also documented our data collection process, providing valuable insights for future research in this domain. In future work, we plan to extend the dataset, including with new languages. We also plan to use more powerful LLMs. While our study has made major contributions to claim normalization, it is critical to recognize and address its potential limitations. During our data collection process, we excluded claims about images and videos. Yet, we believe that including multimodal information may help improve claim normalization. Another key problem is that each fact-checking organization adheres to its own set of editorial norms, procedures, and subjective interpretations of claims. These variations in writing style and judgments make it challenging to establish a standardized claim normalization. Addressing this issue will necessitate attempts to develop consensus or guidelines among fact-checking organizations in order to ensure greater consistency and coherence in claim normalization. By acknowledging and addressing these limitations, we may endeavour to improve the reliability and soundness of claim normalization systems in the future.

Part 3

Identifying Check-Worthy Claims

6. CLAIM CHECK-WORTHINESS

Determining the check-worthiness of claims is an integral process in fact-checking. Despite ongoing advancements in claim check-worthiness detection, there remains a notable transparency gap in elucidating why a claim is deemed check-worthy. Addressing this gap, we introduce an explainable approach to claim check-worthiness by incorporating *'rationality labels.'* Closely mirroring the complex decision-making process of human cognition, these labels evaluate claims on multiple aspects, including factual verifiability, societal impact, and the likelihood of causing public unrest. Unlike existing systems that provide simplistic binary judgments on check-worthiness, our framework, `CheckMate`, delves into the nuanced human-like reasoning behind the assessment of claim check-worthiness for fact-checking. We also introduce `CheckIt`, the first claim check-worthiness dataset of 5920 tweets with rationality labels. This multi-dimensional approach offers a significant advancement over traditional binary evaluation systems. We compare our proposed approach with several baseline systems and comprehensively analyze the results, including quantitative and qualitative comparisons.

6.1 Introduction

Rapid evolution of online social media, driven by human communication needs and technological advancements, has transformed it from a mere leisure activity into a thriving business sector. The freedom of speech and expression provided by these platforms has resulted in a tremendous increase in social media users. However, this unrestricted freedom comes at a high cost of factuality and accountability, allowing the dissemination of misinformation to pose a profound threat to public discourse, democratic processes, and societal cohesion (74; 245). False claims might emanate accidentally, but in most cases, they are actively spread by those who seek to benefit from them. One such instance is the $45th$ Presidential election in the United States. The entire world witnessed the developing influence of fake news and false claims. It was reported that about 25% of Americans visited a fake news website, possibly impacting the election's outcome (217). Another prominent example is the global COVID-19 pandemic. When the entire world went into lockdown, the virtual world encountered a great closeness, transforming social media platforms into the primary conduits for information consumption and dissemination. Around the same time, enormous social media posts with unverified bogus claims about the pandemic began to arise, frequently spurring life-threatening remedies (36). Such claims had an unprecedented impact, resulting in monetary damage and the loss of priceless human lives. A study revealed that at least 800 individuals died worldwide in the first quarter of 2020 due to misinformation about COVID-19 (246).

In response to such events, various initiatives have been taken for fact-checking (140; 100; 106; 247; 248; 249). Unfortunately, the daily publication of millions of blog articles and social media posts makes manual as well as automated fact-checking nearly absurd, necessitating a more refined strategic approach to pinpoint the most heinous claims. For example, consider a post that reads *'Alcohol cures all kinds of illness, including coronavirus.'* can lead to massive re-sharing and consequential unrest. On the other hand, a claim stating, *'I like eating oranges, it contains Vitamin-C.'* is not a check-worthy claim. In such cases, automated filtration of crucial claims for immediate fact-checking could be paramount. Hence, a crucial element within a fact-checking pipeline, including claim detection and claim verification, is the

system that detects claims worthy of fact-checking, which aids in prioritizing claims for verification (142).

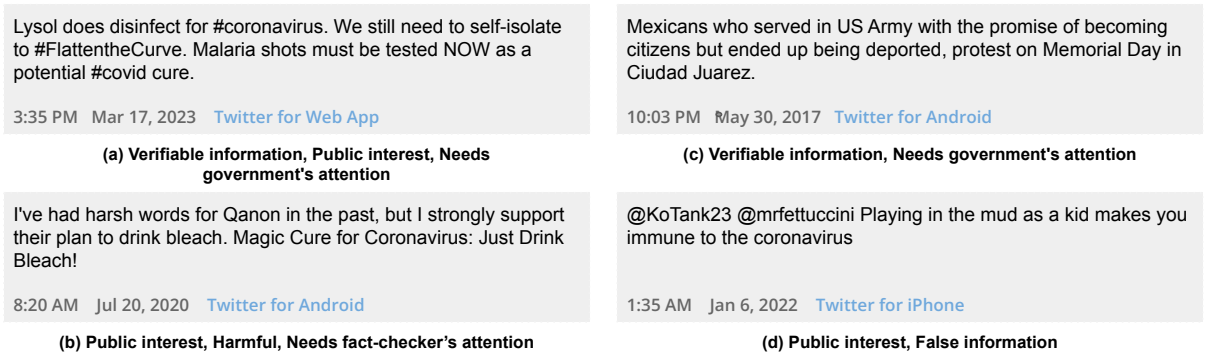


Figure 6.1: Examples of check-worthy tweets, with various rationales justifying their check-worthiness.

Over the past few years, research efforts have been devoted to crafting systems that assess the check-worthiness of claims (140; 139). Numerous studies have pitted these systems, engaging in multiple shared tasks organized by CheckThat! lab (144; 143; 142). Although these systems have demonstrated remarkable prowess in identifying check-worthy claims, they lack explanations about the decisions made. This absence of explanatory insight often necessitates additional time and effort from fact-checkers, making it challenging to understand why a claim is deemed check-worthy. To address this issue, we introduce *'rationality labels'* in claim check-worthiness estimation for enhanced explainability and more informed decisions. These rationality labels focus on three key aspects: (a) Information, (b) Social Impact, and (c) Actionable items. These facets are distilled into a set of six non-disjoint labels – *verifiable information, false information, public interest, harm to society, requiring fact-checker's attention,* and *necessitating government's attention.*

Figure 6.1 demonstrates the varying rationality labels associated with the check-worthiness of claims, influencing their prioritization for fact-checking. Example 1(a) presents a claim, *'Lysol does disinfect for #coronavirus. We still need to self-isolate to #FlattentheCurve. Malaria shots must be tested NOW as a potential #covid cure'*, which is worth fact-checking; on closer inspection, we find the underpinning reasons as verifiable information, public interest, needs government attention. In contrast, the check-worthy claim in Example 1(b) is harmful to society, as it propagates a false cure for COVID-19. Leveraging these rationality labels in claim check-worthiness determination fosters several benefits for various stakeholders, including the general public, fact-checkers, and policymakers. Fact-checking resources are often limited. By determining these rationality labels, fact-checkers can better allocate resources, focusing on claims that have a greater impact or pose a risk to the public. For content moderators using automated tools, these labels can enhance efficiency by offering a deeper understanding of check-worthy labels. This fine-grained assessment can also enable social media platforms to act appropriately toward these claims, promoting transparency.

Task Formulation: The objective is to develop a unified classification system tasked to identify check-worthy claims. Additionally, we seek to identify the underlying rationale behind the decision. We strive to enhance the claim check-worthiness objectives with additional fine-grained rationality labels that would aid in defining the plausible reasons for the check-worthiness determination. Inspired by Alam et al. (147), we devise six rationality labels that align most with the final claim check-worthy decision, keeping in sight the viewpoints of policymakers, society, and manual fact-checkers. Let us consider $D = \{(x_1, y_1, R_1), (x_2, y_2, R_2), ..., (x_n, y_n, R_n)\}$ be the set of n labeled instances. Each instance in the dataset is a triplet where $x_i = \{w_1, w_2, ..., w_m\}$ denotes the input claim comprising of m

words, $y_i \in \{$*check-worthy, non check-worthy*$\}$ denotes the check-worthiness label. Accompanying the check-worthiness label, $R_i = \{$*verifiable information, false information, public interest, harmful, need fact-checkers' attention, government's attention*$\}$ signifies the set of k rationality labels of claim x_i. For $y_i =$ *non check-worthy*, we do not consider rationality labels and hence $R_i = \phi$. These rationality labels are not mutually exclusive and can overlap, reflecting the multifaceted and often interrelated reasoning factors that inform the assessment of a claim's checkworthiness. The objective is to build a multi-label classifier F that maps an input text instance x_i to the set of rationality labels R_i and discerning the overall check-worthy label y_i.

Our Contributions: Through this work, we make the following contributions:

- We propose a novel task of explainable claim check-worthiness and curate a large manually annotated dataset comprising over $5.9k$ claims, with check-worthiness and rationality labels.
- We propose `CheckMate`, a unified framework for explainable claim check-worthiness detection. This unified approach not only produces a claim check-worthiness label but also furnishes accompanying rationality labels, enhancing the depth and transparency of the assessment.
- We present an extensive comparison and error analysis to quantify the efficacy of our proposed approach to the state-of-the-art baseline systems.

6.2 Related Work

Through rapid information release and consumption, social media has greatly aided the democratization of information distribution. Simultaneously, the lag in information verification has become rampant and flooded, severely affecting people's lives, social stability, and even national security. Over 200 fact-checking organizations worldwide, including PolitiFact, Snopes, and Full Fact, have launched initiatives to perform manual fact-verification.[1] Unfortunately, these efforts are inadequate, given the magnitude of misinformation disseminated through multiple online communication channels. To combat this deluge of misinformation, several studies have addressed this issue of fact-checking through several automated downstream tasks (100; 232; 250). Recent studies have focused on various subtasks around fact-checking, ranging from automatic identification of claims to extracting evidence to establish claim veracity. It isn't easy for humans and automated systems to fact-check every claim, which necessitates segregation of check-worthy content online, saving the valuable time and resources of manual fact-checkers and automated systems. Thus, the first and most significant step in fact-checking is determining whether a piece of text is worth fact-checking. The task of identifying check-worthy claims has drawn the attention of numerous researchers (137; 138; 139). Until now, these studies have primarily been divided into the following two categories.

Check-Worthy Claims in Political Debates. The first category focuses on identifying check-worthy claims in debates or political speeches. With a pioneering attempt by Hassan et al. (140), claim check-worthiness detection has emerged as a major research area in recent years. They developed ClaimBuster, the first-of-its-kind system to target a claim's check-worthiness in a debate. It was trained on a massive manually annotated dataset of US election debates from 1960 to 2012, totalling 30 debates and $28,029$ transcribed sentences. Each statement made during a political discussion was categorized into one of three categories: non-factual, unimportant factual, or check-worthy factual. Focusing on the 2016 US Presidential debates, Gencheva et al. (137) obtained binary annotations for check-worthiness from various

[1] `www.politifact.com`, `www.snopes.com`, `www.fullfact.in`

fact-checking organizations. Later, they developed ClaimRank, which was trained on additional data, including Arabic content. Vasileva et al. (139) used a multi-task learning neural network. The task was to predict whether a sentence would be selected for fact-checking by each of the nine different fact-checking organizations, namely CNN, The Washington Post, PolitiFact, FactCheck, ABC, NPR, NYT, Chicago Tribune, and The Guardian. The CheckThat! Lab has organized similar shared tasks since 2018, focusing on political debates and speeches (142; 143; 144).

Research on estimating claim-worthiness in debates and political speeches has been thoroughly studied and made substantial progress. Due to the lack of structure and proper linguistic properties, these approaches do not work well for noisy social media texts.

Check-Worthy Claims in Social Media. The social media expansion has boosted the preponderance of false and misleading claims; as a result, current research has turned to identifying claim check-worthiness in social media. To address this, in their recent editions, CheckThat! Labs organized shared tasks on identifying check-worthy claims with a primary focus on social media texts (142; 143; 144). These shared tasks attracted multiple systems modelled to handle noisy text from social media platforms. The 2020 edition featured three central tasks: detecting previously fact-checked claims, evidence retrieval, and actual fact-checking of claims. Along with English, they offered the tasks in Arabic and Spanish. For Arabic, several teams improved pre-trained models, such as AraBERT and multilingual BERT. In the case of the English task, multiple systems harnessed the strength of pre-trained Transformers, specifically BERT and RoBERTa. Other methods extracted tweet embeddings from tweets using pre-trained models like GloVe and Word2Vec, which were then fed into a neural network or an SVM. The top-ranked system in the 2021 edition also leveraged transformer-based models (142). The first shared task in the 2022 edition was anticipating which Twitter posts should be fact-checked, emphasizing COVID-19 and politics. It was put forth in Arabic, Bulgarian, Dutch, English, Spanish, and Turkish (143). A total of 19 teams competed, with the majority of submissions achieving significant improvements over baselines using Transformer-based models such as BERT and GPT-3. Alam et al. (147) created a multi-question annotation schema of COVID-19 tweets structured around seven claim check-worthiness questions.

While better models for detecting claim check-worthiness are constantly being developed, there is a lack of literature on the explainability aspects of these binary decisions. Inspired by the annotation schema proposed by Alam et al. (147), we take it a step further and attempt to back the binary label decision with rationality labels, forging a new task of fine-grained claim check-worthiness. We define these rationality labels in the following section.

6.3 Dataset

Several check-worthy claim detection datasets have been released in recent years (138; 142; 143). Nevertheless, each classifies the entire sentence or document as check-worthy and does not furnish reasonable grounds to quantify a claim as check-worthy. To fill this void, we propose `CheckIt`, a large-scale manually annotated Twitter corpus with check-worthiness and rationality labels. We provide a brief overview of our proposed taxonomy for rationality labels, followed by a detailed description of the annotation process.

6.3.1 Rationality Labels

We develop the initial version of a taxonomy designed to categorize rationality labels for the task of explainable claim check-worthiness. Drawing inspiration from the well-founded literature provided by

Alam et al. (147), we structure our taxonomy to align with their established theoretical framework. We present a detailed overview of the proposed rationality labels and their definitions and examples in Table 6.1. We focus on identifying various rationality labels that offer insights into the different dimensions and reasoning factors behind a check-worthy label. These labels indicate the multifaceted aspects contributing to the overall evaluation of a claim's check-worthiness and relevance in fact-checking. We formally define these labels as follows:

- R_1**. Verifiable information:** This label is affixed to claims containing factual details that can be verified, such as definitions, statistics, and law references, prompting a closer inspection of their accuracy.
- R_2**. False information:** Statements falling under this category are scrutinized to ascertain whether they seem to present inaccuracies, contributing to determining their check-worthiness.
- R_3**. Public interest:** Claims under this label encompass information of broad public concern, including details about potential COVID-19 cures, updates on case numbers, government measures, and discussions involving rumors or conspiracy theories.
- R_4**. Harmful:** This label is applied to claims that could potentially cause harm to society, prompting a focused investigation.
- R_5**. Needs fact-checker's attention:** Statements in this category demand the expertise of a qualified fact-checker for verification, signifying a nuanced level of complexity beyond the purview of novice verification.
- R_6**. Needs government's attention:** Claims designated under this label trigger deliberation on whether the target text warrants bringing to the attention of government entities, underscoring their potential impact on a larger scale.

In essence, our proposed rationality labels are a comprehensive exploration into the specific attributes that render a claim check-worthy, providing an in-depth understanding of its rationale. Furthermore, these rationality labels encapsulate varied dimensions, and hence, they are not mutually exclusive.

	Rationality Label	Definition	Example
Information	R_1. **Verifiable information**	Containing verifiable factual details, such as definitions, statistics, or law references.	*Approximately 40% of the global population uses Twitter.*
	R_2. **False information**	Statements or claims that are subject to scrutiny to determine if they contain inaccuracies	*Drinking bleach can cure COVID-19.*
Social Impact	R_3. **Public interest**	Information relevant to the broad public, encompassing topics such as health, safety, and government measures.	*Forty-one laborers trapped in a tunnel in Uttarakhand for 17 days have been rescued.*
	R_4. **Harmful**	Claims that could potentially cause harm to individuals, communities, or society at large.	*Skipping vaccines is the key to natural immunity. Say no to vaccines!*
Actionable	R_5. **Needs fact-checker's attention**	Claims demanding the expertise of a qualified fact-checker for the verification.	*COVID-19 is the same as the common cold or the seasonal flu.*
	R_6. **Needs government's attention**	Statements that merit attention from authorities who can address or manage the potential consequences at a larger scale.	*Critical health crisis ignored by government officials! They're hiding the truth.*

Table 6.1: Comprehensive taxonomy for rationality labels in the explainable check-worthiness task, including definitions and examples.

6.3.2 Collection

We source the tweets from our Twitter dataset CURT, which provides tweets marked with claim spans.[2] According to our observations, each claim within a tweet has a different relevance and must be validated separately. Thus, we annotate each claim rather than the entire tweet for our task.

6.3.3 Annotation

We progressed through three iterations of refinements to attune the dataset better. Three annotators carried out the annotations. During the first iteration, each annotator assesses a batch of 50 randomly selected claims and labels them for rationality and check-worthiness labels. Following that, in the second iteration, they collectively discuss and resolve ambiguous cases and annotate an additional set of 50 samples. Finally, they annotate 100 claims during the final phase of our pilot annotation. After the final iteration, we obtain an average inter-annotator agreement score of 0.73 using the Cohen Kappa. Some annotated samples from our dataset, CheckIt, are shown in Table 6.2. The first example contains two check-worthy claims within the same tweet, but the rationales for their check-worthiness vary considerably. A closer examination reveals that the latter claim has a more detrimental effect on society than the former. The second tweet contains three claims, only two of which are worth fact-checking. The first one, *'#ozone is being used to destroy #COVID-19,'* is a verifiable claim of significant public interest. It should be fact-checked as quickly as possible as it discusses COVID-19 protection. The second claim also includes a verifiable claim, but it may not be of broad interest to the general public.

Tweet	**Claim**	**Rationality Labels**	**Check-Worthy**
Healthy people don't spread #covid_19, people who have recovered and are immune don't spread the virus. Do you know who else doesn't spread the virus? People wearing masks. #Masks4All	Healthy people don't spread #covid_19	*false information, public interest*	✓
	people who have recovered and are immune don't spread the virus	*verifiable information, false information, public interest, harmful, needs fact-checker's attention*	✓
#ozone is being used to destroy COVID19. Vaccines for RNA viruses cause mutation, so if you want a worse pandemic than this or to die from an injection, that's all that will be on offer. There's a known cure.	#ozone is being used to destroy COVID-19	*verifiable information, false information, public interest, harmful, needs fact-checker's attention*	✓
	Vaccines for RNA viruses cause mutation	*verifiable information, needs fact-checker's attention*	✓
	there is a known cure		✗
@username From what I've read, the coronavirus is no more dangerous than the common cold or the seasonal flu. Symptoms are usually so mild that one doesn't even notice they are sick. The people who run into problems are the immune-compromised (like the very old or the already ill)	coronavirus is no more dangerous than the common cold or the seasonal flu	*verifiable information, false information, public interest, harmful, needs fact-checker's attention*	✓
	Symptoms are usually so mild that one does not even notice they are sick	*verifiable information, public interest, needs fact-checker's attention*	✓
	The people who run into problems are the immune-compromised		✗

Table 6.2: A few examples from our dataset CheckIt, labeled for claim check-worthiness and rationality labels.

[2] A *claim span* is the specific phrase that makes an assertion. For ease, we will use *claim* to denote these *claim span* within the tweets.

6.3.4 Statistics and Analysis

We annotated 5920 claims in total. We split `CheckIt` into three sections – training set, validation set, and test set – with a 70:15:15 split. The summary of dataset statistics is shown in Table 6.3. The statistics of six rationality labels are depicted in Table 6.4. Each rationality label has a broadly consistent distribution, except for *harmful* and *needs government's attention* R_6, where the distribution is slightly skewed towards *No*. The slightly skewed distribution towards *No* label in *harmful* indicates that most online claims may not seen as directly harmful. This implies that a significant portion of the questionable claims may not be flagged immediately by content moderators and require more nuanced analysis besides the harmfulness. The label *needs government's attention* examines whether a claim requires government intervention. It is critical to recognize that most of the misinformation on the internet targets the general public rather than necessitating direct government intervention. Thus, the distributions seen in these two labels reflect these inherent biases observed in the context of social media platforms.

Dataset	Check-worthy	Non check-worthy
Training	2677	1467
Validation	574	314
Testing	573	315
Total	3824	2096

Table 6.3: Dataset statistics of `CheckIt`.

Rationality Label	Training		Validation		Testing	
	Yes	No	Yes	No	Yes	No
Verifiable information	2336	1808	509	379	516	372
False information	2535	1609	602	286	662	226
Public interest	2910	1234	609	279	630	258
Harmful	1690	2454	358	530	438	450
Needs FC's attention	2076	2068	380	508	443	445
Needs govt's attention	1711	2433	368	520	416	472

Table 6.4: Statistics on six rationality labels. FC denotes fact-checker. These rationality labels are not mutually exclusive.

6.3.5 Comparison with Existing Datasets

Existing datasets in computational argumentation cover a wide range of aspects of claim analysis. For instance, the LESA dataset (251) focuses on claim detection on Twitter, comprising approximately 9894 instances. FEVER dataset (252), with a substantial scale of $185,445$ claims, involves altering sentences from Wikipedia and verifying them without knowledge of their original context. They make valuable contributions to automatic claim verification work but cannot replace real-world datasets. MultiFC dataset (106) aims to address the limitations of artificially constructed claims by curating naturally occurring claims from 26 fact-checking websites in English. It includes $34,918$ claims, along with evidence pages and contextual information. Additionally, a new task introduced by Sundriyal et al. (232) produces CURT, a large-scale Twitter corpus featuring token-level claim spans in over $7,500$ tweets.

Datasets for claim check-worthy estimation from the CheckThat! shared tasks are the closest counterparts to our work (142; 143). However, these datasets for social media claims are much smaller and do not include justifications or rationals for the checkworthiness label. On the other hand, our proposed dataset, `CheckIt`, provides fine-grained rationality labels to approximately $5.9k$ claims, making it the first

of its kind. Table 6.5 compares our proposed dataset CheckIt with other publicly available claim-related datasets.

Dataset	Instances	Source	Granularity	Task
Student Essays (163)	90	Student essays	Document	Claim Detection
LESA (251)	9894	Tweets	Sentence	Claim Detection
FEVER (100)	185k	Wikipedia	Sentence	Claim Verification
MultiFC (106)	36534	Fact-checking sites	Sentence	Claim Verification
CURT (232)	7555	Tweets	Phrase	Claim Span Identification
CLEF-2021 Task 1A (142)	1312	Tweets	Sentence	Claim Check-worthiness
CLEF-2022 Task 1A (143)	3040	Tweets	Sentence	Claim Check-worthiness
CheckIt	5920	Tweets	Phrase	Fine-grained Check-worthiness

Table 6.5: Description of existing benchmark claim datasets and task compared with our proposed dataset, CheckIt.

6.4 Methodology

Previously, the narrative on claim check-worthiness detection was primarily contextual or linguistic. We propose an integrated model, CheckMate, to achieve our goal of utilising both. It combines linguistic features obtained from part-of-speech tags and dependency parsing with contextual factors extracted from the transformer-based model, BERT. A high-level architecture of our proposed model is shown in Figure 6.2. It has two main components – CoNet, consisting of the BERT framework, a module to optimize six attention heads of BERT (one attention head per rationality label) and obtain the contextual features, and LiNet to incorporate linguistic features emanating from POS tags and dependency trees. The following subsections provide a formal task definition of the fine-grained extension of the claim check-worthiness problem and intricate details of the aforementioned modules of the proposed model.

Figure 6.2 depicts the high-level architecture of our proposed model, CheckMate, for explainable claim check-worthiness. The model consists of two primary modules: a Contextual Network (CoNet) and a Linguistic Network (LiNet). CoNet is a transformer-based architecture intended to extract contextual elements of the text while also optimizing the six attention heads of BERT corresponding to the six rationality labels. LiNet considers linguistic features retrieved from part-of-speech (POS) tags and dependency trees (Dep).

6.4.1 Contextual Network

To extract the contextual features of the input claim, the CoNet module includes a BERT layer followed by a self-attention layer. We begin by pre-training BERT with our claim check-worthy corpus and then compute six attention vectors using BERT's six attention heads to fine-tune BERT during training. We hypothesize that assigning one attention head to each rationality label will assist the model in catering to each label specifically.

To begin with, we tokenize every i^{th} input text $x_i = \{w_1, w_2, ..., w_m\}$. x_i is transformed into a high-dimensional vector through an embedding function E, where $E(x_i) \in \mathbb{R}^d$ and d represents the embedding dimension. To preserve the language's sequential nature, we incorporate positional encodings P into these embeddings, effectively allowing the model to maintain awareness of the tokens' positions within the sequence. CoNet's multi-headed attention mechanism uses learned projection matrices W^Q,

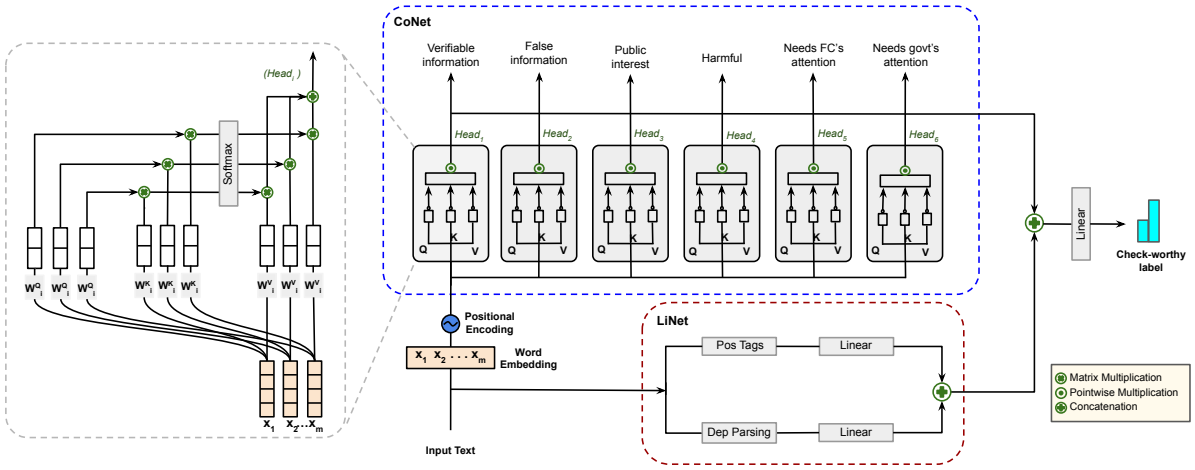


Figure 6.2: The architecture of our proposed model, `CheckMate`, for detecting fine-grained claim check-worthiness. The model is divided into two modules: CoNet and LiNet. CoNet aims to capture contextual features of the input text, whereas LiNet seeks linguistic features.

W^K, and W^V for each head j to create interaction between queries Q, keys K, and values V. The attention function for head j is computed as follows:

$$\text{Att}_j(Q, K, V) = \text{softmax}\left(\frac{QW_j^Q \cdot (KW_j^K)^T}{\sqrt{d_k}}\right) VW_j^V \tag{6.1}$$

where d_k is the dimensionality of the key vectors. The outputs from each attention head are concatenated to generate a unified contextual representation C for the entire sequence.

6.4.2 Linguistic Network

Building upon the utilization of linguistic features for better comprehending claims as shown by Gupta et al. (251), we also incorporate linguistic features into our model. The LiNet module seeks to extract linguistic information from the input text by employing two critical linguistic constructs: part-of-speech tags and dependency parse trees. For each i^{th} input text $x_i = \{w_1, w_2, ..., w_m\}$, we acquire a series of associated POS tags, yielding the set $\{p_1, p_2, ..., p_m\}$. After that, dependency parsing is used to abstract the input text's grammatical structure. It discovers the related words and derives a directed relation, $d(w_j, w_k)$, between any two tokens w_j and w_k, in which w_j is the head and w_k is the dependent.[3] These vectors are then processed through their respective linear layers, and the resulting transformed vectors are concatenated.

Following this, we integrate contextual features retrieved from self-attended vectors of CoNet along with the linguistic features obtained from LiNet. The aggregated feature vector is subjected to a linear transformation L, projecting it onto the label space y, i.e. check-worthy or non check-worthy.

[3]These dependency relations are obtained using Python's SpaCy library.

6.5 Experiments and Results

Experimental Setup. We set the maximum length of input posts to 128 tokens. We optimize the performance of `CheckMate` on the validation set by adjusting certain hyperparameters. After careful consideration, we select the optimum configuration for all hyperparameters. This configuration includes a dropout rate of 0.25, a learning rate of 0.001, and a batch size of 8. The model is trained for a maximum of 20 epochs, using early stopping with a patience of 5. We utilize *Adam* as the optimizer, employing a decay rate of 0.001, and a linear scheduler with warm-up. For baseline systems, we maintain consistency by keeping most of the hyperparameters the same as mentioned in the original papers.

Evaluation Metrics. We select F1 scores as the primary metric for evaluating claim check-worthiness because they offer a balanced assessment of a model's precision and recall. To better comprehend the performance of the systems to detect check-worthy claims, we also report accuracy and class-wise F1 scores. We hinge on the macro-F1 score corresponding to each rationality label.

Baseline Systems. To evaluate the performance of our proposed system, `CheckMate`, we conduct experiments under two setups. In the coarse-grained setup, we compare binary claim check-worthiness predictions with state-of-the-art systems. In the fine-grained setup, we treat six rationality labels (R_1–R_6) as independent classification tasks and benchmark against several traditional models. We compare `CheckMate` against the following baselines: ▷ **AI Rational** (253): Ranked first at CLEF-2022, using fine-tuned transformer models (RoBERTa-large performed best). ▷ **PoliMi FlatEarthers** (254): Ranked third at CLEF-2022 using a fine-tuned GPT-3 model. ▷ **NLPIR@UNED** (255): CLEF-2021 winner with a transformer-based approach; BERTweet yielded best results. ▷ **Fight for 4230** (256): CLEF-2021 runner-up using BERTweet with custom preprocessing (e.g., link removal, normalization). ▷ **Accenture** (113): CLEF-2020 winner; their top model used RoBERTa with pooling and classification layers. ▷ **Team Alex** (115): Second place at CLEF-2020 with preprocessing-focused approaches, including hashtag normalization. ▷ **BERT** (160): Fine-tuned for binary claim check-worthiness detection. ▷ **RoBERTa** (114): Fine-tuned robust BERT variant for the same task. ▷ **XLNet** (182): An auto-regressive transformer model, also fine-tuned for our task. ▷ **BERTweet**: A RoBERTa-based model pre-trained on 850M English Tweets, fine-tuned for check-worthiness detection.

Apart from the existing state-of-the-art systems, we also explore several traditional machine learning models, like Logistic Regression (LR), Multinomial Naive Bayes (MNB), Random Forest (RF), and Support Vector Machine (SVM). We leverage the pre-trained model of Word2Vec to obtain embeddings for the tweets, which we feed into these models.

Experimental Results. Our experiments reveal that `CheckMate` outperforms all baselines across most evaluation measures. We further examine all systems aiming to answer the research questions listed below.

R1. How effectively does the `CheckMate` detect check-worthy claims compared to existing baseline systems? Even for humans, let alone automated systems, it can be challenging to discern check-worthy claims due to the high degree of subjectivity. We summarize the aggregated results on the test set in Table 6.6. Evidently, `CheckMate` effectively outperforms all existing baseline systems. When considering binary check-worthiness labels, `CheckMate` achieves the highest macro-F1 score of 78.02% compared to other baseline systems. Most of the baseline systems have low F1 scores for non-check-worthy claims. However, our model receives an F1-score of 70.45% for the non check-worthy class, indicating strong

Model	Check-worthiness Label				Rationality Labels					
	Acc	***m*-F1**	***cw*-F1**	***ncw*-F1**	R_1	R_2	R_3	R_4	R_5	R_6
AI Rational	0.7905	0.7606	0.8453	0.6760	-	-	-	-	-	-
PoliMi FlatEarthers	0.7579	0.7292	0.8173	0.6411	-	-	-	-	-	-
NLP&IR@UNED	0.7905	0.7674	0.8408	0.6941	-	-	-	-	-	-
Fight for 4230	0.7860	0.7585	0.8401	0.6769	-	-	-	-	-	-
Accenture	0.7939	0.7647	0.8476	0.6817	-	-	-	-	-	-
Team Alex	0.7872	0.7653	0.8369	0.6937	-	-	-	-	-	-
LR	0.7570	0.7055	0.8287	0.5823	0.7313	0.6816	0.7640	0.7362	0.6913	0.7296
MNB	0.7620	0.6960	0.8376	0.5544	0.7183	0.6499	0.7624	0.7136	0.6968	0.7293
RF	0.7430	0.6981	0.8145	0.5817	0.7109	0.6743	0.7689	0.6863	0.6606	0.7110
SVM	0.7699	0.7347	0.8314	0.6053	0.7359	0.7010	0.7743	0.7277	0.6862	0.7247
BERT	0.7714	0.7443	0.8275	0.6611	0.7346	0.6933	0.7857	0.5379	0.7021	0.7208
RoBERTa	0.7838	0.7572	0.8376	0.6768	0.7723	**0.7855**	0.7984	0.7866	0.7497	0.7182
XLNet	0.8018	0.7778	0.8508	**0.7047**	0.7782	0.7631	**0.8050**	0.7527	0.7444	0.7501
BERTweet	0.7894	0.7681	0.8384	0.6979	0.7700	0.7684	0.7845	0.7740	0.7556	**0.7666**
GPT-3	0.7568	0.7205	0.8212	0.6197	0.3503	0.4885	0.5772	0.4341	0.6545	0.5337
CheckMate	**0.8063**	**0.7802**	**0.8559**	0.7045	**0.7858**	0.7733	0.7985	**0.8095**	**0.7658**	0.7421

Table 6.6: Performance comparison of CheckMate using our dataset, CheckIt. For binary classification, we report accuracy (Acc), macro-F1 (*m*-F1), and class-wise F1 scores for check-worthy (*cw*) and non check-worthy (*ncw*) classes. For rationality labels, we report F1 scores for each label (R_1-R_6). The **bold** indicates the best, while underlined numbers represent the second-best results for each metric.

performance in the minority class as well. CheckMate accounts for a 0.38% boost in macro-F1 score over the second-best performer, XLNet. There is also a considerable gain in accuracy – our model has an accuracy of 80.63%, which is 0.56% higher than the second-best performer, XLNet. A palpable gain of about $5 - 10\%$ across all metrics is ascertained when we transit from the classical machine learning architectures to the transformer-based models of BERT, RoBERTa, etc. This, in turn, emphasizes the significance of employing contextual word embeddings and transformer-based architectures for the task at hand.

R2. Does joint learning improve performance across rationality labels compared to traditional classification models? To assess the effectiveness of joint learning of our model on rationality labels, we compare it with eight traditional classification baselines. For this evaluation, we treat the six rationality labels as individual classification tasks within the baseline model. The results of this comparison are provided in the last six columns of Table 6.6. The results show that no single model consistently outperforms all six rationality labels; however, our model performed well for most of them. Our model, CheckMate, achieved the highest scores on three rationality labels (R_1, R_4, and R_5) and had competitive results on R_2 and R_3. These experimental results underscore the importance of considering label-specific factors when choosing the most suitable model for the given fine-grained labelling task.

R3. Compared to dedicated models, how does LLM perform in check-worthiness and rationality labelling for social media posts? We investigate the check-worthiness of social media posts in a fine-grained manner using a robust large language model, specifically GPT-3, via the OpenAI API.[4] We

[4] https://platform.openai.com/docs/api-reference

task the model to determine whether *the provided post is worth fact-checking or not*. In the event of an affirmative response, we further sought plausible rationales from the set of suggested rationality labels that we provided. This comprehensive approach aims to leverage the language model's capabilities to assess the credibility of social media posts and provide insightful justifications for the determination. However, our findings revealed a substantial performance gap between GPT-3 and our model. GPT-3 produced a macro-F1 score of 72.05% for claim check-worthiness; our model outperformed it with a huge margin, scoring 78.02%. Furthermore, for all rationality labels, GPT-3 performed poorly. It struggled most with label R_1, representing verifiable information, recording a macro-F1 score of 35.03%, the lowest of any baseline. These stark disparities highlight the limitations of using large language models for subjective and nuanced tasks like check-worthiness estimation. These findings highlight the need for sophisticated dedicated models designed specifically for tasks of this nature.

Ablation Study on Rationality Labels. We perform an ablation study to ascertain the impact of the absence of each rationality label on our model's performance in categorizing claims into check-worthy and non check-worthy classes. Table 6.7 presents the results of the ablation study when one of the six rationality labels (R_1 through R_6) is excluded from `CheckMate`. The scores clearly demonstrate that each label has a significant influence on the overall performance of the final check-worthiness estimation. Removing these labels negatively affects the systems. By excluding R_1, which assesses the verifiability of claims, the model achieves an accuracy of 78.27%, and the macro-F1 score 76%. The influence is moderate, suggesting that R_1 plays a balanced yet significant role in the model. However, eliminating R_2, specifically targets *'false information,'* significantly impairs the model's performance. The accuracy decreases to 59.80%, while the macro-F1 score reduces to 57.19%. This suggests that the presence of R_2 is crucial for detecting erroneous information, and its absence greatly weakens the model's capacity to categorize and handle check-worthy claims appropriately. Other labels also have a substantial impact on the model's efficacy. This emphasizes the essential importance of each rationality label and offers useful suggestions for improving and strengthening the model's capacity.

	Acc	***m*-F1**	***cw*-F1**	***ncw*-F1**
`CheckMate`	0.8063	0.7802	0.8559	0.7045
$-R_1$	0.7827	0.7600	0.8338	0.6862
$-R_2$	0.5980	0.5719	0.6775	0.4664
$-R_3$	0.7444	0.6964	0.8171	0.5757
$-R_4$	0.6273	0.6179	0.6777	0.5581
$-R_5$	0.7162	0.7064	0.7600	0.6529
$-R_6$	0.6700	0.6699	0.6777	0.6621

Table 6.7: Ablation result for `CheckMate`. The symbol $(-)$ signifies the absence of the respective rationality label from the main model.

6.6 Error Analysis

Due to the highly subjective nature of claims, detecting check-worthy claims over social media platforms poses a difficult challenge. To dive deep into the errors encountered by the models, we perform a qualitative analysis of the error cases. Table 6.8 shows a few randomly selected error instances from the test dataset along with their ground labels and output labels as predicted by `CheckMate`. To compare, we also examine predictions from the best-performing baseline, XLNet. Both XLNet and `CheckMate` struggle to identify check-worthy claims in some cases, but `CheckMate` performs exceptionally well in the majority. In example x_1, which states *Methanol, Ethanol, and Bleach are poisons and drinking them can lead to*

disability and death, we see that both systems classify the check-worthiness label correctly but partially misclassify the rationality labels. `CheckMate` adeptly identifies the information as verifiable; however, erroneously assigns the label necessitating fact-checkers attention to the post, potentially influenced by the presence of impactful terms such as *'death'* and *'disability.'* It is noteworthy that XLNet also encounters challenges in correctly discerning all the rationales in this instance. While correctly identifying the content of interest to the general public, it overlooks categorizing it as verifiable information, which may mislead fact-checkers into prioritizing other claims for verification. This oversight can potentially lead to inefficiencies in fact-checking efforts and delay the identification and correction of misinformation.

Example x_2, asserts that *virus will die from cow-dung cakes and agarbatti*, which is definitely a check-worthy claim as it contains false information and can harm society. Both models proficiently identify several rationales but falter in pinpointing the need for fact-checkers' attention; a divergence in their approaches becomes apparent. XLNet introduces an additional rationale – *needs government's attention*. This raises intriguing questions about the model's interpretative lens, prompting consideration of whether this expansion aligns with a broader understanding of the content or introduces a spurious element. Conversely, `CheckMate` adopts a more cautious stance, offering a partially correct classification and refraining from making further inferences. This prudent approach opens avenues for discussing the trade-off between assertiveness and precision in model predictions. It prompts consideration of whether a more conservative model, avoiding unwarranted inferences, may yield a more reliable and nuanced decision-making process. This nuanced understanding lays the groundwork for refining the models and advancing the discourse on the complexities of rationality label prediction in fine-grained setups. We witness similar behaviour in the next example, where our model correctly identifies all the rationality labels while XLNet generates extra rationality labels.

	Claim	**Ground Truth**	**Predictions**	
			`CheckMate`	**XLNet**
x_1	Methanol, Ethanol, and Bleach are poisons and drinking them can lead to disability and death.	Check-worthy *(verifiable information, public interest)*	Check-worthy *(verifiable information, needs fact-checker's attention)*	Check-worthy *(public interest)*
x_2	Put cow products like cow-dung cakes and agarbatti made from that. Upon using these, the virus will die immediately.	Check-worthy *(verifiable information, false information, public interest, harmful, needs fact-checker's attention)*	Check-worthy *(verifiable information, false information, public interest, harmful)*	Check-worthy *(verifiable information, false information, public interest, needs governments' attention)*
x_3	The virus blocks red blood cells from absorbing and distributing oxygen.	Check-worthy *(verifiable information, public interest, needs fact-checkers' attention)*	Check-worthy *(verifiable information, public interest, needs fact-checker's attention)*	Check-worthy *(false information, needs fact-checker's attention)*
x_4	No single VIRUS can ever wipe out the entire human race from this earth.	Non check-worthy	Non check-worthy	Check-worthy *(needs fact-checker's attention)*
x_5	If we all stop breathing, the spread would stop.	Non check-worthy	Check-worthy *(public interest)*	Non check-worthy

Table 6.8: Error analysis of the check-worthy labels of `CheckMate`. Rationality labels for check-worthy claims are given in brackets and italics. For comparison, we also present the predictions of the best-performing baseline system, XLNet. Errors are highlighted in red.

The last two examples in Table 6.8 are non check-worthy claims. In the x_4, the claim asserts that *No single VIRUS can ever wipe out the entire human race from this earth.* `CheckMate` correctly labels this claim as *non check-worthy*. On the other hand, XLNet marks the claim as *check-worthy*, which contradicts the ground truth. The claim contains the phrase *'wipe out the entire human race'* is open to interpretation

regarding the severity or scope of the event. This possibly demonstrates the inability of the model to understand the broader context or implications of the claim. It might not recognize the rhetorical nature of the statement or consider the relevant scientific or historical context. We witness similar behaviour from our model in example x_5, which states *if we all stop breathing, the spread would stop.* It follows that our system has difficulty dealing with real-world knowledge and sarcastic statements. The phrase *'spread would stop'* causes the system to pay attention to it and incorrectly predicts it as a check-worthy claim. Given that it discusses one of the life-threatening COVID-19 cures, `CheckMate` indicates that the claim is check-worthy, as it might interest the general public. However, the rambling writing style can confuse both systems and yield inaccurate predictions.

Delving deep into these error cases reveals that the fine-grained setup for rationality labels introduces a nuanced layer of complication compared to the relatively straightforward coarse-grained counterpart for check-worthiness. Notably, our model has a commendable level of accuracy, making correct predictions in most cases. In most instances, our model demonstrates accuracy by making at least one correct prediction for rationality labels.

6.7 Summary

The intricacies surrounding claim check-worthiness underscore several challenges for both human evaluators and automated systems. The inherent subjectivity of such claims is exacerbated by the frequent lack of linguistic soundness in these texts, presenting a major challenge. This challenge is further amplified in the context of web-based short texts. By addressing this fundamental problem, our work sets the stage for a nuanced analysis that seeks to improve model performance and sheds light on the underlying rationales of determining claim check-worthiness. In this work, we introduced a pioneering task of explainable claim check-worthiness, leveraging rationality labels. The lack of a suitable annotated dataset has posed a major challenge in explainable claim check-worthiness detection within online social media platforms. To address this limitation, we curated an extensive Twitter corpus comprising over $5k$ manually annotated claims for precise identification of rationality label and check-worthiness label. We benchmarked our dataset with a unified architecture, `CheckMate`, that outperformed existing state-of-the-art check-worthiness systems and illustrated how the abetting rationale factors through rationality labels could be used to comprehend more reliable claim check-worthy systems. The results showed that our suggested model outperformed the best-performing baselines by $\geq 0.31\%$ in macro F1-score. We illustrated how leveraging different determining factors can lead to more informed decisions about check-worthy labels. Moreover, the incorporation of rationales can assist manual fact-checkers in prioritizing fact-checking efforts more efficiently. While our current focus has been on English, future endeavours will extend to developing fine-grained claim check-worthiness detection models for other languages, particularly focusing on low-resource languages. In our subsequent work, we will also emphasise expanding the dataset and broadening the scope by incorporating multimodal elements into the detection framework.

Part 4

Verifying Check-Worthy Claims

7. DISCOURSE-BASED VERIFICATION

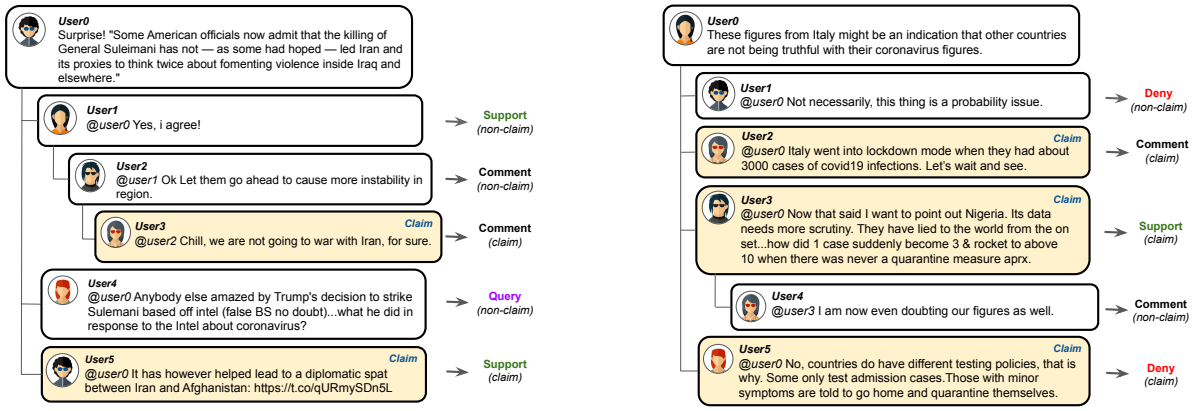
Misinformation spreads rapidly on social media, causing serious damage by influencing public opinion, promoting dangerous behavior, or eroding trust in reliable sources. It spreads too fast for traditional fact-checking, stressing the need for predictive methods. We introduce `CrowdShield`, a crowd intelligence-based method for early misinformation prediction. We hypothesize that the crowd's reactions to misinformation reveal its accuracy. Furthermore, we hinge upon exaggerated assertions/claims and replies with particular positions/stances on the source post within a conversation thread. We employ Q-learning to capture the two dimensions – stances and claims. We utilize deep Q-learning due to its proficiency in navigating complex decision spaces and effectively learning network properties. Additionally, we use a transformer-based encoder to develop a comprehensive understanding of both content and context. This multifaceted approach helps ensure the model pays attention to user interaction and stays anchored in the communication's content. We propose `MisT`, a manually annotated misinformation detection Twitter corpus comprising nearly 200 conversation threads with more than $14K$ replies. In experiments, `CrowdShield` outperformed ten baseline systems, achieving an improvement of $\sim 4\%$ macro-F1 score. We conduct an ablation study and error analysis to validate our proposed model's performance.

7.1 Introduction

The proliferation of misinformation over social media platforms has become a critical challenge. The sheer amount of misinformation across platforms significantly impacts public opinion and can have tangible real-world consequences. Notable events such as the 45th Presidential US Elections, the COVID-19 pandemic, and the Ukraine-Russia war highlight the adverse impact of online misinformation on society (257; 258; 259). Manual fact-checkers struggle to identify misinformation due to the enormous volume of content generated and shared on social media daily. As a result, researchers have begun investigating automated methods addressing this problem (260; 261; 262; 251). Moreover, the rapid spread of misinformation on social media poses another formidable challenge in containing the misinformation early before it reaches a wide audience. By identifying misinformation before it gains substantial traction, we can mitigate its negative effects and uphold the integrity of online discourse. The concept of early detection has been well-studied for online information disorders such as hate speech (263), rumors (264), fake news (99), etc. At the same time, developing robust methods to detect misinformation early is also essential to protect the accuracy of information and promote meaningful discussions. Recently, major advancements have been in the early detection of misinformation (99; 265). These efforts aim to strengthen digital platforms against the harmful effects of misinformation and ensure that accurate and reliable information prevails online. This work aims to devise an approach for early misinformation detection in social media utilizing *crowd intelligence*.

Conventional approaches to discern misinformation typically rely on linguistic patterns or external knowledge sources to ascertain whether the information is true. Content-based detection hinges heavily on linguistic cues, emotions, or sentiments (86; 266). However, these methods come with certain shortcomings. Firstly, the messages on social media sites like Twitter and Facebook are short and informal. As a result, linguistic features extracted from them tend to be insufficient for deep learning algorithms.

Moreover, they often use excessive informal text and slang, making following proper grammar and syntax challenging. Another line of research concentrates on evidence-based misinformation identification (267; 268; 269). These systems are generally reliable and depend on established evidence to validate the posts. However, these approaches can be time-consuming and unable to keep up with the swift propagation of misinformation. At the same time, they often lack sufficient evidence to verify newly posted information. Recent studies have sought innovative methods to detect misinformation that combine temporal features extracted from user responses and propagation networks (270; 98; 91). We also hypothesize that the *'crowd intelligence'* through these user responses hold a wealth of direct or indirect cues, which can be leveraged to effectively assess the credibility of the source posts.



(a) Non-misinformation

(b) Misinformation

Figure 7.1: Illustrative examples of Twitter posts and their subsequent conversational thread, with the original post containing false and true information, respectively. The hierarchical arrangement of replies reflects users' interaction and stances toward the source post. Replies highlighted in yellow denote claims made by the users in response to the source post.

Crowd Intelligence. Social media users often express their opinions on posts by supporting or contradicting them, which aids in determining the credibility of the information posted (271; 272). Building on this, we introduce the concept of *crowd intelligence*. This refers to the collective decision-making capabilities of social media users that emerge from the interaction and aggregation of diverse inputs. Researchers highlight that stances such as support and denial play a crucial role in predicting veracity, whereas comments – statements lacking any explicit stance – are deemed irrelevant for this task (273). However, we hypothesize that not all comments are irrelevant, and they may contain crucial claims essential for assessing the truthfulness of the information posted. As defined by Toulmin (53), *a claim is an assertive statement with or without evidence*. These assertive statements in user responses can be strong indicators for determining the veracity of the source post, giving additional context for the post. For instance, in Figure 7.1a, where User3's comment lacks an explicit stance but makes an important claim that cannot be overlooked. Thus, by harnessing crowd intelligence through user stances and claims, we can significantly enhance the effectiveness of misinformation prediction. Figure 7.1 vividly demonstrates the utility of analyzing the source posts and how others orient to them. It depicts two posts from Twitter and some initial user replies. The first example is a tweet about the aftermath of General Suleimani's death and its ineffectiveness in reducing violence from Iran and its proxies. Importantly, this tweet is not misinformation. The reactions are mostly favourable or add additional commentary, indicating a shared understanding of the situation described in the tweet. This shared viewpoint implies a fundamental level of public acceptance, which is critical in determining the veracity of information shared on social media at an early stage. Furthermore, claims in this conversation thread elevate the conversation above mere agreement. Responses like *'we are not going to war with Iran, for sure'* are significant as they go beyond

straightforward agreement to make independent claims reinforcing the original post. These claims are critical; they carry an inherent legitimacy as an authority contrasted with passive responses such as *'Yes, I agree.'* These claims serve as anchors in the discourse, providing substantive support for the tweet's message while highlighting a higher level of community engagement. In contrast, in the second case, the tweet is misinformation. The discourse surrounding this tweet is characterized by a broad spectrum of stances, most notably the presence of direct denials and contradictory claims that serve as significant red flags here. These claims directly contradict the tweet's assertions, showing denial of the source post. This exemplifies that crowd intelligence is critical in determining social media content's credibility.

To address the early misinformation prediction problem, we propose `CrowdShield`, a deep-learning framework that utilizes the post's semantic and user responses to predict false information. At a high level, `CrowdShield` first adopts a Q-network to learn propagation representation through the user's stances towards the source post. Next, the model utilizes a transformer-based model to learn the semantic features from the source post and the relevant replies. Finally, the learned semantic and propagation representations are combined to determine whether the post is misinformation or not. We also create a real-world corpus, `MisT`, comprising 199 Twitter posts and 14,436 user replies. Each source post is manually annotated for misinformation, and user replies are labelled with claim and stance labels. Through this work, we make the following contributions:

- We present a novel model for early misinformation detection that takes advantage of crowd intelligence reflected through user reply stances and claims.
- We present a novel dataset that includes Twitter threads meticulously annotated for the veracity of the source post. Additionally, we mark stances expressed in replies to the source post and claim labels associated with each reply.
- We perform exhaustive experiments and confirm significant improvements over several baselines. We also provide detailed qualitative and quantitative analysis.

7.2 Related Work

The widespread dissemination of false or misleading information on social media platforms has attracted much scrutiny in recent years. A plethora of research has been conducted focusing on various aspects of identifying and mitigating misinformation, including fake news, rumors, etc (82; 83; 84; 85). Researchers have investigated several approaches for identifying misinformation, including methods that analyze the content, examine the network, and predict falsehoods in advance.

Content-based approaches rely heavily on linguistic clues such as writing styles, lexical aspects, sentiment analysis, and subject relevance (83; 84; 85). For example, Castillo et al. (86) discovered that highly reputable social media messages have more URLs and longer text lengths than less credible ones. Similarly, Pawan et al. (87) introduced the WELFake model, which detects bogus news by combining word embedding vector and linguistic data. Anshika et al. (85) proposed a deep learning model that leverages syntactic, grammatical, semantic, and readability aspects of news stories to identify bogus news. Validated over several publicly accessible datasets, Jain et al. (84) created the Confake algorithm using a comprehensive set of content-based features and word vector attributes taken from news items. These content-based methods analyze the credibility of individual posts in isolation, often neglecting the high correlation between a post's veracity and the replies it garners from other users.

To overcome the limitations of content-based methods, current research has concentrated on propagation-based approaches. These methods employ social context knowledge to identify false information by examining the dissemination of information on social networks, the people responsible for disseminating

it, and the relationships between these propagators (88; 89; 90; 91). Ma et al. (88) developed a neural network with a tree structure that uses false news cascades to detect disinformation. Dhawan et al. (89) introduced GAME-ON, a system that utilizes Graph Neural Networks to enable detailed interactions inside and between several modalities for the purpose of detecting multimodal fake news. Song et al. (90) created a temporal learning model based on graphs to capture the changing patterns of tweets that are organized as tree-structured data. Kang et al. (92) constructed a news detection graph that links bogus news with many third-party information sources. Sun et al. (93) proposed a hypergraph learning model that utilizes a unique hyperedge walking technique and hyperedge expansion method to create comprehensive representations for entire graphs. These studies indicate that models utilizing network information outperform those that merely rely on content. In addition to propagation features, user-specific attributes, including the number of followers and the stance of other users toward the post, have also been employed to detect misinformation. The significance of other user's reactions in the verification of information has been emphasized in numerous studies (86; 94; 95). Li et al. (96) conducted a study on the semantic aspects of false information, analyzing their dissemination and user characteristics, using 421 false statements and 1.47 million related tweets. One of the most significant discoveries is that individuals are inclined to disseminate falsehoods when they are uncertain about their veracity without incorporating their personal opinions. According to (96), users are inclined to endorse unverified information. In contrast, highly reputable users, such as news organizations, endeavour to publish well-reasoned statements that are supported by evidence and appear to be certain. These findings underscore that models using network information and user responses are highly efficient for misinformation detection on social media.

With the quick dissemination of information on social media, false information can reach thousands of users in minutes, causing significant confusion, panic, and even physical harm. It is essential for real-time systems to identify misinformation at its inception before its rampant dissemination. Early detection algorithms have been the subject of numerous investigations (97; 98; 99). Kwon et al. (97) analyzed feature stability over time. They discovered that user and linguistic features are more effective than structured and propagation features in determining the veracity of information during the early stages. Using network embedding techniques on social network graphs, Liu and Wu (99) constructed user representations using network embedding approaches on the social network graph. The critical importance of this area in real systems necessitates additional research despite the existence of current research.

An effective solution to address the issue of misinformation involves a detailed exploration of user responses to posted information, considering their stances and the strength of their claims. The positions held by users on the information can offer valuable insights into the accuracy of a certain post. The combined intelligence of users can help identify inconsistencies or provide support for factual content. Traditional network or user-based approaches often neglect these factors, instead primarily concentrating on disparities in structure. Addressing these limitations, we propose a framework that integrates linguistic and crowd intelligence to enhance the integrity and reliability of online information by swiftly identifying misinformation. Our framework is designed to meet the critical need for early detection, providing a robust solution to mitigate the rapid spread of misinformation at the earliest possible.

7.3 Dataset

Over the past few years, several misinformation detection datasets have been released (271; 274). However, none of these datasets come with both stance and claim labels for the user replies. The only dataset that closely aligns with our research objectives is the SemEval-2019 Task 7 dataset (274). It contains 446 conversation threads and 8142 user replies, containing rumors gathered from Twitter and Reddit during eight distinct events. It contains stance labels (i.e., *support*, *deny*, *comments*, and *query*) for replies and

veracity labels (i.e., *true*, *false*, and *unverified*) for source posts. However, similar to other cases, it also does not contain claim labels. For our task, we annotate claim labels at the user reply level. Given the age of these events, we further meticulously cross-checked all source posts for veracity labels to ensure the accuracy and reliability of the data. Furthermore, we revisit all instances labeled as *unverified* to reassess their truthfulness with new evidence. Out of 110 unverified tweets, 47 have been classified as misinformation, while 63 as non-misinformation. Similarly, 8 out of the 13 unverified Reddit posts are categorized as misinformation and 5 as true.

Despite its usefulness, the SemEval-2019 dataset has some limitations – it primarily focuses on well-known and limited topics. Also, the dataset contains posts related to events that are now outdated and of limited relevance. Therefore, we additionally curate a new Twitter corpus, `MisT` (**Mis**information detection on **T**witter) with a much diverse set of recent topics, hence making it more representative of general social media interactions.

7.3.1 Collection

We collect random trending Twitter threads from various periods, user demographics, and thematic categories. This ensures that the dataset is representative of a cross-section of Twitter discourse and encompasses various interests and perspectives. We leverage Twitter's API to access publicly available conversation threads while adhering to ethical guidelines and platform policies regarding data usage and user privacy.[1] We collect the data between October 4, 2022 to December 31, 2022. Each instance selected for inclusion consists of a source post, which serves as the initiating message or topic of discussion, followed by a series of user replies, forming a cohesive conversational thread.

7.3.2 Annotation

For annotation, we employ four annotators experienced in social media and linguistics. We annotate the collected samples focusing on two dimensions – *source post* and *user replies*.

- Source Post: We meticulously examine and annotate every source post to indicate whether it contains misinformation or adheres to factual accuracy. For each tweet, the annotators can assess the content's veracity based on their prior knowledge or available external sources.
- User Replies: We analyze each reply of the conversation thread and annotate them for their stance towards the source post. Stances can take one of the four values: support, deny, comment, and query. Note that we use *'root'* to denote the stance of the source post. Besides stance annotation, user replies are subject to claim annotation to identify whether they assert something or not. We employ the annotation guidelines provided by Gupta et al. (251) to annotate the replies for claims.

The annotation process involves three rounds of review and refinement to ensure consistency and agreement among annotators. Any discrepancies were resolved through consensus discussions and adjudication by senior annotators or domain experts. This iterative annotation process aimed to achieve high-quality annotations with a high degree of accuracy, reliability, and relevance to the task of misinformation detection and analysis on social media platforms. Finally, we obtain Cohen's Kappa (275) inter-annotator agreement (IAA) score as 0.67 for claim labels and 0.79 for stance labels.

[1] https://x.com/en/privacy

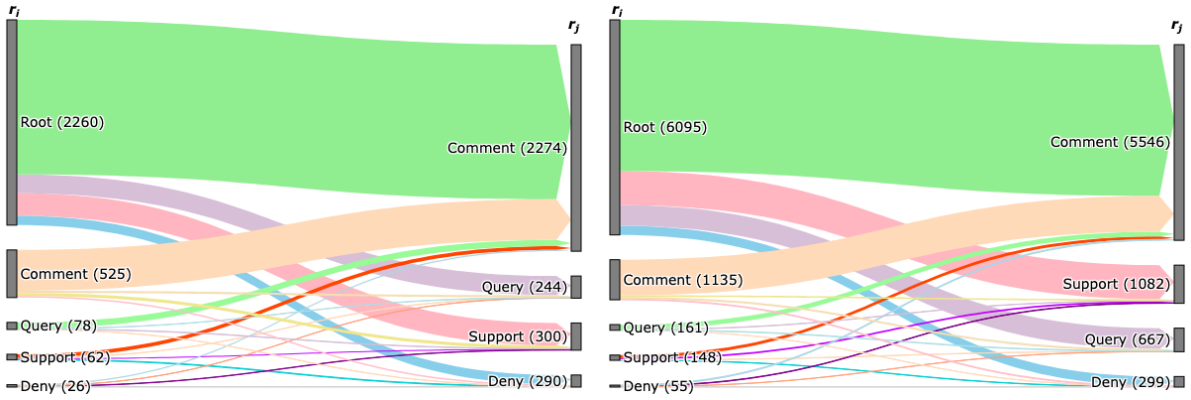
7.3.3 Statistics and Analysis

MisT comprises 199 tweets with 14,436 user replies. Table 7.1 contains detailed statistics for both MisT and SemEval datasets. It is important to note that the SemEval dataset has a higher prevalence of misinformation than ours, while the number of misinformation instances in MisT is significantly low. This is due to the fact that our dataset is *neutrally-seed* and does not focus on specific and controversial events, such as those highlighted in the SemEval dataset. The qualitative analysis of the dataset is delineated as follows.

	MisT		**SemEval**$_{Twitter}$		**SemEval**$_{Reddit}$	
	Mis	**Non-mis**	**Mis**	**Non-mis**	**Mis**	**Non-mis**
Train	40	119	190	135	12	28
Test	10	30	24	32	14	11
Total	50	149	214	167	26	39

Table 7.1: Dataset statistics showing misinformation (mis) and non-misinformation (non-mis) labels.

Stance Evolution within Replies. We analyze how stances change between consecutive responses within the conversation threads. Figure 7.2 demonstrates the evolution of stance from reply r_i (left vertical) to subsequent reply r_j (right vertical) in conversation threads. The first vertical represents the stance of a reply, with the caveat that stances for source posts are labelled as 'root' due to their inherent nature. Conversely, the second vertical denotes the stance of the immediate following reply. Upon analysis, we observe a notable trend across misinformation and true information – most direct replies to source posts are categorized as 'comments,' lacking explicit stances, thus making it crucial to analyze further discourse to understand the veracity of the social media threads. However, a distinction emerges regarding the level of support for the source post between true and false misinformation. Specifically, in the case of misinformation, the support for the source post is notably lower than true information, as evidenced by the flow from 'root' to 'support' in Figure 7.2 (b).



(a) Misinformation (b) Non-misinformation

Figure 7.2: Analysis of stance evolution in replies within the conversation threads. The diagram depicts the change of stance from reply r_i (shown on the left vertical axis) to subsequent reply r_j (shown on the right vertical axis) in conversation threads.

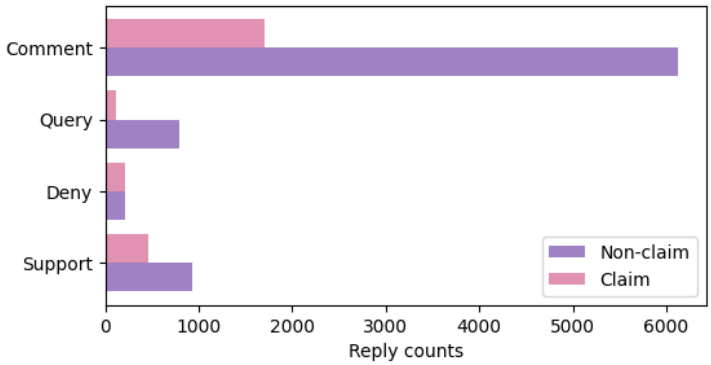


Figure 7.3: Correlation between different stances towards source posts and their assertive nature, as indicated by claim labels, across all conversation threads. The purple bar indicates whether the response is a non-claim, whereas the pink bar indicates whether it is a claim.

Relationship between Stance and Claims. In this analysis, we delve deep into the relationship between reply stances and the presence of claims within them. We discover intriguing patterns that shed light on user behaviour and communication dynamics. Figure 7.3 depicts the key findings from this analysis. Initially, we observe a significant predominance of the 'comment' stance in responses without explicit claims. This implies that users frequently discuss, share opinions, and provide context without making factual claims. However, in replies containing claims, the prevalence of the 'comment' stance reduces to more than half compared to non-claims. However, the claim count remains higher when compared to other stances, indicating that not all explicit stances are useful and rejecting all comments can be harmful. The 'deny' stance is notably uncommon across all responses, whether they contain claims or not. This suggests that users prefer presenting alternative viewpoints or evidence rather than outright rejecting statements. However, it is worth noting claims and non-claims appear equally within the 'deny' stance, emphasizing its importance as a stance to consider when analyzing user responses.

7.4 Methodology

There are several definitions of misinformation across existing papers. In this work, we define misinformation as *'a piece of information propagating through social media which is ultimately verified as false or inaccurate.'* Formally, we define the task as follows. Given a social media post p and its associated set of replies $R = \{r_1, r_2, \ldots, r_n\}$. Each reply r_i is labeled with stance label s_i (i.e., support, deny, query, or comment) towards p and claim c_i (i.e., claim or non-claim). We aim to predict whether p disseminates misinformation as early as possible. This involves analyzing the initial few replies, denoted as τ, and leveraging the stance s_i and claim c_i labels associated with these replies.

In this chapter, we propose a novel framework, `CrowdShield`, to comprehend the social media conversation thread propagation dynamics using deep Q-learning, an off-policy reinforcement learning technique. Our primary objective is to understand the stances observed within the replies of a thread. We opt for deep Q-learning because it can effectively navigate complex decision spaces and learn network properties well, thereby shedding light on the underlying discourse patterns. Our model can iteratively explore the state-action space by leveraging deep Q-learning, gradually refining its understanding of the underlying thread propagation dynamics. Furthermore, the off-policy nature of deep Q-learning enables the decoupling of the exploration and exploitation phases, facilitating more robust and efficient learning. Figure 7.4 illustrates the overall architecture of `CrowdShield`.

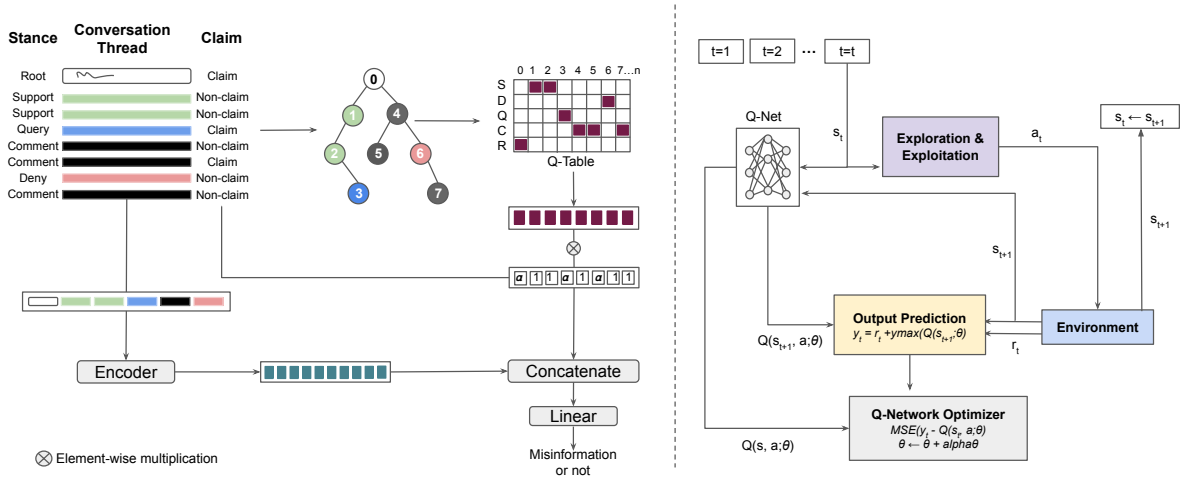


Figure 7.4: Illustrative model diagram for our proposed framework for early misinformation prediction. The right side of the diagram shows the Q-table update mechanism. In the Q-table, S, D, Q, and C denote support, deny, query, and comment, respectively.

To begin with, we define the foundational elements of our methodology:

State Space: The *state* is defined as the content and context of the current reply. Thus, the state space is determined by the set of all possible replies and their content. Mathematically, a state s can be represented as,

$$s = (i, c, p) \tag{7.1}$$

where i denotes the reply id, c represents the content of the current reply, p denotes time of posting. s_0 represents the source post in the conversation thread. Consequently, the state space (S) is represented as follows,

$$S = \{(i_0, c_0, p_0), (i_1, c_1, p_1), (i_2, c_2, p_2), ...(i_n, c_n, p_n)\} \tag{7.2}$$

where each tuple (i_t, c_t, p_t) represents a specific state in the state space and n denotes the total number of replies (states) in the conversation thread.

Action Space: Complementing the state space, the action space plays a pivotal role in orchestrating the traversal of the conversation thread. Our architecture conceptualizes the stances expressed within the replies as *action*. Formally, at time step, t, the action a_t signifies the stance of the corresponding reply within the thread towards the source post. Thus, the action space encompasses all possible stances a reply can exhibit. Notably, the source post, the initial point of the conversation thread, is characterized by the *Root* stance. Consequently, the action space (A) is delineated as follows,

$$A = \{support, deny, query, comment, root\} \tag{7.3}$$

with a_0 representing the state of the source post, always being assigned the *root* stance.

7.4.1 Q-Learning Phase

In the Q-learning phase, the primary objective is to accurately predict the Q-values for each state-action pair within the model's architecture. In essence, these Q-values represent the expected rewards of having a particular state at the given action. We employ a deep Q-learning network, an effective hybrid neural network framework, which would take in the current state of the environment as input and output the

estimated Q-values for each possible action. The learning process begins with initializing the Q-table entries to zero. This ensures that the learning starts from a neutral and unbiased perspective. As the model interacts with the environment, it transitions from one state to another based on the actions taken, and it receives rewards that reflect the efficacy of each action. Learning directly from consecutive samples is inefficient due to the strong correlations between the samples and the chances of overfitting; thus, randomizing the samples limits these correlations and reduces the variance of the updates. To solve this, we leverage exploration and exploitation within the framework through ϵ-greedy policy. This policy allows the agent to explore new actions with a certain probability, determined by ϵ. The value of ϵ ranges between 0 and 1. Exploration involves randomly selecting actions, while exploitation entails choosing the action with the maximum Q-value obtained from the Q-network table. The Q-network table is updated using the Bellman equation:

$$Q(s_t, a_t) = R_t + \epsilon \max_a Q(s_{t+1}, a) \tag{7.4}$$

where $Q(s_t, a_t)$ is the Q-value for the state-action pair at time step t, r_t is the immediate reward obtained at that state, ϵ is the discount factor that balances the importance of immediate and future rewards, and s_{t+1} is the next state after taking action. The agent learns to make decisions that optimize long-term rewards by maximizing the Q-value for each state-action pair. We define reward function R as the sum of the claim reward R_{claim} and the stance reward R_{stance}. The claim reward R_{claim} is 1 if the content of the reply is a claim, denoted by c, and 0 otherwise. The stance reward R_{stance} is defined based on the stance taken in the reply. Specifically, we formally define R_{claim} and R_{stance} as,

$$R_{\text{claim}} = \begin{cases} 1 & \text{if } c \text{ is a claim} \\ 0 & \text{otherwise} \end{cases} \tag{7.5}$$

$$R_{\text{stance}} = \begin{cases} 1 & \text{if } stance = \text{ 'support' or 'query' or 'root'} \\ -1 & \text{if } stance = \text{ 'deny'} \\ 0 & \text{if } stance = \text{ 'comment'} \end{cases} \tag{7.6}$$

Therefore, the total reward R ranges between -1 to 2 and is given by,

$$R = R_{\text{claim}} + R_{\text{stance}} \tag{7.7}$$

To optimize the parameters of the Q-network, we use gradient descent with the Mean Squared Error loss. The target value y_t is calculated as follows,

$$y_t = \mathbb{E}_{s'} \left\{ R + \epsilon \max_a Q(s_{t+1}, a, w_{t-1}) \right\} \tag{7.8}$$

The loss function $L_t(w_t)$ is then defined as,

$$L_t(w_t) = \left[y_t - Q(s, a, w_t) \right]^2 \tag{7.9}$$

where $Q(s, a, w_t)$ is the predicted Q-value for state s and action a with current parameters w_t.

By minimizing this loss using gradient descent, the parameters of the Q-network are updated to approximate the true Q-values better. We employ the Adam optimizer during backpropagation to efficiently update the parameters and improve the overall performance of the Q-network.

7.4.2 Training Phase

After the Q-learning phase, we obtain a Q-table. The values in the Q-table represent the expected future rewards for taking a particular action in a given state. To construct a feature vector for each conversation thread, we extract the values from the Q-table. This involves capturing the Q-values corresponding to the given stance exhibited within the thread for each reply. We define the feature vector for thread i as follows,

$$F_i = [Q(s_0, a_0), Q(s_1, a_1), \cdots, Q(s_n, a_n)] \tag{7.10}$$

where $Q(s_j, a_j)$ denotes the Q-value for jth reply in the thread, (s_j, a_j) denotes its state-action pair for the reply, and n is the total number of replies in the thread.

For each reply in the thread, we create a binary claim vector C_i, of the same size as the thread. If the reply is a claim, the corresponding element in $c_{ij} \in C_i$ is marked as 1 else 0. So, the claim vector for the ith thread would be:

$$C_i = [c_{i0}, c_{i1}, c_{i2}, \cdots, c_{in}] \tag{7.11}$$

where

$$c_{ij} = \begin{cases} 1 & \text{if } jth \text{ reply in } ith \text{ thread is a claim} \\ 0 & \text{otherwise} \end{cases}$$

With the objective of assigning more weightage to claims in the veracity determination process, we first multiply each value in the claim vector C_i by the scalar claim weight factor α. Then, we compute the Hadamard product of the resulting vector with the feature vector F_i,

$$F_i = (\alpha \bullet C_i) \otimes F_i \tag{7.12}$$

where α = 1 denotes equal emphasis on claims and non-claims in the reply thread. The value of α can be empirically set to an appropriate real number.

Next, we combine the source post and its corresponding replies having *support*, *deny*, and *query* stances and obtain rich semantic representation S for the combined text. We utilize a pretrained language model – BERT (211) – and fine-tune it on our data to generate text representation.

$$S_i = BERT(combined\_text_i) \tag{7.13}$$

Subsequently, we concatenate Q-feature vector F_i with the semantic feature vector S_i to form the final feature vector V_i for each thread,

$$V_i = [F_i, S_i] \tag{7.14}$$

Finally, we employ a linear layer to facilitate final classification that can be computed as,

$$\hat{y} = \sigma(\mathbf{W}V + b) \tag{7.15}$$

where $\hat{y}$ is the final predicted misinformation label. $\mathbf{W}$ represents the weight matrix and b represents the bias vector of the linear layer. .

7.5 Experiments and Results

Experimental Setup. In the Q-training phase, we use a single-layer network for the deep Q-learning network and train it over 1000 episodes. To balance exploration and exploitation, we empirically set the discount factor (ϵ) at 0.2, resulting in less exploration during training. On our dataset, we fine-tune a BERT model obtained from HuggingFace.[2] The fine-tuned BERT model is then used to generate text embeddings for the final feature vectors. We train this model for 20 epochs, stopping early based on macro-F1 and setting patience to 3. The final classifier employs a feed-forward layer with softmax activation. We use the Adam Optimizer with a learning rate of 0.001 and a batch size of 8 for training. Experimentally, we find the best α value to be 2 (shown in Table 7.3). We keep a development set containing 10% of the training data to optimize and fine-tune the models, allowing us to monitor and adjust performance during training iterations. For all baseline systems, we replicate the models using the descriptions provided in their original papers.

Evaluation Metrics. For evaluation, we primarily use macro-F1 scores from *scikit-learn* library. In addition, we report class-wise recall, precision, and F1 scores to provide a more detailed understanding of model performance.

Baseline Models. We employ the following baseline systems. ▷ **LSTM** (276): A long short-term memory network trained for binary classification tasks. ▷ **RNN** (277): A recurrent neural network is any network whose neurons send feedback signals to each other. ▷ **BERT** (211): A bidirectional transformer-inspired auto-encoder language model fine-tune for our misinformation detection task. ▷ **RoBERTa** (114): A robustly optimized BERT approach with improved training methodology. We fine-tune it on our data. ▷ **EventAI** (278): The system achieved the highest ranking in the SemEval 2019 Task 7 on rumor verification. To verify rumors, they utilized information from several aspects, including the content of the rumor, the reliability of the source, the credibility of the user, and the stance of the user. We exclude the user credibility features due to the absence of relevant user information. ▷ **GCN** (279): A graph convolutional network is a neural network adapted to leverage the structure and properties of graphs. We also augmented it with word embeddings. ▷ **RPL** (280): A zero-shot framework based on a prompt learning system to detect rumors falling in different domains. It uses a hierarchical prompt encoding mechanism for learning language-agnostic contextual representations for both prompts and data. ▷ **ACLR** (281): The architecture proposes an adversarial contrastive learning framework to detect rumors. ▷ **GMVCN** (98): The system encodes the multiple views of the conversation thread based on GCN and leverages convolutional neural networks to capture consistent and complementary information. ▷ **GPT-4** (282): A state-of-the-art large language model intended to address intricate queries and provide responses that are both coherent and contextually pertinent across a diverse array of tasks.

Experimental Results. Table 7.2 presents consolidated results on our dataset `MisT`. We ran two sets of experiments to evaluate text classification models. Initially, models were trained solely on the source post (SP), with no additional information from conversation threads (CT). Second, we incorporated baseline systems that utilized conversation threads into the misinformation identification process, employing a variety of methodologies to capture response dynamics. The first two rows of Table 7.2 show the performance of binary classification systems for misinformation detection with only the post. Notably, BERT outperformed other systems on the Twitter dataset, with the highest macro-F1 score of 0.60. To improve the source post, we added replies tagged as 'support,' 'deny,' or 'query,' as well as claims, to the

[2] https://huggingface.co/docs/transformers/en/model_doc/bert

Model	Input	Non-misinformation			Misinformation			Macro-F1
		Pecision	Recall	F1	Pecision	Recall	F1	
LSTM	*SP*	0.7838	0.9667	0.8657	0.6667	0.2000	0.3077	0.5867
RNN	*SP*	0.7586	0.7333	0.7458	0.2727	0.3000	0.2857	0.5157
BERT	*SP*	0.8000	0.8000	0.8000	0.4000	0.4000	0.4000	0.6000
RoBERTa	*SP*	0.7812	0.8333	0.8065	0.3750	0.3000	0.3333	0.5699
BERT$_{stance}$	*SP* ∪ *CT*	0.7188	0.7667	0.7419	0.1250	0.1000	0.1111	0.4265
BERT$_{claim}$	*SP* ∪ *CT*	0.7692	1.0000	0.8696	1.0000	0.1000	0.1818	0.5257
BERT$_{stance_claim}$	*SP* ∪ *CT*	0.7429	0.8667	0.8000	0.2000	0.1000	0.1333	0.4667
EventAI	*SP* ∪ *CT*	0.7500	1.0000	0.8571	0.0000	0.0000	0.0000	0.4286
GCN	*SP* ∪ *CT*	0.8000	0.6667	0.7273	0.3333	0.5000	0.4000	0.5636
ACLR	*SP* ∪ *CT*	0.7742	0.9600	0.8571	0.6667	0.2222	0.3333	0.5952
RPL	*SP* ∪ *CT*	0.7500	0.7500	0.7500	0.3000	0.3000	0.3000	0.5250
GMVCN	*SP* ∪ *CT*	0.6957	0.5333	0.6038	0.1765	0.3000	0.2222	0.4130
GPT-4	SP	0.7619	0.5333	0.6275	0.2632	0.5000	0.3448	0.4861
GPT-4	*SP* ∪ *CT*	0.7667	0.7667	0.7667	0.3000	0.3000	0.3000	0.5333
CrowdShield	*SP* ∪ *CT*	0.8125	0.8667	0.8387	0.5000	0.4000	0.4444	0.6416
– {*Q-learning*}	*SP* ∪ *CT*	0.7879	0.8667	0.8254	0.4286	0.3000	0.3529	0.5892
– {*Text Features*}	*SP* ∪ *CT*	0.7576	0.8333	0.7937	0.2857	0.2000	0.2353	0.5145

Table 7.2: Experimental results of CrowdShield and its variants (last two rows) on our dataset MisT. Input to the system: *SP* denotes Source Post only, and *SP* ∪ *CT* denotes Source Post and the corresponding Conversation Thread.

BERT model. Among these configurations, the one that included only replies classified as claims alongside the source post produced the best results, with a macro-F1 score of 0.52 macro-F1 (rows 5-7). This emphasizes the significance of claims in conversation threads as indicators of misinformation. Interestingly, including these replies did not significantly improve the overall BERT performance compared to using the source post alone, highlighting the challenge posed by the potential noise in the replies. Moreover, compared to the existing baseline systems that include the conversation threads in their frameworks, CrowdShield achieved the highest macro-F1 score, exhibiting approximately 5% improvement over the best baseline – ACLR (281). Furthermore, CrowdShield obtained an F1 score of 0.44 for the true class, indicating superior performance in detecting misinformation compared to the other systems. GPT-4 initially performed poorly for the task, scoring 0.48. However, when replies were included, its performance improved significantly, with a nearly 4.7% increase. This improvement could be attributed to the context provided by the conversation threads. By taking into account interactions within the thread, GPT-4 was likely able to understand the conversation's context and nuances better, resulting in improved performance in identifying and addressing misinformation.

Early Detection Efficiency. Detecting misinformation early is crucial to mitigate its societal impact. We establish detection milestones, such as the number of reply posts, to identify content for evaluation only up to these points. We progressively analyze test data chronologically until reaching the desired number of posts. Figure 7.5 illustrates the effectiveness of our approach CrowdShield, compared to two best-performing baselines that use conversation threads – ACLR (281) and BERT$_{claim}$ (denoted as BERT) for early detection. For fewer replies (10, 20, and 30), CrowdShield consistently outperforms both BERT and ACLR. Our Q-learning-based method consistently outperforms other approaches from the beginning, demonstrating its effectiveness in the early detection of misinformation. Our model achieves a high macro-F1 score shortly after its initial dissemination. Using only the first ten responses, our model's

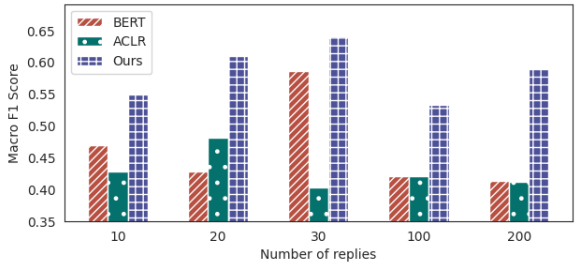


Figure 7.5: Macro-F1 scores are presented for our model `CrowdShield` (indicated by the violet bar with vertical lines) compared to the top baseline systems include conversation threads: ACLR (represented by the green bar with dots) and BERT (depicted by the red bar with diagonal lines). The evaluation was conducted across varying numbers of replies within the conversation thread.

macro-F1 score is 0.54, compared to 0.42 for ACLR and 0.46 for BERT. This vital margin demonstrates our model's ability to understand user response patterns and detect misinformation early on. As the number of replies increases, our approach continues to excel. Our model achieves peak performance following 30 responses to the source post. This is crucial for early misinformation detection, as quick identification can prevent the spread of false information. ACLR shows lower performance at early stages, but it gets the best when the entire thread is provided (cf table 7.2), showing its lack of ability to perform early misinformation detection. On the other hand, BERT performs well with few replies as well, indicating huge limitations in its capacity to utilize the increasing context provided by more replies effectively. Our model's macro-F1 score remains considerably greater than ACLR's and BERT's throughout, especially for fewer replies. This demonstrates both the robustness of our approach and its ability to maintain high accuracy as more user responses are analyzed. In conclusion, `CrowdShield` outperforms existing methods in early detection scenarios and provides a reliable tool for combating misinformation.

Ablation Study. The last two rows of Table 7.2 present the results of our ablation study, highlighting the contributions of different components of our model. We systematically remove key features of `CrowdShield` to determine their impact on overall performance. First, we assess the impact of eliminating the *Q-learning* component. This removal causes a noticeable drop in performance across nearly all metrics. Specifically, the macro-F1 score drops from 0.64 to 0.58. The substantial drop emphasizes the importance of our model's Q-learning mechanism, which effectively captures and applies dynamic patterns of user responses to predict misinformation. Next, we investigate the impact of removing the *Text Features* from the model. The performance drop is even more pronounced in this scenario, with the macro-F1 score dropping by more than ten points. This drastic reduction suggests that linguistic cues are critical for understanding and identifying misinformation. The text embeddings will likely capture nuanced contextual information and subtleties in user language required for accurate misinformation detection. Our ablation study results confirm the critical roles of Q-learning and text embeddings in our model. The Q-learning component helps model discourse dynamics and learn from user interactions, both critical for early detection. Text embeddings, conversely, provide a thorough understanding of the linguistic features in posts and responses, which is critical for distinguishing between accurate and false information.

Additionally, we evaluate our model's performance for different claim weights, denoted by α. As shown in Table 7.3, our experiment results show that setting α to 2 results in the best performance. This weighting strategy recognizes the critical role assertive statements play in determining the integrity of the source post. We observe a substantial decrease in performance when we set it to 1, treating claims

α	Non-misinformation			Misinformation			M-F1
	P	R	F	P	R	F	
1	0.8000	0.8000	0.8000	0.4000	0.4000	0.4000	0.6000
2	0.8125	0.8667	0.8387	0.5000	0.4000	0.4444	0.6416
3	0.7879	0.8667	0.8254	0.4286	0.3000	0.3529	0.5892

Table 7.3: Performance of CrowdShield on MisT with varying values of claim weight (α).

and non-claims as equal importance. This finding highlights the importance of emphasizing claims in our model for accurately predicting misinformation. Claims frequently contain more decisive and informative content that aids in determining the veracity of the information, and failing to give them sufficient weight reduces the model's effectiveness. When we increase α to 3, we see a decrease in performance, suggesting that while claims are essential, over-emphasizing them can potentially obscure valuable contextual information provided by non-claims.

7.6 Error Analysis

To reaffirm our point, detecting misinformation on online social media is finicky due to the profoundly subjective nature of claims mentioned in them and the folksiness of these platforms. Glancing at Table 7.2, we can see that all the systems, including ours, inevitably make mistakes in identifying these posts as misinformation as the F1-score for the true label is much lesser than that of the false. To further comprehend the performance of our proposed model CrowdShield, now we strive to analyze the errors committed by our proposed model qualitatively, CrowdShield. Table 7.4 highlights randomly sampled error cases from our test dataset, their gold labels and predictions from CrowdShield. In addition, for comparison, we furnish the predictions from the best-performing baseline, BERT.

Consider that first post (p_1) about Iran converting a mall into an emergency hospital and the impact of U.S. sanctions; our model successfully identified the content as 'False' (not misinformation), aligning with the gold label. This accuracy suggests that our model robustly understands global contexts and can effectively differentiate between emotionally charged language and factual information. It leveraged thread information to contextualize the statement within a broader discussion, enhancing its judgment accuracy. In contrast, the baseline model, BERT, incorrectly labelled the post as 'True' (misinformation), indicating potential weaknesses. This error could stem from the baseline's insufficient handling of geopolitical nuances and crisis-related content. Integrating conversation thread into the analysis helped our model to provide more contextually aware assessments.

In the second example, the post underscores the seriousness of the coronavirus outbreak and the need for collective responsibility; our model incorrectly classified the post as 'True' (misinformation), which diverges from the gold label 'False' (not misinformation). This error suggests our model may have misinterpreted the post's underlying intent or factual content, possibly due to the emotionally charged call to action or misunderstanding contextual nuances in the user comments. On the other hand, the baseline model BERT accurately classified the post as 'False' (not misinformation), showing better alignment with the gold label. This indicates that the BERT model may have a more effective mechanism for interpreting the factual nature of health-related advisories. Despite being an error relative to the gold label, there is a nuanced aspect to consider that suggests some potential benefits of such false positives, with varied user comments. Public health messaging requires high accuracy and sensitivity, especially during a crisis. A model that errs on the side of caution by marking potentially exaggerated or emotionally charged statements as misinformation might help maintain a stringent check on the quality and reliability of information spread during crucial times. This cautious approach ensures that only well-

	Example	GOLD	OURS	BERT
p_1	**Source Post: Iran turned its most luxurious mall into an emergency hospital, with 3,000 beds to treat coronavirus patients (Meanwhile, the US government keeps escalating its criminal, illegal sanctions on Iran, trying to destroy its economy and prevent it from buying medical equipment).** └ r_1: My heart goes out to them. *(comment, non-claim)* └ r_2: Stay Strong! Keep Safe! Best wish *(comment, non-claim)* └ r_3: And in USA https://t.co/izvTq3ZLct *(comment, non-claim)* └ r_4: I have a feeling this is going to turn into real bad karma for the US. *(comment, non-claim)* └ r_5: Triage in America if the photo is to believed. https://t.co/u3MtLrQ4GZ *(comment, non-claim)*	FALSE	FALSE	TRUE
p_2	**Source Post: There are still people who aren't taking this #coronavirus outbreak seriously. Remember, it's not just about you. Spare a thought for family & loved ones, especially the elderly. You may be young & healthy but there are those whose immune system is compromised. Don't be selfish!** └ r_1: Subhanallah may Allah Almighty save us all from this calamity.....Amin *(comment, non-claim)* └ r_2: Indeed shiekh, i agree. *(support, non-claim)* └ r_3: Is rather unfortunate and sad and this people happen to be people that people looks upto, rely on and respect their words. *(support, non-claim)* └ r_4: tell them pls sir. Our scholar here are telling us #coronavirus its lie and a western propaganda we shouldn't accept it. Shall we sir? *(query, non-claim)* └ r_5: God bless you! I watched a video saying this yesterday. I was just like wow! They think this virus is a joke. *(support,claim)*	FALSE	TRUE	FALSE
p_3	**Source Post: As of yesterday Canada has completed 1.5 times more coronavirus tests than US. On a per capita basis? 13 times** └ r_1: We should not really be comparing our health care system or our Federal Government to the US. It's not a level playing field. *(deny, non-claim)* └ r_2: Not sure of the intent of your tweet. The table is tilted in favour of? *(query, non-claim)* └ r_3: In favour of Canadians. We do not have to worry about the cost of going to the hospital. *(comment, non-claim)* └ r_4: The other half of my position is that our Federal government is...functioning... Dr. Theresa Tam has a clear and consistent message. *(query, non-claim)* └ r_5: Source of the info? *(query, non-claim)*	TRUE	TRUE	FALSE

Table 7.4: Error analysis of the veracity labels of social media posts with the first five replies (chronological). Errors are highlighted in red.

substantiated, clear, and responsibly communicated messages gain widespread traction, thus preventing panic or misinformation.

The last example, about Canada's coronavirus testing compared to the US, we observe that our model correctly classified the claim as 'True,' aligning with the gold label, while the baseline model incorrectly flagged it as 'False.' Comments like *'We should not really be comparing our health care system or our Federal Government to the US'* and *'Not sure of the intent of your tweet'* might suggest doubt or denial in the conversation thread that challenges the factual accuracy of the initial claim made in the source

post. Even though labeled as non-claims, these comments have aided our model in questioning the post's veracity. It indicates that our model may better handle the context or use more reliable external sources to affirm the truthfulness of statistical claims, maintaining its accuracy despite mixed sentiments in the comments.

7.7 Summary

In conclusion, our research makes a substantial contribution to the growing field of combating the widespread spread of misinformation on social media platforms. We presented a novel framework for early misinformation detection that combines crowd intelligence with sophisticated reinforcement learning mechanisms. By incorporating user stances and claims, our model can quickly detect and flag potentially deceptive content, fostering a sense of community and facilitating informed public debate. We demonstrated how using user stances and assertions on social media posts can help to make better decisions about the truth or falsity of posts at an early stage. Furthermore, improved efficiency in misinformation detection may help manual fact-checkers prioritize fact-checking more quickly. One of the major obstacles to misinformation detection in online social media is a lack of a sufficiently comprehensive annotated dataset. As a result, we created `MisT`, a Twitter corpus of manually annotated conversation threads for misinformation detection, complete with replies, stances, and claim labels. We developed a unified architecture, `CrowdShield` which outperformed existing state-of-the-art systems and demonstrated how the abetting RL could be used to understand crowd intelligence better. The results showed that our proposed model outperformed the best-performing baselines by $\geq 0.4\%$ in the macro-F1 score. We acknowledge that we have focused on English in this work; therefore, in future work, we will seek to develop misinformation detection models for other languages, particularly low-resource languages. In the future, we plan to expand the dataset and improve the performance of the minority class. Finally, we intend to expand our efforts to include multimodality, like images, memes, URLs, etc.

8. QUESTION-BASED VERIFICATION

In an era of misinformation, the need for effective fact-checking mechanisms is more urgent than ever. The complexity of real-world claims often necessitates diverse evidence and multi-step reasoning. Despite advancements in automating fact-checking, primarily for synthetic claims from platforms like Wikipedia, a significant gap remains in verifying complex quantitative assertions. To address this critical issue, we introduce QLAIM, a pioneering multi-domain dataset focused exclusively on quantitative claims. It includes 33k fact-checked claims featuring comparative, statistical, interval, and temporal entities, accompanied by detailed metadata and supporting evidence. In conjunction with QLAIM, we present Q2FC, a comprehensive fact-checking framework designed to replicate the investigative rigour of human fact-checkers. Our approach employs controlled question generation to create precise queries that guide the verification process and retrieve relevant responses. This enhances the explanatory power of our model while ensuring data efficiency through clear, human-like inquiries. Empirical evaluations show that our framework significantly outperforms recent fact-checking baselines.

8.1 Introduction

In today's linked world, spreading misinformation online poses a significant challenge, particularly in high-stakes settings such as political elections and public health emergencies (283). The unrestricted spread of false narratives, misleading claims, and distorted statistics can devastate societal systems, causing political turmoil, economic instability, and decreased public trust in fundamental institutions. The sheer volume and speed with which misinformation spreads online makes it a pressing worldwide concern. For the past few decades, platforms such as PolitiFact,[1] Snopes,[2] etc., have worked hard to combat misinformation. However, the rapid growth of social media content has exposed significant limitations in manual fact-checking, which is both time-consuming and labour-intensive. The reliance on manual verification limits their scalability (284). To address this, innovative fact-checking systems are being devised (105; 100; 285; 286; 248; 64). These systems provide a more scalable response to the misinformation epidemic. Despite tremendous progress, one especially difficult aspect of fact-checking remains underdeveloped – verifying statements containing numerical values, comparative references, or temporal expressions, which we collectively refer to as ***quantitative claims***. Several studies advocate that when numbers are included, they affect decision-making (287; 288; 289). This phenomenon is often termed as *numeric-truth effect*. Claims containing numerical data can wield a disproportionately persuasive influence on the audience. This highlights the need for more nuanced techniques to understand and analyze such claims.

We formally define a quantitative claim as *an assertive statement that includes or implies quantitative information ranging from statistical, temporal, or comparative measurements.* These claims require a nuanced understanding for accurate interpretation. Misrepresentation of numeric information, whether intentional or unintentional, can lead to significant misunderstandings. For instance, a seemingly minor change in a percentage in example A in Figure 8.1 can drastically alter the perceived effectiveness

[1] https://www.politifact.com/
[2] https://www.snopes.com/

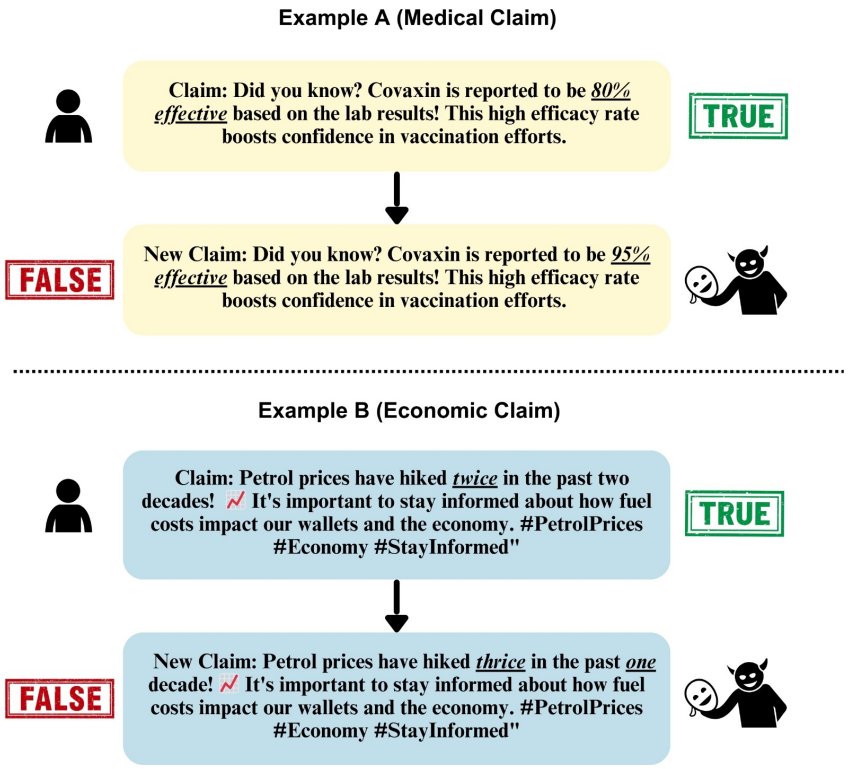


Figure 8.1: Illustrative examples of two quantitative claims from different domains and their subsequent fabricated *false* claims. The quantitative entities are underlined, and malicious users alter them to spread misinformation.

of a policy or product. Thus, the complexity inherent in quantitative data necessitates specialized methodologies for validation and verification. This altered claim can easily bypass existing fact-checking systems, primarily using textual similarity for veracity prediction. In public health, a false claim stating that a vaccine is *'95% effective'* can lead to a significantly different public perception than a claim stating it is *'80% effective.'* While both percentages indicate high effectiveness, a decrease from 95% to 80% may lead some individuals to question the vaccine's reliability, potentially impacting vaccination rates and herd immunity. In example B of Figure 8.1, the economic claim states, *'Petrol prices hiked twice in the past two decades.'* However, if it were presented as *'thrice,'*, the difference in numerical representation could lead to more concern about rising costs. This numerical value disparity not only influences public perception but may also influence policy deliberations and consumer behaviour, emphasizing the crucial relevance of accurate numerical assertions in moulding understanding and reactions to economic concerns.

Over the past few decades, there has been a surge in the development of large-scale, open-source fact-checking datasets (100; 105; 63). This, in turn, has dramatically accelerated the advancement of automated fact-checking. These datasets form the basis for training models capable of interpreting and assessing claims and comparing them to reliable evidence. Despite the prevalence and importance of quantitative claims, existing fact-checking datasets fall short of adequately addressing these claims. The AVERITEC dataset, one of the most recent fact-checking datasets, contains only around 30% of numerical claims (64), while the FEVEROUS dataset has only around 10% (101). As a result, a dedicated dataset for verifying quantitative claims is required. Recognizing this critical gap, we introduce QLAIM, a novel dataset specifically designed to tackle the complexities of verifying quantitative claims in a real-world context. Additionally, we present Q2FC, an innovative framework specifically crafted to verify quantitative claims, surpassing existing fact-checking systems in their effectiveness. With this work, we offer several substantial contributions such as:

- We shed light on the crucial class of *quantitative claims*.
- We develop a carefully curated, high-quality dataset QLAIM, comprising more than 33k quantitative claims.
- We introduce a comprehensive framework Q2FC for fact-checking, focusing on the quantitative entities of the claims.

8.2 Related Work

In recent years, there have been notable breakthroughs in automated fact-checking. Researchers have sought to develop models that can reliably detect misinformation across various domains. The typical pipeline for such systems involves several stages: claim detection (251; 197), evidence retrieval (101), and veracity prediction (248; 64; 79). The focus of most of these efforts has been on textual claims, verified against structured or unstructured data sources. Graph-based models have also been used to facilitate the reasoning over multiple pieces of evidence (76; 77). Although such models achieve sizable performance gains, they lack explainability, and they rely on large amounts of training data. Recent studies indicate that LLMs can perform well and be dependable for verification tasks despite the possibility of hallucinations (78). Lee et al. (79) demonstrated that the inherent knowledge of LLMs can be leveraged for fact verification. Previous research suggests that incorporating external information improves performance on reasoning-intensive tasks (80; 81). Recent advancements aim to simplify complex claims into manageable sub-questions to enhance evidence retrieval, showing promise in improving fact-checking accuracy, particularly for claims involving implicit reasoning or multiple verification steps.

Several fact-checking datasets are available, with the FEVER dataset being one of the most recognized, primarily focusing on textual claims sourced from Wikipedia (100). While FEVER has been instrumental in training fact-checking models, only about 10% of the claims in its FEVEROUS (101) extension involve numerical reasoning. This ongoing deficiency is a common limitation across many datasets, which often rely on synthetic claims or oversimplified representations that fail to reflect real-world complexities. Datasets like TabFact (102), which includes claims derived from Wikipedia tables, and SciTab (103), designed for scientific claims, do not adequately capture the nuances of verifying numerical content in broader contexts. Efforts to compile real-world claims have been more effective in the political sphere, with datasets like ClaimDecomp (104), LIAR (105), and MultiFC (106) featuring claims verified by professional fact-checkers. However, these datasets do not prioritize numerical claims specifically, and their treatment of statistical and temporal expressions is limited. In contrast, earlier works focusing on quantitative claims have had a narrow scope. Recently, Venktesh et al. (107) introduced the QuanTemp dataset that focuses on validating numbers across various domains. However, our proposed dataset goes beyond this by encompassing a broader range of quantitative aspects. While QuanTemp targets verifying explicit numbers only, our dataset also addresses the complexities of comparative, statistical, interval, and temporal elements. This comprehensive approach allows for a more nuanced understanding of claims, making our dataset better suited for developing advanced fact-checking models that tackle real-world quantitative assertions' intricacies.

In recent times, question-answering has emerged as a potential strategy for fact-checking. Yang et al. (155) developed a model that does not require a specific dataset of annotated question-answer pairings. Two recent datasets by Fan et al. (156) and Chen et al. (104) have framed fact-checking as a question-answer task. However, the question-answer pairs used by Fan et al. (156) were primarily designed to give relevant context rather than to capture the entire fact-checking process, resulting in insufficient evidence. Ousidhoum et al. (157) emphasized context dependence as a major issue in

question-answer generation, stating that many queries cannot be generated only from the claim since they refer to entities and events stated only in the original fact-checking articles. Chen et al. (104) sought to establish evidentiality sufficiency but did not demonstrate success. Furthermore, their proof was drawn straight from fact-checking publications published after the assertions, which raises the possibility of temporal leaking. Recently, Schlichtkrull et al. (64) illustrated that reasoning about evidence could be efficiently modelled using an approach based on questions and answers. They engaged human annotators to formulate questions and supply akin responses that included evidence supporting the claim. In contrast, we propose automated quantitative entity-based question generation comparable to human annotators. This strategy has several advantages, as it allows for explanations beyond the facts, corresponds more closely with how humans analyze numerical data and makes it easier to generate more controlled queries at a scale.

8.3 Dataset

In this section, we delve into the meticulous process of creating the QLAIM dataset, explicitly designed to address the challenges of verifying quantitative claims in real-world contexts. The creation of this dataset involves several carefully designed stages, each aimed at ensuring that the final dataset is representative of a wide array of domains and highly suitable for automated fact-checking.

8.3.1 Collection

We initiate our data collection by sourcing claims from trusted fact-checking organizations through the ClaimReview Schema[3], which is licensed under the Creative Commons Attribution-ShareAlike License (version 3.0). We adhere to the terms of this license. The initial collection encompasses a staggering 278,636 fact-checked claims spanning multiple languages and domains. However, to ensure the dataset's consistency and usability, we translate non-English claims into English using Google Translate and drop the languages not identified. One of the primary challenges in collecting these claims is the diversity of labelling conventions used by different fact-checking organizations. To address this, we standardize the labels for all claims to fall under three categories – True, False, or Not Enough Information. Claims with ambiguous labels or those lacking clear classifications were excluded. This standardization mirrors approaches in prior works, ensuring the dataset's compatibility with existing fact-checking pipelines. This refinement process yields a final set of 105,432 claims. After this filtering, we hone in on quantitative claims identification, detailed further in the following subsection.

8.3.2 Quantitative Entity Labelling

A core innovation in constructing the dataset was the identification of quantitative segments within claims, which we term *Quantitative Entity Labelling* (QEL). Quantitative claims are defined as those that include any explicit or implicit numerical, statistical, or temporal content that requires verification against an external source of evidence. We experiment with three off-the-self tools for this task: (a) *Regular Expression (RegEx)*, (b) *Named Entity Recognition (NER) tagging*, and (c) *Part of Speech (POS) tagging*. For the RegEx approach, we identify numerical values and choose a set of pre-defined terms often linked with quantitative statements, such as *increase, decrease, twice, double*, etc. This enables us to successfully catch patterns indicating numerical changes or comparisons within the claims. We employ a pre-trained

[3] https://schema.org/ClaimReview

NER model from the spaCy library to identify quantitative entities. This model has 18 entity types, of which we pick all the numeric types – Date, Time, Percent, Quantity, Ordinal, and Cardinal. A detailed description of each tag and examples is shown in Table 8.1. For Part of Speech (POS) tagging, we use a BERT-based sequence tagger (290) to assess the text's grammatical structure and assign relevant tags. We use Cardinal tags to detect phrases with quantitative features. This approach not only facilitates the identification of numerical references but also enhances our understanding of the context in which these numbers appeared, allowing for a more nuanced interpretation of claims related to quantitative data.

Entity Type	Description	Example
DATE	Absolute or relative dates or periods	Covid-19 was announced as 'Pandemic' on 29 February 2020.
TIME	Times smaller than a day	The meeting is scheduled for 3 pm.
PERCENT	Percentage (including "%")	About 50% of the population voted in the last assembly elections.
MONEY	Monetary values, including unit	The book's price, is $19.99.
QUANTITY	Measurements such as weight or distance	The package weighs 10 kg.
ORDINAL	"first," "second," etc.	She finished in first in the race.
CARDINAL	Numerals that do not fall under another type	There are 25 students in the class.

Table 8.1: Broad overview of NER numeric entity types, along with their descriptions and examples.

Human Evaluation of QEL Approaches. To evaluate the effectiveness of the different QEL approaches, we conduct a human assessment. We seek three annotators to manually identify quantitive entities in 100 random samples from our dataset. The intersection of their selected entities is investigated. We compare these human-labelled entities to those produced with the aforementioned tools. For evaluation, we use the Fuzzy Score[4] and the Jaccard Score.[5] The results of human evaluation are shown in Table 8.2. Since the scores from the Regular Expression (RegEx) and Named Entity Recognition (NER) tagging closely match the human markings, we decided to incorporate both into our QEL process and merge their outputs to form a thorough list of quantitative entities for every claim. Ultimately, any claims without these quantitative entities are removed from the final dataset.

QEL Mechanism	Fuzzy	Jaccard
RegEx	0.7723	0.4272
NER-tagging	<u>0.8460</u>	<u>0.6843</u>
POS-tagging	0.6466	0.2140
RegEx ∪ NER	**0.9168**	**0.7929**

Table 8.2: Human evaluation results of automatic quantitative entity labelling. The best scores are in **bold**, and the second best is <u>underlined</u>.

8.3.3 Statistics and Analysis

Finally, our dataset contains 33,422 quantitative claims. We partition it into an 80:10:10 split for training, development, and testing. Notably, most fact-checked claims are False, highlighting a larger tendency in the fact-checking arena, where fact-checkers frequently prioritize debunking misinformation over

[4] https://github.com/seatgeek/thefuzz
[5] https://en.wikipedia.org/wiki/Jaccard_index

validating true claims. Table 8.3 shows examples from our QLAIM dataset. It shows how we pull out numbers and quantities from different claims. The Claim column contains statements involving numerical or temporal details, while the Quantitative Entity column highlights the extracted numerical entities associated with each claim. Note that there can be multiple entities in one claim. As an example, in the first example, the quantity *[50x, September]* represents how much change happened and when it happened. These are vital to understand the factual basis of the claims.

Claim	**Quantitative Entity**
Virus levels are now 50x higher among secondary school pupils than they were in September	*[50x, September]*
A ghost bus filled with FBI informants dressed as Trump supporters deployed onto our Capitol on January 6th	*[January 6th]*
Over 800 pounds of fentanyl were seized at our Southern Border in October 2023. This is Biden 2019s Border Crisis	*[800 pounds, October 2023]*

Table 8.3: Examples from our dataset, QLAIM, along with their quantitative entities.

Dataset	**Train**	**Dev**	**Test**
Number of claims	26737	3342	3343
Avg. claim length	128.49	129.34	130.33
Avg. questions per claim	1.55	1.54	1.53
Fact-check rating			
▷ False	21631	2704	2705
▷ True	4259	532	533
▷ Not enough information	847	106	105

Table 8.4: Descriptive statistics for the QLAIM dataset.

8.4 Methodology

The more conventional fact-checking methods are built with evidence extraction at their core. Recently, (64) established that reasoning about evidence can be represented through questions and answers. Unlike them, who compose these questions manually, we use an automated approach to generate these human-like questions, focusing on the quantitative elements of the claim. We propose **Q2FC**, (Questioning Quantity for Fact-Checking), based on our perspective on assimilation of the correct questions and evidence. The framework is shown in Figure 8.2.

Q2FC's backbone comprises three sequential modules – controlled question generation, knowledge-grounded response generation, and veracity assessment. First, we denote the input claim as c, which is inherently quantitative. The process begins by generating a set of queries Q that specifically focus on the quantitative entities e present in c. This results in an ordered set of question-entity pairs $Q = \{(q_1, e_1), (q_2, e_2), \ldots, (q_m, e_m)\}$, where each query q_i corresponds to one quantitative entity e_i. Notably, typical question generation algorithms frequently fail to provide the specialized inquiries required for quantitative claims; however, our methodology bridges this gap by ensuring that the generated questions are tailored to extract meaningful information about quantitative entities. Once the queries are defined, we use Large Language Models (LLMs) to retrieve responses for each query, leveraging their sophisticated capabilities to give accurate and contextually relevant information. Finally, we compare the retrieved responses to the original claim c to determine its validity, indicating if the quantitative claim is supported. This systematic methodology enables a strong and efficient verification process for quantitative claims. The following subsections provide more information about each module.

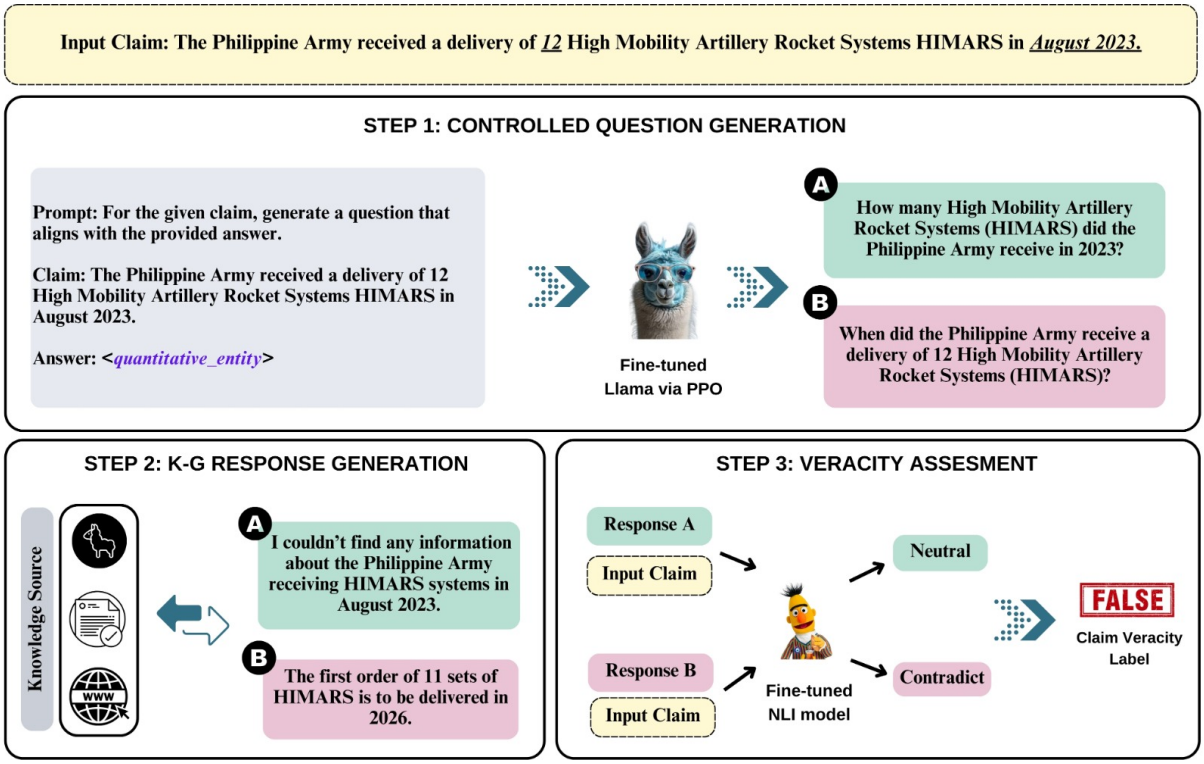


Figure 8.2: A schematic illustration of our approach, Q2FC, for a sample input claim, emphasizing quantitative components.(i) *Step 1*: Create questions for each quantitative entity. (ii) *Step 2*: Use a specific knowledge source to generate a response to the question. (iii) *Step 3*: Determine the final truthfulness label based on the alignment of the claims and the retrieved responses. 'K-G' stands for Knowledge-Grounded.

8.4.1 Controlled Question Generation

Traditional methods of question generation often rely on supervised learning, which may lead to a lack of adaptability and insufficient contextual understanding. In contrast, we employ reward-based controlled question generation to enhance the generation of contextually relevant questions. We generate questions in a zero-shot manner. We then utilize a fine-tuned T5 model[6] to answer them by inputting both the original claim and the generated question. We then use Natural Language Inference (NLI) to determine whether the quantitative entity can be retrieved as the answer. The NLI scores are used as a reward. The Proximal Policy Optimization (PPO) algorithm (291) is employed to iteratively update the model based on these rewards derived from external evaluations. This dynamic learning approach allows the model to adapt and improve over time, effectively generating high-quality questions that are relevant and aligned with the provided context.

The model generates a set of candidate questions for each claim in each training iteration based on prompts derived from the claim-quantitative entity pairs. The reward-based learning allows for continuous refinement of the model's parameters based on the rewards received, facilitating the generation of high-quality questions that are not only contextually aligned but also maximally informative.

[6] https://huggingface.co/MaRiOrOsSi/t5-base-finetuned-question-answering

8.4.2 Knowledge-Grounded Response Generation

Once the questions are framed, we seek to extract their responses from the LLMs. For this module, we employ three distinct response generation setups that depend on the nature of the available knowledge source (K) – closed-book setup, limited-book, and open-book.

- *Close-Book Setup.* In this setup, the model operates without external knowledge sources (K=ϕ), seeking to examine whether LLMs can respond to questions based purely on their encoded information. This strategy takes advantage of insights gathered during pre-training, forcing the model to rely solely on internal knowledge, revealing its capabilities and limitations in responding to created questions.
- *Limited-Book Setup.* In this setup, K comprises a set of fact-checked articles that can explicitly support or refute the claim in question. This setup is similar to traditional claim verification, where the model assesses the veracity of the claim by cross-referencing it against the provided gold evidence. We base our approach on the retrieval-augmented generation (RAG) (292) framework for responding to the generated questions by utilizing verified sources.
- *Open-Book Setup.* In this setting, the model uses web-retrieved documents as its knowledge source. We first create a corpus of the top five documents relevant to each claim in our dataset to optimise the retrieval process. We employ Google Search API[7] for the retrieval. Our model then indexes relevant documents based on the generated question via Facebook AI Similarity Search (FAISS) (293). It then uses the Retrieval-Augmented Generation (RAG) framework (292) to formulate responses. This multi-step process effectively leverages external knowledge, allowing the model to access a broader range of data outside its internal parameters.

8.4.3 Veracity Assessment

Once we have obtained the responses to the questions, we produce final veracity labels through Natural Language Inference (NLI). NLI determines whether a given *'hypothesis'* and *'premise'* logically imply, contradict, or are neutral to one another. Due to its ability to evaluate the logical relationships between statements, it has proved to be a valuable tool for verifying claims. Hereby, we also seek to determine whether our generated responses support or contradict the initial claims. To achieve this, we leverage a cross-encoder model based on DistilRoBERTa, which has been fine-tuned on the Stanford Natural Language Inference (SNLI) dataset[8]. If all of the corresponding responses to a claim support the original assertion, we consider it *True*. We mark the claim *False* if it contradicts any response. When a question stays unanswered, it indicates insufficient information to establish a decision, leading to the *Not Enough Information* label.

8.5 Experiments and Results

In this subsection, we present all the results for various experimental setups based on the following research questions.

[7] https://console.cloud.google.com/marketplace/product/google/customsearch.googleapis.com
[8] https://nlp.stanford.edu/projects/snli/

R1. How well are LLMs equipped to generate questions based on quantitative information? Questions generated for a specific claim can take many different textual forms in real life, largely due to the diverse writing styles employed by various news organizations. Capturing these variances is critical for developing a solid baseline that enables a thorough examination. Manually generating prospective questions is the best way, but it is time-consuming and labour-intensive. Therefore, we seek to investigate how successfully existing LLMs can generate questions centred on quantitative entities when requested. We use the following prompt:

```
For the answer <quantitative-entity>, generate a question for the given claim
<claim>.
```

For each claim, we generate multiple questions based on the number of quantitative entities it contains. We employ several LLMs for this task, including BART, T5, Flan-T5, Gemma-7B, and Llama3. To assess the quality of the questions generated by these models, we do a thorough manual evaluation with 5 annotators examining 75 randomly selected claims and their corresponding generated questions. Each annotator independently marks each claim-question pair using three fundamental linguistic dimensions.

- **Grammatical Correctness:** Assesses the syntactic quality of the generated queries.
- **Factual Alignment:** Assesses how well the questions match the factual content of the claims.
- **Relevance:** Determines how closely the questions are relevant to the quantitative parts of the claim.

Each criterion is scored on a scale of 0 to 5, with 0 indicating the lowest quality and 5 indicating the highest. The average scores of manual evaluation are reported in Table 8.5. Our empirical findings show that Llama3 outperforms the other models across all dimensions. Thus, we use it for further experiments with our framework. Questions generated by the systems for a sample are shown in Table 8.6. BART perform worst among all the systems. While T5 and Flan-T5 demonstrate strong performance, particularly in grammatical correctness. In contrast, Gemma performs notably worse across all metrics. Despite being a capable general-purpose model, Gemma underperforms in the quantitative question generation task due to several limitations. Unlike models such as T5 and LLaMA3, which have been fine-tuned for language generation and instruction-following tasks, Gemma lacks task-specific adaptation. This affects its ability to generate grammatically correct and semantically relevant questions, especially when dealing with structured or numerical content.

Model	Grammatical Correctness	Factual Alignment	Relevance
BART	0.4667	0.4445	0.5112
T5	4.8667	3.6445	3.6889
Flan-T5	4.3442	3.1871	2.1542
Gemma	1.8889	1.1334	1.0000
Llama3	**5.0000**	**4.3556**	**4.2445**

Table 8.5: Average scores of manual evaluation for generated questions.

R2. Which answering setup is most effective for generating responses to quantitative questions? To understand how the availability of diverse knowledge sources facilitates the answering capabilities, we examine the performance of Q2FC across three answering settings, as stated in the preceding section. The experimental results are shown in the bottom three rows of Table 8.7. Each setup provides varying levels of access to external information, which is critical in determining the accuracy and reliability of the responses, particularly for quantitative questions.

Claim	Over *30000* scientists worldwide declare climate change is a hoax.
BART	Is climate change a real hoax according to scientists?
T5	How many scientists believe climate change is a hoax?
Flan-T5	What is the name of the group that declared climate change a hoax?
Gemma	Can you name the scientists who are against climate change?
Llama3	What is the approximate number of scientists who have allegedly declared climate change a hoax?

Table 8.6: Example of a claim and generated questions from various models.

In the close-book arrangement, the model depends solely on its ingrained understanding from previous large-scale training. Although this method is quick and efficient, the static nature of the stored information limits the model's ability to respond accurately to complex or time-sensitive quantitative queries. As a result, the close-book arrangement generates a weighted-F1 score of 0.7043 and an accuracy rate of 0.7586, the lowest among the three setups. On the other hand, the limited-book arrangement grants access to selected articles that have been fact-checked. This significantly enhances the model's capacity to verify information and maintain precision. Using data that has undergone fact-checking, the model secures accuracy and dependability, enabling it to generate highly accurate responses. This configuration performs best among the setups, achieving the highest weighted-F1 score of 0.7107 and an accuracy of 0.7739. The primary factor contributing to its superior performance is the availability of verified sources, which makes the limited-book arrangement the most effective for answering quantitative questions. The open-book setup extends its reach by incorporating information from various online sources, such as Google searches. This provides the model access to more current and diverse knowledge. However, the quality of information varies, as not all sources are equally reliable or verified. At the same time, the advantage of real-time data integration is evident; including less reliable sources slightly impacts the overall accuracy. The open-book arrangement achieves a weighted-F1 score of 0.7056 and an accuracy rate of 0.7613, placing it between the limited-book and close-book arrangements in terms of performance.

Model	**wt-F1**	**Acc**
PROGRAMFC (248)	0.7019	0.7087
AVERITEC (64)	0.6398	0.6186
Q2FC ***(Close-Book)***	0.7043	0.7586
Q2FC ***(Limited-Book)***	**0.7107**	**0.7739**
Q2FC ***(Open-Book)***	0.7056	0.7613

Table 8.7: Experimental results for veracity labels. wt-F1 denotes weighted-F1 scores, while Acc denotes Accuracy. The last three rows show the results of our model with three distinct response generation setups.

R3. How does our framework's performance compare to recent fact-checking systems? As shown in Table 8.7, our Q2FC model consistently outperforms the recent fact-checking systems – PROGRAMFC and AVERITEC, across all answering setups. PROGRAMFC, which achieves a weighted-F1 score of 0.7019 and an accuracy rate of 0.7087, and AVERITEC, with a weighted-F1 score of 0.6398 and an accuracy of 0.6186, were both surpassed by all configurations of our Q2FC model. Notably, the setup for Q2FC close-book that only uses initial model knowledge provides a weighted-F1 score of 0.7043 and an accuracy level of 0.7586, surpassing both PROGRAMFC and AVERITEC even without access to additional resources. The Q2FC limited-book configuration, which uses articles for fact-checking, demonstrates the best overall performance. It achieves a weighted-F1 score of 0.7107 and an accuracy rate as high as 0.7739. The reliability of the checked sources contributes to the configuration's superior performance, further widening the gap compared to both PROGRAMFC and AVERITEC. Lastly, the Q2FC open-book

setup, which incorporates wider internet sources such as Google searches, also exceeds PROGRAMFC, attaining a weighted-F1 score of 0.7056 and an accuracy rate of 0.7613. Despite variations in the quality of external sources, the open-book configuration maintains higher precision than both earlier systems, highlighting the benefits of combining diverse knowledge resources. Our Q2FC model outperforms the recent fact-checking systems –PROGRAMFC and AVERITECfor quantitative fact-checking. The limited-book setup, leveraging reliable external articles, provides the highest precision and accuracy against these baselines. This underscores the importance of credible information sources in enhancing the outcomes of quantitative fact-checking systems.

8.6 Summary

In an era driven by misinformation, the need for effective fact-checking approaches has become more pressing than ever. Through this work, we addressed the significant challenges of verifying quantitative claims, which is generally overlooked in traditional automated fact-checking systems. We developed a novel dataset, QLAIM, comprising more than 33k fact-checked quantitative claims. It encompassed a diverse range of topics and ensured comprehensive coverage of a wide range of numerical contexts – comparative, statistical, interval, and temporal entities. We propose a novel framework, Q2FC, that introduces a fresh approach to fact-checking quantitative claims by mirroring the meticulous investigative methods employed by human fact-checkers. We employ controlled question generation to create quantitative entity-based queries that drive the verification process. Empirical results show that our technique outperforms existing baselines. As we continued to deal with the issues posed by misinformation, our results paved the path for advances in automated fact-checking, ultimately contributing to a more educated public.

9. CONCLUSION

One of the most pressing issues of our time is the proliferation of online misinformation. In the digital world, unverified claims are the primary source of misinformation. Although there is a lot of discussion about misinformation, it is essential to recognize that its spread fundamentally begins with these unverified claims. In order to effectively tackle this issue, we need to concentrate on understanding and addressing these claims at their core. This thesis sought to confront this challenge by introducing a systematic framework designed to dissect and evaluate claims found in online content. By providing a structured approach to analyzing these claims, the thesis aimed to combat the pervasive issue of misinformation better. The framework developed in this thesis consisted of four key stages: (1) claim detection, (2) claim simplification, (3) claim check-worthiness, and (4) claim verification. Each stage served a distinct purpose in addressing various aspects of misinformation, from identifying claims to verifying their factual accuracy. This comprehensive framework aimed to establish a more efficient, scalable, and automated system for fact-checking claims within social media. Ultimately, it sought to improve the quality of information available online and reduce the propagation of falsehoods.

We first addressed the challenge of claim detection, which proved to be a complex task within social media's unstructured and noisy environment. The highly subjective and context-dependent nature of claims made their identification particularly difficult. Often, claims were embedded within lengthy, convoluted sentences or obscured by broader discussions, further complicating their detection. To overcome this, we introduced two novel methods to detect and isolate these claims for effective evaluation. These methods focused on linguistic markers to identify the core assertions requiring validation, thereby establishing a solid foundation for the subsequent stages of the framework. The lack of sufficiently large, high-quality annotated datasets is a significant bottleneck in claim detection on social media platforms. To address this issue, we developed a substantial dataset of approximately 10,000 manually annotated tweets specifically for claim detection. This dataset played a critical role in training and evaluating our detection models, providing a robust resource for future research and practical applications in the field.

Once claims were detected, the next step was simplification. In social media, claims often exist within dense, ambiguous, or poorly structured text, making it essential to isolate the primary assertion from surrounding noise for effective fact-checking. Recognizing this challenge, we introduced two critical methods for simplifying complex claims: extractive and abstractive. These techniques broke down lengthy or convoluted social media posts into more manageable segments, facilitating the more straightforward assessment of a claim's veracity. Claim span identification focused on pinpointing the specific portion of text that constituted the claim, while claim normalization involved rephrasing these assertions into more precise and straightforward language. By simplifying claims, we reduced cognitive load and enhanced the clarity of the information, preparing them for the subsequent steps in the fact-checking pipeline. The methods proposed in this thesis aligned with recent advancements in automated fact-checking.

In an age of pervasive misinformation, fact-checking every claim is neither feasible nor necessary. Instead, prioritizing claims based on their relevance—particularly those related to health, politics, public safety, and other societal issues—is essential. Thus, in the third stage, we focused on selecting check-worthy claims, which was crucial for ensuring that resources were directed toward claims with significant potential impact. This stage introduced an intelligent filtering mechanism that allowed the system to eliminate trivial or inconsequential claims, concentrating efforts on those that warrant immediate attention. By assessing the potential consequences of a claim, the selection process streamlines the fact-checking workflow and enhances its efficiency. This approach aligned with recent advancements in automated fact-checking systems, emphasizing the importance of identifying check-worthy claims. For instance,

automated methods have been developed to assist human fact-checkers in selecting claims that the public has a vested interest in verifying, thereby maximizing the impact of their efforts.

The final stage, claim verification, is the most critical component of the fact-checking process. Verification involved cross-referencing claims with reliable and authoritative sources to confirm or refute their accuracy. This stage went beyond merely assessing the superficial aspects of a claim; it required gathering and analyzing shreds of evidence to establish factual correctness. For verification, we employed two methods: discourse-based and evidence-based. The crowd's reactions and comments on misinformation often provide valuable insights into its accuracy. In this context, we introduced a framework for misinformation detection that leverages the collective intelligence of online users. By integrating user stances and claims, our model was able to rapidly identify and flag potentially deceptive content, promoting community engagement and encouraging informed public discourse.

The approach laid out in this thesis, consisting of detection, simplification, selection, and verification, offered a comprehensive and systematic framework for automated fact-checking. By focusing on claims as the fundamental unit of misinformation, this work aimed to provide a scalable solution to keep pace with the overwhelming daily content on social media platforms. The framework sought to increase the reliability of online information, support transparency in the fact-checking process, and reduce the impact of false claims on public discourse.

9.1 Limitations and Self-Critical Reflection

While this thesis set out to develop a practical and modular framework for automated fact-checking on social media, it is essential that we also reflect on the limitations and challenges we encountered along the way. One of the central challenges we faced was the gap between strong performance on benchmark datasets and the unpredictability of real-world scenarios. While our models performed well in controlled settings, social media content is messy, informal, and constantly evolving. As discussed in Chapters 3 and 4, elements like numerical values, coded language, and satire often made claims harder to detect and interpret accurately. Our simplification module helped clarify many complex or convoluted claims. Still, we observed edge cases where simplification risked changing the original meaning or stripping away important nuance, particularly in ambiguous or multi-layered claims.

We aimed for topical diversity in our datasets, but did not explicitly enforce domain balancing. As a result, topics like politics and health, naturally dominant in social media discourse, might be slightly over-represented. Another critical limitation stemmed from our verification pipeline, like other existing systems, we assumed that reliable evidence was available for every check-worthy claim. While, in reality, evidence may be incomplete or extracted from unreliable sources. These gaps underscore the limitations of automated verification and highlight the importance of assessing the trustworthiness of evidence and not just its presence.

Reflecting on these limitations has been just as crucial as reporting results. They remind us that building automated fact-checking systems is not purely a technical challenge; it's also deeply social and ethical. While our solutions are far from perfect, we believe they offer a thoughtful foundation for future research that prioritises technical performance, fairness, and human oversight.

9.2 Future Work

While this thesis has made significant strides in developing a framework for combating misinformation, several challenges remain that require further exploration and refinement. The nature of misinformation is complex, and future work can build on the findings of this research in several key areas:

- **Real-Time Feedback from Users and Experts:** In this thesis, we predominantly assess models using quantitative criteria like accuracy and F1-scores. However, it is equally important to consider user interactions and feedback for fact-checking systems to be effective in real-world applications. Conducting large-scale user testing can help us understand how these systems can be easily integrated into online platforms and how people perceive and interact with them. Incorporating real-time feedback from both users and professional reviewers may significantly enhance the system's ability to adapt and develop over time.

- **Extending to Other Domains and Languages:** Misinformation is not confined to English-language content. It is a global issue affecting speakers of many languages, each with linguistic nuances and cultural context. Future work should focus on extending the proposed framework to support the verification of claims in multiple languages. Developing models and tools that can handle domain-specific misinformation (e.g., health, politics, science) in diverse linguistic and cultural contexts will broaden the framework's applicability and reach.

- **Multimodal Content and Evidence Sources:** A natural extension of this work is to incorporate multimodal content—such as images, videos, and infographics—into the claim detection and verification process. Misinformation often takes the form of visual content, which can be as misleading as text-based claims. Expanding the framework to include analysis of multimedia content would enhance its robustness. Moreover, verifying claims across broader evidence sources, such as tables, graphs, and info-boxes, would provide a more comprehensive approach to fact-checking.

- **Verifying the Evidence:** This thesis primarily used Wikipedia and Google search as knowledge sources for claim verification. However, verifying claims in the real world often requires consulting multiple evidence sources, including domain-specific databases, academic papers, or expert opinions. Future work should focus on developing systems that can aggregate and reason about evidence from diverse sources while also considering varying degrees of trustworthiness. This becomes especially important when dealing with emerging topics or specialist fields, where knowledge may not be as readily available in mainstream sources. Additionally, there may be instances where claims contradict established norms or scientific consensus, requiring careful reasoning to determine whether such claims should be challenged or accepted.

In summary, while the work presented in this thesis offers a foundational framework for automated fact-checking, there are numerous opportunities for extending and refining this research. By addressing the evolving nature of language, incorporating user feedback, expanding to new domains and languages, and integrating multimodal evidence sources, future work can significantly enhance the effectiveness of misinformation detection and verification systems. The ongoing challenges of misinformation are vast, but with continued research and development, we can move closer to building more reliable, scalable, and context-aware solutions for combating misinformation in the digital age.

REFERENCES

[1] Alam, S. Shaar, F. Dalvi, H. Sajjad, A. Nikolov, H. Mubarak, G. D. S. Martino, A. Abdelali, N. Durrani, K. Darwish, and P. Nakov, "Fighting the covid-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society," 2020.

[2] S. Gupta, P. Singh, M. Sundriyal, M. S. Akhtar, and T. Chakraborty, "LESA: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Apr. 2021, pp. 3178–3188.

[3] H. Herget. [Online]. Available: https://education.nationalgeographic.org/resource/key-components-civilization/

[4] M. Bauwens, "P2p and human evolution: Peer to peer as the premise of a new mode of civilization," *Ensaio, rascunho*, vol. 1, pp. 1–73, 2005.

[5] E. C. Baek, C. Scholz, M. B. O'Donnell, and E. B. Falk, "The value of sharing information: a neural account of information transmission," *Psychological science*, vol. 28, no. 7, pp. 851–861, 2017.

[6] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 30, 2023.

[7] E. Broda and J. Strömbäck, "Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review," *Annals of the International Communication Association*, vol. 48, no. 2, pp. 139–166, 2024.

[8] J. Posetti and A. Matthews, "A short guide to the history of 'fake news' and disinformation," *International Center for Journalists*, vol. 7, no. 2018, pp. 2018–07, 2018.

[9] M. Fraser, *In truth: a history of lies from ancient Rome to modern America*. Rowman & Littlefield, 2020.

[10] W. Mieder, *" Proverbs speak louder than words": folk wisdom in art, culture, folklore, history, literature and mass media*. Peter Lang, 2008.

[11] C. A. Watson, "Information literacy in a fake/false news world: An overview of the characteristics of fake news and its historical development," *International Journal of Legal Information*, vol. 46, no. 2, pp. 93–96, 2018.

[12] I. Kaminska, "A lesson in fake news from the info-wars of ancient rome," *Financial Times*, vol. 17, 2017.

[13] A. Augustyn, "Printing press," Dec. 2018. [Online]. Available: https://www.britannica.com/technology/printing-press

[14] C. Fletcher, "Religious change and print culture in the reformation," 2016, accessed: 05 November 2024. [Online]. Available: https://dcc.newberry.org/?p=14405

[15] I. K. Vida, “The" great moon hoax" of 1835,” *Hungarian Journal of English and American Studies (HJEAS)*, pp. 431–441, 2012.

[16] J. L. Hilton, “Lucian and the great moon hoax of 1835,” *Akroterion*, vol. 50, no. 1, pp. 87–107, 2005.

[17] F. Hudson, *American Journalism 1690-1940*. Psychology Press, 2000.

[18] C. Carey, “Breaking the news: Telegraphy and yellow journalism in the spanish-american war,” *American Periodicals*, pp. 130–148, 2016.

[19] W. Mieder, “Proverbs in nazi germany: The promulgation of anti-semitism and stereotypes through folklore,” *The Journal of American Folklore*, vol. 95, no. 378, pp. 435–464, 1982.

[20] T. Holt, *The deceivers: Allied military deception in the Second World War*. Simon and Schuster, 2007.

[21] K. Narayanaswami, “Analysis of nazi propaganda,” *HIST S-1572: The Holocaust in history, literature, and film*, 2011.

[22] N. Lande, *Spinning History: Politics and Propaganda in World War II*. Skyhorse Publishing, Inc., 2017.

[23] B. Nietzel, “Propaganda, psychological warfare and communication research in the usa and the soviet union during the cold war,” *History of the Human Sciences*, vol. 29, no. 4-5, pp. 59–76, 2016.

[24] M. Galeotti, “Active measures: Russia’s covert geopolitical operations,” *Security Insights*, vol. 31, pp. 1–7, 2019.

[25] J. Orr, *Panic diaries: a genealogy of panic disorder*. Duke University Press, 2006.

[26] D. Westerman, P. R. Spence, and B. Van Der Heide, “Social media as information source: Recency of updates and credibility of information,” *Journal of computer-mediated communication*, vol. 19, no. 2, pp. 171–183, 2014.

[27] A. Geiger, “Social media outpaces print newspapers in the u.s. as a news source,” Apr. 2024. [Online]. Available: https://www.pewresearch.org/short-reads/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/

[28] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, and S. Liu, “Fake news on social media: the impact on society,” pp. 443–458, 2024.

[29] C. Melchior and M. Oliveira, “A systematic literature review of the motivations to share fake news on social media platforms and how to fight them,” pp. 1127–1150, 2024.

[30] K. Sullivan, “Tiktok video of titanic submersible implosion features fake screams,” *Newsweek*, Jun. 2023. [Online]. Available: https://www.newsweek.com/titan-submersible-implosion-screams-tiktok-1811033

[31] S. Subramanian, “The macedonian teens who mastered fake news,” Jan. 2020. [Online]. Available: https://www.wired.com/2017/02/veles-macedonia-fake-news/

[32] R. Jordan, R. Michael, and W. Andrew, “https://www.bloomberg.com/features/2016-how-to-hack-an-election/,” 2016. [Online]. Available: https://www.bloomberg.com/features/2016-how-to-hack-an-election/

[33] M. S. Islam, T. Sarkar, S. H. Khan, A.-H. M. Kamal, S. M. Hasan, A. Kabir, D. Yeasmin, M. A. Islam, K. I. A. Chowdhury, K. S. Anwar *et al.*, "Covid-19–related infodemic and its impact on public health: A global social media analysis," *The American journal of tropical medicine and hygiene*, vol. 103, no. 4, p. 1621, 2020.

[34] V. Suarez-Lledo and J. Alvarez-Galvez, "Prevalence of health misinformation on social media: systematic review," *Journal of medical Internet research*, vol. 23, no. 1, p. e17187, 2021.

[35] G. Fletcher and M. Griffiths, "Digital transformation during a lockdown," *International journal of information management*, vol. 55, p. 102185, 2020.

[36] S. B. Naeem and R. Bhatti, "The covid-19 'infodemic': a new front for information professionals," *Health information and libraries journal*, vol. 37, no. 3, p. 233—239, Sep. 2020. [Online]. Available: https://europepmc.org/articles/PMC7323420

[37] A. Brinkley, "Henry luce and the launch of time magazine," May 2014. [Online]. Available: https://www.vanityfair.com/news/2010/05/time-magazine-henry-luce?srsltid=AfmBOoo7jIe4O2aDoVUHaktMB8jAdqWASerCXvp8_IOArGLtxjtfRaiv

[38] L. Graves and M. Amazeen, "Fact-checking as idea and practice in journalism," *Oxford Research Encyclopedia of Communication*, 2019.

[39] C. Dickey, "The rise and fall of facts," *Columbia Journalism Review*, vol. 23, 2019.

[40] S. H. Smith, *The Fact Checker's Bible: A Guide to Getting It Right*. Anchor, 2004.

[41] B. Adair, "Politifact wins pulitzer," Apr. 2009. [Online]. Available: https://www.politifact.com/article/2009/apr/20/politifact-wins-pulitzer/

[42] W. contributors, "List of fact-checking websites," Nov. 2024. [Online]. Available: https://en.wikipedia.org/wiki/List_of_fact-checking_websites

[43] B. Adair, "Reporters' lab to launch project to promote claimreview - duke reporters' lab," Oct. 2018. [Online]. Available: https://reporterslab.org/reporters-lab-to-launch-project-to-promote-claimreview/

[44] B. D. Oladokun, J. E. Aruwa, G. A. Ottah, and Y. A. Ajani, "Misinformation and disinformation in the era of social media: The need for fact-checking skills," *Journal of Information and Knowledge*, pp. 1–7, 2024.

[45] M. Pérez-Escolar, E. Ordóñez-Olmedo, and P. Alcaide-Pulido, "Fact-checking skills and project-based learning about infodemic and disinformation," *Thinking Skills and Creativity*, vol. 41, p. 100887, Jun. 2021. [Online]. Available: https://doi.org/10.1016/j.tsc.2021.100887

[46] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The covid-19 social media infodemic," *Scientific Reports*, vol. 10, no. 1, Oct. 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-73510-5

[47] N. Hassan, B. Adair, J. T. Hamilton, C. Li, M. Tremayne, J. Yang, and C. Yu, "The quest to automate fact-checking," in *Proceedings of the 2015 computation+ journalism symposium*. Citeseer, 2015.

[48] B. Adair, C. Li, J. Yang, and C. Yu, "Progress toward "the holy grail": The continued quest to automate fact-checking," in *Computation+ Journalism Symposium,(September)*, 2017.

[49] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 178–206, 2022.

[50] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. Da San Martino *et al.*, "Automated fact-checking for assisting human fact-checkers," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence,{IJCAI-21}*. International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 4551–4558.

[51] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks," *PloS one*, vol. 10, no. 6, p. e0128193, 2015.

[52] T. Schuster, A. Fisch, and R. Barzilay, "Get your vitamin c! robust fact verification with contrastive evidence," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 624–643.

[53] S. E. Toulmin, *The uses of argument*. Cambridge university press, 2003.

[54] R. Levy, Y. Bilu, D. Hershcovich, E. Aharoni, and N. Slonim, "Context dependent claim detection," in *Proc. of COLING*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 1489–1500. [Online]. Available: https://aclanthology.org/C14-1141

[55] M. Lippi and P. Torroni, "Context-independent claim detection for argument mining," in *Proc. of IJCAI*, ser. IJCAI'15. AAAI Press, 2015, p. 185–191.

[56] A. Wührl and R. Klinger, "Claim detection in biomedical Twitter posts," in *Proceedings of the 20th Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics, Jun. 2021, pp. 131–142. [Online]. Available: https://aclanthology.org/2021.bionlp-1.15

[57] N. Hassan, F. Arslan, C. Li, and M. Tremayne, "Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1803–1812.

[58] S. Vasileva, P. Atanasova, L. Màrquez, A. Barrón-Cedeño, and P. Nakov, "It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, ser. RANLP '19. Varna, Bulgaria: INCOMA Ltd., Sep. 2019, pp. 1229–1239. [Online]. Available: https://aclanthology.org/R19-1141

[59] D. Wright and I. Augenstein, "Claim check-worthiness detection as positive unlabelled learning," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 476–488.

[60] S. Zhi, Y. Sun, J. Liu, C. Zhang, and J. Han, "Claimverif: a real-time claim verification system using the web and fact databases," in *Proc. of CIKM*, ser. CIKM '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 2555–2558. [Online]. Available: https://doi.org/10.1145/3132847.3133182

[61] L. Wu, Y. Rao, L. Sun, and W. He, "Evidence inference networks for interpretable claim verification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 058–14 066. [Online]. Available: https://www.aaai.org/AAAI21Papers/AAAI-911.WuL.pdf

[62] J. Ma, W. Gao, S. Joty, and K.-F. Wong, "Sentence-level evidence embedding for claim verification with hierarchical attention networks," in *Proc. of ACL*. ACL, 2019, pp. 2561–2571.

[63] Y. Jiang, S. Bordia, Z. Zhong, C. Dognin, M. Singh, and M. Bansal, "Hover: A dataset for many-hop fact extraction and claim verification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3441–3460.

[64] M. Schlichtkrull, Z. Guo, and A. Vlachos, "Averitec: A dataset for real-world claim verification with evidence from the web," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[65] G. Pennycook, A. Bear, E. T. Collins, and D. G. Rand, "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings," *Management science*, vol. 66, no. 11, pp. 4944–4957, 2020.

[66] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[67] S. Lewandowsky, U. K. Ecker, and J. Cook, "Beyond misinformation: Understanding and coping with the "post-truth" era," *Journal of applied research in memory and cognition*, vol. 6, no. 4, pp. 353–369, 2017.

[68] S. Altay, M. Berriche, and A. Acerbi, "Misinformation on misinformation: Conceptual and methodological challenges," *Social media+ society*, vol. 9, no. 1, p. 20563051221150412, 2023.

[69] S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, "Misinformation and its correction: Continued influence and successful debiasing," *Psychological science in the public interest*, vol. 13, no. 3, pp. 106–131, 2012.

[70] C. Wardle and H. Derakhshan, *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Council of Europe Strasbourg, 2017, vol. 27.

[71] R. L. Derr, "The concept of information in ordinary discourse," *Information processing & management*, vol. 21, no. 6, pp. 489–499, 1985.

[72] L. Monsees, "Information disorder, fake news and the future of democracy," *Globalizations*, vol. 20, no. 1, pp. 153–168, 2023.

[73] M. Pérez Escolar, D. Lilleker, A. J. Tapia Frade *et al.*, "A systematic literature review of the phenomenon of disinformation and misinformation," 2023.

[74] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[75] S. M. Alzanin and A. M. Azmi, "Detecting rumors in social media: A survey," *Procedia computer science*, vol. 142, pp. 294–300, 2018.

[76] J. Zhou, X. Han, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Gear: Graph-based evidence aggregating and reasoning for fact verification," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 892–901.

[77] G. Barnabò, F. Siciliano, C. Castillo, S. Leonardi, P. Nakov, G. Da San Martino, and F. Silvestri, "Deep active learning for misinformation detection using geometric deep learning," *Online Social Networks and Media*, vol. 33, p. 100244, 2023.

[78] J. Guan, J. Dodge, D. Wadden, M. Huang, and H. Peng, "Language models hallucinate, but may excel at fact verification," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 1090–1111.

[79] N. Lee, B. Z. Li, S. Wang, W.-t. Yih, H. Ma, and M. Khabsa, "Language models as fact checkers?" in *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, 2020, pp. 36–41.

[80] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, "Active retrieval augmented generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 7969–7992.

[81] S. Yao, J. Zhao, D. Yu, I. Shafran, K. R. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

[82] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

[83] N. Ye, D. Yu, X. Ma, Y. Zhou, and Y. Yan, "Tai: a lightweight network for content-based fake news detection," *Online Information Review*, 2024.

[84] M. K. Jain, D. Gopalani, and Y. K. Meena, "Confake: fake news identification using content based features," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 8729–8755, 2024.

[85] A. Choudhary and A. Arora, "Linguistic feature based learning model for fake news detection and classification," *Expert Systems with Applications*, vol. 169, p. 114171, 2021.

[86] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of WWW*, 2011, pp. 675–684.

[87] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "Welfake: word embedding over linguistic features for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 881–893, 2021.

[88] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on twitter with tree-structured recursive neural networks," in *Proceedings of ACL*, 2018.

[89] M. Dhawan, S. Sharma, A. Kadam, R. Sharma, and P. Kumaraguru, "Game-on: Graph attention network based multimodal fusion for fake news detection," *Social Network Analysis and Mining*, vol. 14, no. 1, p. 114, 2024.

[90] C. Song, K. Shu, and B. Wu, "Temporally evolving graph neural network for fake news detection," *Information Processing & Management*, vol. 58, no. 6, p. 102712, 2021.

[91] J. Wu, W. Xu, Q. Liu, S. Wu, and L. Wang, "Adversarial contrastive learning for evidence-aware fake news detection with graph neural networks," *IEEE TKDE*, 2023.

[92] Z. Kang, Y. Cao, Y. Shang, T. Liang, H. Tang, and L. Tong, "Fake news detection with heterogenous deep graph convolutional network," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2021, pp. 408–420.

[93] X. Sun, H. Yin, B. Liu, Q. Meng, J. Cao, A. Zhou, and H. Chen, "Structure learning via meta-hyperedge for dynamic rumor detection," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 9, pp. 9128–9139, 2022.

[94] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," in *Proceedings of the 24th CIKM*, 2015, pp. 1867–1870.

[95] O. Enayet and S. R. El-Beltagy, "Niletmrg at semeval-2017 task 8: Determining rumour and veracity support for rumours on twitter." in *Proceedings of the 11th SemEval-2017*, 2017, pp. 470–474.

[96] Q. Li, X. Liu, R. Fang, A. Nourbakhsh, and S. Shah, "User behaviors in newsworthy rumors: A case study of twitter," in *Proceedings of ICWSM*, vol. 10, no. 1, 2016, pp. 627–630.

[97] S. Kwon, M. Cha, and K. Jung, "Rumor detection over varying time windows," *PloS one*, vol. 12, no. 1, p. e0168344, 2017.

[98] Y. Wu, J. Yang, L. Wang, and Z. Xu, "Graph-aware multi-view fusion for rumor detection on social media," in *IEEE ICASSP*. IEEE, 2024, pp. 9961–9965.

[99] Y. Liu and Y.-F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Proceedings of AAAI*, vol. 32, no. 1, 2018.

[100] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Fever: a large-scale dataset for fact extraction and verification," *arXiv:1803.05355*, pp. 809–819, 2018.

[101] R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, and A. Mittal, "The fact extraction and verification over unstructured and structured information (feverous) shared task," in *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, 2021, pp. 1–13.

[102] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, and W. Y. Wang, "Tabfact: A large-scale dataset for table-based fact verification," *arXiv preprint arXiv:1909.02164*, 2019.

[103] X. Lu, L. Pan, Q. Liu, P. Nakov, and M.-Y. Kan, "Scitab: A challenging benchmark for compositional reasoning and claim verification on scientific tables," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 7787–7813.

[104] J. Chen, A. Sriram, E. Choi, and G. Durrett, "Generating literal and implied subquestions to fact-check complex claims," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3495–3516.

[105] W. Y. Wang, ""liar, liar pants on fire": A new benchmark dataset for fake news detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 422–426.

[106] I. Augenstein, C. Lioma, D. Wang, L. C. Lima, C. Hansen, C. Hansen, and J. G. Simonsen, "Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 4685–4697.

[107] V. Venktesh, A. Anand, A. Anand, and V. Setty, "Quantemp: A real-world open-domain benchmark for fact-checking numerical claims," in *47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024*. Association for Computing Machinery (ACM), 2024, pp. 650–660.

[108] S. Rosenthal and K. McKeown, "Detecting opinionated claims in online discussions," in *IEEE ICSC*, ser. ICSC '12, IEEE. USA: IEEE Computer Society, 2012, pp. 30–37. [Online]. Available: https://doi.org/10.1109/ICSC.2012.59

[109] J. Daxenberger, S. Eger, I. Habernal, C. Stab, and I. Gurevych, "What is the essence of a claim? cross-domain claim identification," in *Proc. of EMNLP*, 2017, pp. 2055–2066.

[110] R. Levy, S. Gretz, B. Sznajder, S. Hummel, R. Aharonov, and N. Slonim, "Unsupervised corpus–wide claim detection," in *Proc. of Workshop on Argument Mining*, I. Habernal, I. Gurevych, K. Ashley, C. Cardie, N. Green, D. Litman, G. Petasis, C. Reed, N. Slonim, and V. Walker, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 79–84. [Online]. Available: https://aclanthology.org/W17-5110

[111] T. Chakrabarty, C. Hidey, and K. McKeown, "Imho fine-tuning improves claim detection," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 558–563. [Online]. Available: https://aclanthology.org/N19-1054

[112] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, and F. Haouari, "Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media," in *Proc. of Advances in Information Retrieval*. Cham: Springer International Publishing, 2020, pp. 499–507.

[113] E. Williams, P. Rodrigues, and V. Novak, "Accenture at checkthat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models," *arXiv:2009.02431*, 2020.

[114] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv:1907.11692*, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[115] A. Nikolov, G. D. S. Martino, I. Koychev, and P. Nakov, "Team alex at clef checkthat! 2020: Identifying check-worthy tweets with transformer models," *arXiv:2009.02931*, 2020.

[116] G. S. Cheema, S. Hakimov, and R. Ewerth, "Check_square at checkthat! 2020: Claim detection in social media via fusion of transformer and syntactic features," *arXiv: 2007.10534*, 2020.

[117] O. Zaidan, J. Eisner, and C. Piatko, "Using "annotator rationales" to improve machine learning for text categorization," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York: Association for Computational Linguistics, Apr. 2007, pp. 260–267. [Online]. Available: https://aclanthology.org/N07-1033

[118] D. Trautmann, J. Daxenberger, C. Stab, H. Schütze, and I. Gurevych, "Fine-grained argument unit recognition and classification," in *Proc. of AAAI*, vol. 34, no. 05, 2020, pp. 9048–9056.

[119] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, pp. 14 867–14 875, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17745

[120] J. Pavlopoulos, J. Sorensen, L. Laugier, and I. Androutsopoulos, "SemEval-2021 task 5: Toxic spans detection," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 59–69. [Online]. Available: https://aclanthology.org/2021.semeval-1.6

[121] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov, "SemEval-2020 task 11: Detection of propaganda techniques in news articles," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1377–1414. [Online]. Available: https://aclanthology.org/2020.semeval-1.186

[122] G. Chhablani, A. Sharma, H. Pandey, Y. Bhartia, and S. Suthaharan, "NLRG at SemEval-2021 task 5: Toxic spans detection leveraging BERT-based token classification and span prediction techniques," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 233–242. [Online]. Available: https://aclanthology.org/2021.semeval-1.27

[123] S. Coope, T. Farghly, D. Gerz, I. Vulić, and M. Henderson, "Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 107–121. [Online]. Available: https://aclanthology.org/2020.acl-main.11

[124] J. Rusert, "NLP_UIOWA at Semeval-2021 task 5: Transferring toxic sets to tag toxic spans," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 881–887. [Online]. Available: https://aclanthology.org/2021.semeval-1.119

[125] K. Pluciński and H. Klimczak, "GHOST at SemEval-2021 task 5: Is explanation all you need?" in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 852–859. [Online]. Available: https://aclanthology.org/2021.semeval-1.114

[126] Q. Zhu, Z. Lin, Y. Zhang, J. Sun, X. Li, Q. Lin, Y. Dang, and R. Xu, "HITSZ-HLT at SemEval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 521–526. [Online]. Available: https://aclanthology.org/2021.semeval-1.63

[127] V. A. Nguyen, T. M. Nguyen, H. Quang Dao, and Q. Huu Pham, "S-NLP at SemEval-2021 task 5: An analysis of dual networks for sequence tagging," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 888–897. [Online]. Available: https://aclanthology.org/2021.semeval-1.120

[128] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[129] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. of ACL*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: ACL, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[130] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[131] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 9332–9346. [Online]. Available: https://aclanthology.org/2020.emnlp-main.750

[132] P. Utama, J. Bambrick, N. Moosavi, and I. Gurevych, "Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 2763–2776. [Online]. Available: https://aclanthology.org/2022.naacl-main.199

[133] E. Durmus, H. He, and M. Diab, "FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Jul. 2020, pp. 5055–5070. [Online]. Available: https://aclanthology.org/2020.acl-main.454

[134] A. Fan, D. Grangier, and M. Auli, "Controllable abstractive summarization," in *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, A. Birch, A. Finch, T. Luong, G. Neubig, and Y. Oda, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 45–54. [Online]. Available: https://aclanthology.org/W18-2706

[135] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, L. Màrquez, C. Callison-Burch, and J. Su, Eds. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 379–389. [Online]. Available: https://aclanthology.org/D15-1044

[136] Y. Kikuchi, G. Neubig, R. Sasano, H. Takamura, and M. Okumura, "Controlling output length in neural encoder-decoders," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1328–1338. [Online]. Available: https://aclanthology.org/D16-1140

[137] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, and I. Koychev, "A context-aware approach for detecting worth-checking claims in political debates," in *Proc. of RANLP*, 2017, pp. 267–276.

[138] I. Jaradat, P. Gencheva, A. Barrón-Cedeño, L. Màrquez, and P. Nakov, "Claimrank: Detecting check-worthy claims in arabic and english," in *Proc. of NAACL: Demonstrations*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 26–30. [Online]. Available: https://aclanthology.org/N18-5006

[139] S. Vasileva, P. Atanasova, L. Màrquez, A. Barrón-Cedeño, and P. Nakov, "It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction," in *Proc. of RANLP*, 2019, pp. 1229–1239.

[140] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak *et al.*, "Claimbuster: The first-ever end-to-end fact-checking system," *Proc. of the VLDB Endowment*, vol. 10, no. 12, pp. 1945–1948, 2017.

[141] A. Patwari, D. Goldwasser, and S. Bagchi, "Tathya: A multi-classifier system for detecting check-worthy statements in political debates," in *Proc. of CIKM*, 2017, pp. 2259–2262.

[142] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour *et al.*, "Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news," in *CLEF (Working Notes)*, 2021.

[143] P. Nakov, A. Barrón-Cedeño, G. da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani *et al.*, "Overview of the clef–2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection," in *CLEF (working notes)*. Springer, 2022, pp. 495–520.

[144] F. Alam, A. Barrón-Cedeño, G. S. Cheema, S. Hakimov, M. Hasanain, C. Li, R. Míguez, H. Mubarak, G. K. Shahi, W. Zaghouani *et al.*, "Overview of the clef-2023 checkthat! lab task 1 on check-worthiness in multimodal and multigenre content," *Working Notes of CLEF*, 2023.

[145] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," *arXiv:2003.00104*, 2020.

[146] Y. S. Kartal and M. Kutlu, "Tobb etu at checkthat! 2020: Prioritizing english and arabic claims based on check-worthiness." in *CLEF (Working Notes)*, 2020.

[147] F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. Da San Martino, A. Abdelali, H. Sajjad, K. Darwish *et al.*, "Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms." in *Proc. of ICWSM*, 2021, pp. 913–922.

[148] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[149] R. Nogueira and K. Cho, "Passage re-ranking with bert," *arXiv preprint arXiv:1901.04085*, 2019.

[150] J. Thorne and A. Vlachos, "Automated fact checking: Task formulations, methods and future directions," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3346–3359.

[151] T. Yoneda, J. Mitchell, J. Welbl, P. Stenetorp, and S. Riedel, "Ucl machine reading group: Four factor framework for fact finding (hexaf)," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 2018, pp. 97–102.

[152] Y. Nie, H. Chen, and M. Bansal, "Combining fact extraction and verification with neural semantic matching networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6859–6866.

[153] J. Ma, C. Chen, C. Hou, and X. Yuan, "Kapalm: Knowledge graph enhanced language models for fake news detection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 3999–4009.

[154] X. Wu, K.-H. Huang, Y. Fung, and H. Ji, "Cross-document misinformation detection based on event graph reasoning," in *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 2022, pp. 543–558.

[155] J. Yang, D. Vega-Oliveros, T. Seibt, and A. Rocha, "Explainable fact-checking through question answering," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8952–8956.

[156] A. Fan, A. Piktus, F. Petroni, G. Wenzek, M. Saeidi, A. Vlachos, A. Bordes, and S. Riedel, "Generating fact checking briefs," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7147–7161.

[157] N. Ousidhoum, Z. Yuan, and A. Vlachos, "Varifocal question generation for fact-checking," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 2532–2544.

[158] T. Govier, *A practical study of argument*. Cengage Learning, 2013.

[159] A. Whalen and K. Laland, "Conformity biased transmission in social networks," *Journal of Theoretical Biology*, vol. 380, pp. 542–549, 2015.

[160] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of NAACL-HLT*. Minneapolis, Minnesota: ACL, 2019, pp. 4171–4186.

[161] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 6341–6350. [Online]. Available: http://papers.nips.cc/paper/7213-poincare-embeddings-for-learning-hierarchical-representations.pdf

[162] A. Peldszus and M. Stede, "Joint prediction in MST-style discourse parsing for argumentation mining," in *Proc. of EMNLP*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 938–948. [Online]. Available: https://www.aclweb.org/anthology/D15-1110

[163] C. Stab and I. Gurevych, "Parsing argumentation structures in persuasive essays," *Computational Linguistics*, vol. 43, no. 3, pp. 619–659, Sep. 2017. [Online]. Available: https://www.aclweb.org/anthology/J17-3005

[164] Carlson, "Coronavirus tweets," Mar. 2020. [Online]. Available: https://www.kaggle.com/carlsonhoo/coronavirus-tweets

[165] S. Smith, "Coronavirus (covid19) tweets - early april," Apr. 2020. [Online]. Available: https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april

[166] S. Celin, "Covid-19 tweets afternoon 31.03.2020." Apr. 2020. [Online]. Available: https://www.kaggle.com/svencelin/covid19-tweets-afternoon-31032020

[167] E. Chen, K. Lerman, and E. Ferrara, "Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set," *JMIR Public Health Surveill*, vol. 6, no. 2, p. e19273, May 2020. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/32427106

[168] U. Qazi, M. Imran, and F. Ofli, "Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 6–15, 2020.

[169] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, p. 37, 1960.

[170] O. Biran and O. Rambow, "Identifying justifications in written dialogs," in *IEEE ICSC*, vol. 5, no. 04, 2011, pp. 162–168.

[171] O. Biran and O. Rambow, "Identifying justifications in written dialogs by classifying text as argumentative," *International Journal of Semantic Computing*, vol. 05, pp. 363–381, Dec. 2011.

[172] I. Habernal and I. Gurevych, "Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse," in *Proc. of EMNLP*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 2127–2137. [Online]. Available: https://www.aclweb.org/anthology/D15-1255

[173] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv:1301.3781*, 2013.

[174] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of NeurIPS*, 2017, pp. 5998–6008.

[175] Q. Qin, W. Hu, and B. Liu, "Feature projection for improved text classification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8161–8171.

[176] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[177] S. Bonnabel, "Stochastic gradient descent on riemannian manifolds," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2217–2229, 2013.

[178] G. Zheng, S. Mukherjee, X. L. Dong, and F. Li, "Opentag: Open attribute value extraction from product profiles," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1049–1058.

[179] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[180] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[181] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[182] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, pp. 5753–5763, 2019.

[183] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[184] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[185] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[186] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[187] E. Ferrara, "What types of covid-19 conspiracies are populated by twitter bots?" *arXiv preprint arXiv:2004.09531*, 2020. [Online]. Available: https://arxiv.org/abs/2004.09531

[188] J. Margolin, "Fbi warns of potential surge in hate crimes against asian americans amid coronavirus," *ABC News*, vol. 27, 2020.

[189] C. Ziems, B. He, S. Soni, and S. Kumar, "Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis," *arXiv preprint arXiv:2005.12423*, 2020. [Online]. Available: https://arxiv.org/abs/2005.12423v1

[190] C. O'Connor and M. Murphy, "Going viral: doctors must tackle fake news in the covid-19 pandemic," *BMJ*, vol. 369, 2020. [Online]. Available: https://www.bmj.com/content/369/bmj.m1587

[191] J. Daxenberger, S. Eger, I. Habernal, C. Stab, and I. Gurevych, "What is the essence of a claim? Cross-domain claim identification," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 2055–2066. [Online]. Available: https://aclanthology.org/D17-1218

[192] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford *et al.*, "Okapi at trec-3," *Nist Special Publication Sp*, vol. 109, p. 109, 1995.

[193] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier, "CORD-19: The COVID-19 open research dataset," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.nlpcovid19-acl.1

[194] A. Hanselowski, H. Zhang, Z. Li, D. Sorokin, B. Schiller, C. Schulz, and I. Gurevych, "UKP-Athene: Multi-sentence textual entailment for claim verification," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, and A. Mittal, Eds. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 103–108. [Online]. Available: https://aclanthology.org/W18-5516

[195] A. Soleimani, C. Monz, and M. Worring, "BERT for evidence retrieval and claim verification," *Advances in Information Retrieval*, vol. 12036, p. 359, 2020.

[196] E. M. Bender, J. T. Morgan, M. Oxley, M. Zachry, B. Hutchinson, A. Marin, B. Zhang, and M. Ostendorf, "Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages," in *Proc. of LSM)*. Portland, Oregon: ACL, Jun. 2011, pp. 48–57. [Online]. Available: https://www.aclweb.org/anthology/W11-0707

[197] M. Sundriyal, P. Singh, M. S. Akhtar, S. Sengupta, and T. Chakraborty, "Desyr: definition and syntactic representation based claim detection on the web," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ser. CIKM '21. Virtual Event, Queensland, Australia: Association for Computing Machinery, 2021, pp. 1764–1773. [Online]. Available: https://doi.org/10.1145/3459637.3482423

[198] A. Barrón-Cedeno, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, and F. Haouari, "Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media," in *European Conference on Information Retrieval*, ser. ECIR '20, Springer. Lisbon, Portugal: Nature Publishing Group, Apr. 2020, pp. 499–507.

[199] C. Stab and I. Gurevych, "Parsing argumentation structures in persuasive essays," *Computational Linguistics*, vol. 43, no. 3, pp. 619–659, Sep. 2017. [Online]. Available: https://aclanthology.org/J17-3005

[200] E. Aharoni, A. Polnarov, T. Lavee, D. Hershcovich, R. Levy, R. Rinott, D. Gutfreund, and N. Slonim, "A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics," in *Proceedings of the First Workshop on Argumentation Mining*. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 64–68. [Online]. Available: https://aclanthology.org/W14-2109

[201] K. Krippendorff, Y. Mathet, S. Bouvry, and A. Widlöcher, "On the reliability of unitizing textual continua: Further developments," *Quality & Quantity*, vol. 50, no. 6, pp. 2347–2364, 2016.

[202] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Natural language processing using very large corpora*. Springer, 1999, pp. 157–176.

[203] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv:1910.01108*, 2019. [Online]. Available: https://arxiv.org/abs/1910.01108

[204] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, p. 282–289. [Online]. Available: https://openreview.net/forum?id=HkbzGjZOZB

[205] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[206] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2249–2255. [Online]. Available: https://aclanthology.org/D16-1244

[207] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *arXiv preprint arXiv:1611.01603*, 2016. [Online]. Available: https://arxiv.org/abs/1611.01603

[208] Y. Tay, A. T. Luu, A. Zhang, S. Wang, and S. C. Hui, "Compositional de-attention networks," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/16fc18d787294ad5171100e33d05d4e2-Paper.pdf

[209] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[210] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015. [Online]. Available: https://arxiv.org/abs/1508.01991

[211] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*. ACL, Jun. 2019, pp. 4171–4186.

[212] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.5

[213] Q. Zhu, Z. Lin, Y. Zhang, J. Sun, X. Li, Q. Lin, Y. Dang, and R. Xu, "HITSZ-HLT at SemEval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 521–526. [Online]. Available: https://aclanthology.org/2021.semeval-1.63

[214] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[215] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4593–4601. [Online]. Available: https://aclanthology.org/P19-1452

[216] M. E. Peters, M. Neumann, L. Zettlemoyer, and W.-t. Yih, "Dissecting contextual word embeddings: Architecture and representation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 1499–1509. [Online]. Available: https://aclanthology.org/D18-1179

[217] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, May 2017. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211

[218] F. Alam, S. Shaar, F. Dalvi, H. Sajjad, A. Nikolov, H. Mubarak, G. Da San Martino, A. Abdelali, N. Durrani, K. Darwish, A. Al-Homaid, W. Zaghouani, T. Caselli, G. Danoe, F. Stolk, B. Bruntink, and P. Nakov, "Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 611–649. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.56

[219] Y. M. Rocha, G. A. de Moura, G. A. Desidério, C. H. de Oliveira, F. D. Lourenço, and L. D. de Figueiredo Nicolete, "The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review," *Journal of Public Health*, pp. 1–10, 2021.

[220] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, and J. Beltrán, "The CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection," in *Proceedings of the 44th European Conference on IR Research: Advances in Information Retrieval*, ser. ECIR '22. Berlin, Heidelberg: Springer-Verlag, 2022, pp. 416–428. [Online]. Available: https://doi.org/10.1007/978-3-030-99739-7_52

[221] I. Khaldarova and M. Pantti, "Fake news," *Journalism Practice*, vol. 10, no. 7, pp. 891–901, 2016. [Online]. Available: https://doi.org/10.1080/17512786.2016.1163237

[222] R. Gangi Reddy, S. C. Chinthakindi, Z. Wang, Y. Fung, K. Conger, A. ELsayed, M. Palmer, P. Nakov, E. Hovy, K. Small, and H. Ji, "NewsClaims: A new benchmark for claim detection from news with attribute knowledge," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6002–6018. [Online]. Available: https://aclanthology.org/2022.emnlp-main.403

[223] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, and I. Koychev, "A context-aware approach for detecting worth-checking claims in political debates," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, ser. RANLP '17, Varna, Bulgaria, 2017, pp. 267–276. [Online]. Available: https://aclanthology.coli.uni-saarland.de/papers/R17-1037/r17-1037

[224] S. Shaar, N. Babulkov, G. Da San Martino, and P. Nakov, "That is a known lie: Detecting previously fact-checked claims," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Association for Computational Linguistics, Jul. 2020, pp. 3607–3618. [Online]. Available: https://aclanthology.org/2020.acl-main.332

[225] S. Shaar, F. Alam, G. Da San Martino, and P. Nakov, "The role of context in detecting previously fact-checked claims," in *Findings of the Association for Computational Linguistics: NAACL 2022*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1619–1631. [Online]. Available: https://aclanthology.org/2022.findings-naacl.122

[226] S. Shaar, N. Georgiev, F. Alam, G. Da San Martino, A. Mohamed, and P. Nakov, "Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2069–2080. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.151

[227] M. Hardalov, A. Chernyavskiy, I. Koychev, D. Ilvovsky, and P. Nakov, "CrowdChecked: Detecting previously fact-checked claims in social media," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, Eds. Online only: Association for Computational Linguistics, Nov. 2022, pp. 266–285. [Online]. Available: https://aclanthology.org/2022.aacl-main.22

[228] L. Pan, X. Wu, X. Lu, A. T. Luu, W. Y. Wang, M.-Y. Kan, and P. Nakov, "Fact-checking complex claims with program-guided reasoning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 6981–7004. [Online]. Available: https://aclanthology.org/2023.acl-long.386

[229] R. Gangi Reddy, S. C. Chinthakindi, Y. R. Fung, K. Small, and H. Ji, "A zero-shot claim detection framework using question answering," in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, Eds. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 6927–6933. [Online]. Available: https://aclanthology.org/2022.coling-1.603

[230] A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, P. Atanasova, W. Zaghouani, S. Kyuchukov, G. Da San Martino, and P. Nakov, "Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims, Task 2: Factuality," in *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, ser. CEUR Workshop Proceedings, L. Cappellato, N. Ferro, J.-Y. Nie, and L. Soulier, Eds. Avignon, France: CEUR-WS.org, 2018.

[231] L. Konstantinovskiy, O. Price, M. Babakar, and A. Zubiaga, "Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection," *Digital threats: research and practice*, vol. 2, no. 2, pp. 1–16, 2021.

[232] M. Sundriyal, A. Kulkarni, V. Pulastya, M. S. Akhtar, and T. Chakraborty, "Empowering the fact-checkers! automatic identification of claim spans on twitter," in *Proc. of EMNLP*, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 7701–7715.

[233] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[234] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, 2022, pp. 24 824–24 837. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[235] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour *et al.*, "Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news," in *Proceedings of the 12th International Conference of the CLEF Association: Information Access Evaluation Meets Multiliguality, Multimodality, and Visualization*, ser. CLEF '2021, Bucharest, Romania (online), 2021, pp. 264–291.

[236] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, G. Da San Martino, A. Barrón-Cedeño, R. Míguez, T. Elsayed, and P. Nakov, "Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates," in *Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum*, ser. CLEF '2021, Bucharest, Romania (online), 2021.

[237] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, and J. Beltrán, "Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection," in *Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*, ser. CLEF '2022, Bologna, Italy, 2022.

[238] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[239] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2020, pp. 11 328–11 339.

[240] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[241] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909

[242] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proceedings of the International Conference on Learning Representations*, ser. ICLR '19', 2019.

[243] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," *ACM Comput. Surv.*, vol. 56, no. 3, Oct. 2023. [Online]. Available: https://doi.org/10.1145/3617680

[244] C. van der Lee, A. Gatt, E. van Miltenburg, and E. Krahmer, "Human evaluation of automatically generated text: Current trends and best practice guidelines," *Computer Speech & Language*, vol. 67, p. 101151, 2021.

[245] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, no. 2, p. 102025, 2020.

[246] A. Coleman, "'hundreds dead' because of covid-19 misinformation," Aug. 2020. [Online]. Available: https://www.bbc.com/news/world-53755067

[247] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, and L. Huang, "End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2733–2743.

[248] L. Pan, X. Wu, X. Lu, A. T. Luu, W. Y. Wang, M.-Y. Kan, and P. Nakov, "Fact-checking complex claims with program-guided reasoning," *arXiv preprint arXiv:2305.12744*, pp. 6981–7004, 2023.

[249] A. Das, H. Liu, V. Kovatchev, and M. Lease, "The state of human-centered nlp technology for fact-checking," *Information processing & management*, vol. 60, no. 2, p. 103219, 2023.

[250] M. Sundriyal, T. Chakraborty, and P. Nakov, "From chaos to clarity: Claim normalization to empower fact-checking," in *Findings of the EMNLP*, 2023, pp. 6594–6609.

[251] S. Gupta, P. Singh, M. Sundriyal, M. S. Akhtar, and T. Chakraborty, "Lesa: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 3178–3188.

[252] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, and A. Mittal, "The fact extraction and VERification (FEVER) shared task," in *Proc. of Workshop on FEVER*. Brussels, Belgium: ACL, Nov. 2018, pp. 1–9.

[253] A. Savchev, "Ai rational at checkthat! 2022: using transformer models for tweet classification," *CLEF (Working Notes)*, 2022.

[254] S. Agrestia, A. Hashemianb, and M. Carmanc, "Polimi-flatearthers at checkthat! 2022: Gpt-3 applied to claim detection," *CLEF (Working Notes)*, 2022.

[255] J. R. Martinez-Rico, J. Martínez-Romo, and L. Araujo, "Nlp&ir@ uned at checkthat! 2021: Check-worthiness estimation and fake news detection using transformer models." in *CLEF (Working Notes)*, 2021, pp. 545–557.

[256] X. Zhou, B. Wu, and P. Fung, "Fight for 4230 at checkthat! 2021: Domain-specific preprocessing and pretrained model for ranking claims by check-worthiness." in *CLEF (Working Notes)*, 2021, pp. 681–692.

[257] M. M. F. Caceres, J. P. Sosa, J. A. Lawrence, C. Sestacovschi, A. Tidd-Johnson, M. H. U. Rasool, V. K. Gadamidi, S. Ozair, K. Pandav, C. Cuevas-Lou *et al.*, "The impact of misinformation on the covid-19 pandemic," *AIMS Public Health*, vol. 9, no. 2, p. 262, 2022.

[258] E. C. Nisbet, C. Mortenson, and Q. Li, "The presumed influence of election misinformation on others reduces our own satisfaction with democracy," *The Harvard Kennedy School Misinformation Review*, 2021.

[259] Y. Chuai, J. Zhao, and G. Lenzini, "Topic diversity and conspiracy theories shape engagement with covid-19 misinformation on x/twitter," *arXiv preprint arXiv:2401.08832*, 2024.

[260] M. Sun, X. Zhang, J. Ma, S. Xie, Y. Liu, and S. Y. Philip, "Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection," *IEEE TKDE*, 2023.

[261] Q. Liao, H. Chai, H. Han, X. Zhang, X. Wang, W. Xia, and Y. Ding, "An integrated multi-task model for fake news detection," *IEEE TKDE*, vol. 34, no. 11, pp. 5154–5165, 2021.

[262] I. B. Schlicht, E. Fernandez, B. Chulvi, and P. Rosso, "Automatic detection of health misinformation: a systematic review," *Journal of Ambient Intelligence and Humanized Computing*, vol. 15, no. 3, pp. 2009–2021, 2024.

[263] K.-Y. Lin, R. K.-W. Lee, W. Gao, and W.-C. Peng, "Early prediction of hate speech propagation," in *IEEE ICDMW*. IEEE, 2021, pp. 967–974.

[264] K. Zhou, C. Shu, B. Li, and J. H. Lau, "Early rumour detection," in *Proceedings of NAACL-HLT*, 2019, pp. 1614–1623.

[265] T. Chen, X. Li, H. Yin, and J. Zhang, "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in *PAKDD*. Springer, 2018, pp. 40–52.

[266] V. Qazvinian, E. Rosengren, D. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proc. of EMNLP*, 2011, pp. 1589–1599.

[267] Z. Yang, J. Lin, Z. Guo, Y. Li, X. Li, Q. Li, and W. Liu, "Towards rumor detection with multi-granularity evidences: A dataset and benchmark," *IEEE TKDE*, 2024.

[268] S. Hangloo and B. Arora, "Evidence-aware fake news detection: A review," in *IEEE ICACCTech*. IEEE, 2023, pp. 81–86.

[269] M. H. Gad-Elrab, D. Stepanova, J. Urbani, and G. Weikum, "Exfakt: A framework for explaining facts over knowledge graphs and text," in *Proceedings of WSDM*, 2019, pp. 87–95.

[270] L. Fang, K. Feng, K. Zhao, A. Hu, and T. Li, "Unsupervised rumor detection based on propagation tree vae," *IEEE TKDE*, 2023.

[271] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," *PloS one*, vol. 11, no. 3, p. e0150989, 2016.

[272] R. Yang, W. Gao, J. Ma, H. Lin, and B. Wang, "Reinforcement tuning for detecting stances and debunking rumors jointly with large language models," *arXiv preprint arXiv:2406.02143*, 2024.

[273] G. Ma, C. Hu, L. Ge, and H. Zhang, "Dsmm: A dual stance-aware multi-task model for rumour veracity on social networks," *Information Processing & Management*, vol. 61, no. 1, p. 103528, 2024.

[274] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, and L. Derczynski, "Semeval-2019 task 7: Rumoureval 2019: Determining rumour veracity and support for rumours," in *Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019*. Association for Computational Linguistics, 2019, pp. 845–854.

[275] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[276] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[277] S. Grossberg, "Recurrent neural networks," *Scholarpedia*, vol. 8, no. 2, p. 1888, 2013.

[278] Q. Li, Q. Zhang, and L. Si, "eventAI at SemEval-2019 task 7: Rumor detection on social media by exploiting content, user credibility and propagation information," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, Eds., Minneapolis, Minnesota, USA, Jun. 2019, pp. 855–859.

[279] D. Bacciu, F. Errica, A. Micheli, and M. Podda, "A gentle introduction to deep learning for graphs," *Neural Networks*, vol. 129, pp. 203–221, 2020.

[280] H. Lin, P. Yi, J. Ma, H. Jiang, Z. Luo, S. Shi, and R. Liu, "Zero-shot rumor detection with propagation structure via prompt learning," in *Proceedings of AAAI*, vol. 37, no. 4, 2023, pp. 5213–5221.

[281] H. Lin, J. Ma, L. Chen, Z. Yang, M. Cheng, and C. Guang, "Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning," in *Findings of NAACL*, 2022, pp. 2543–2556.

[282] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[283] S. Lewandowsky, J. Cook, U. Ecker, D. Albarracin, P. Kendeou, E. J. Newman, G. Pennycook, E. Porter, D. G. Rand, D. N. Rapp *et al.*, "The debunking handbook 2020," 2020.

[284] J. E. Uscinski and R. W. Butler, "The epistemology of fact checking," *Critical Review*, vol. 25, no. 2, pp. 162–180, 2013.

[285] A. Saakyan, T. Chakrabarty, and S. Muresan, "Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2116–2129.

[286] M. Sundriyal, G. Malhotra, M. S. Akhtar, S. Sengupta, A. Fano, and T. Chakraborty, "Document retrieval and claim verification to mitigate covid-19 misinformation," in *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, 2022, pp. 66–74.

[287] A. D. Gurmankin, J. Baron, and K. Armstrong, "The effect of numerical statements of risk on trust and comfort with hypothetical physician risk communication," *Medical Decision Making*, vol. 24, no. 3, pp. 265–271, 2004.

[288] V. F. Reyna, W. L. Nelson, P. K. Han, and N. F. Dieckmann, "How numeracy influences risk comprehension and medical decision making." *Psychological bulletin*, vol. 135, no. 6, p. 943, 2009.

[289] N. Sagara, *Consumer understanding and use of numeric information in product claims*. University of Oregon, 2009.

[290] S. Hassan, D. Nadir, D. Fahim, A. Firoj, R. K. Abdul, and X. Jia, "Analyzing encoded concepts in transformer language models," in *North American Chapter of the Association of Computational Linguistics: Human Language Technologies (NAACL)*, ser. NAACL '22, Seattle, 2022.

[291] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[292] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[293] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.