DEEP LEARNING FOR RUMOUR DETECTION AND CLAIM VERACITY ASSESSMENT ON SOCIAL MEDIA

CHEUNG TSUN HIN

PhD

The Hong Kong Polytechnic University

2024

The Hong Kong Polytechnic University

Department of Electrical and Electronic Engineering

Deep Learning for Rumour Detection and Claim Veracity Assessment on Social Media

CHEUNG Tsun Hin

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Aug 2023

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

___________________________ (Signed)

CHEUNG Tsun Hin (Name of student)

Acknowledgement

First, I would like to express my profound gratitude to Professor Kenneth Lam for his exceptional guidance, unwavering support, and constant encouragement throughout my PhD journey at PolyU. Under his supervision, I have not only gained a wealth of knowledge in my specific research area but also acquired invaluable insights into research methodology, academic writing, and effective communication. Professor Lam's insightful feedback, constructive criticism, and patient mentoring have played a pivotal role in shaping the quality and direction of my research project.

I would also like to extend my heartfelt appreciation to the members of the research group, whose expertise and enthusiasm have greatly enriched my academic experience. The stimulating discussions and sharing of recent advances in our research group meetings have inspired me to enhance the technical design and implementation of my own project.

Lastly, I would like to express my deepest gratitude to my parents, my brother, and my beloved Karen Lo, for their unwavering support, boundless love, and constant encouragement. Their presence in my life has been an incredible source of motivation and has enabled me to pursue my academic dreams. I am profoundly grateful for the sacrifices they have made.

Abstract

Online social networks, such as Twitter, Facebook, and Weibo, have become crucial platforms for news consumption, but they are also prone to the rapid spread of misinformation, leading to public deception. Therefore, the automatic detection and verification of rumours play a vital role in safeguarding society's trust. This thesis investigates deep learning approaches for rumour detection and claim veracity assessment on social media, encompassing multimodal source-based rumour detection, user credibility-enhanced rumour detection, propagation graph-based rumour verification, and the incorporation of external evidence for veracity assessment.

First, we investigate multimodal rumour detection by focusing on classifying user-generated image-text pairs on social media. To handle the diverse multimedia content, we introduce a novel model called the Crossmodal Bipolar Attention Network (CBAN), which incorporates both positive and negative attention mechanisms. Experimental results demonstrate the superior performance of CBAN, compared to existing methods for multimodal rumour detection. Additionally, the proposed CBAN has shown promising performance in other multimodal image-text classification tasks, including sentiment analysis, sarcasm detection, and hate-speech detection.

Moreover, the thesis presents an early detection approach that utilizes textual claims and source author credibility to identify rumours. By leveraging pretrained language models and transforming author-aware rumour detection into a text classification problem, our proposed method enhances detection accuracy.

Additionally, we introduce a Layer-Wise Parameter-Efficient Tuning (LWPET) strategy to optimize pretrained language model parameters, reducing computation and memory requirements during fine-tuning.

In the pursuit of an efficient stream classification framework for early fine-grained rumour classification based on community response, we introduce the Causal Diffused Graph-Transformer Network (CDGTN). CDGTN incorporates Source-Guided Incremental Attention Pooling (SGIAP) and a Stacked Early Classification Loss (SecLoss) to improve early classification effectiveness. Furthermore, we propose a continued inference algorithm based on prefix-sum to enhance efficiency. Experimental results on multiple datasets confirm the effectiveness and efficiency of CDGTN.

To address the challenge of assessing the veracity of claims on social media, particularly those lacking contextual information, we propose the Dual-Stream Cross-Attention Network (DSCAN). DSCAN combines social response and external evidence using a dual attention mechanism. Experimental results demonstrate the significant performance improvement of DSCAN, which is evaluated on extended datasets containing relevant evidence retrieved from web search engines.

Lastly, this thesis explores the integration of recent conversational-based instruction-following language models with external evidence retrieval for fact-checking purposes. This improves the accessibility of the fact-checking system to more general use. By leveraging search engines to retrieve evidence and enhancing the knowledge of a pretrained language model, our approach, called FactLLaMA, achieves state-of-the-art performance in fact-checking tasks by bridging the gap between model knowledge and up-to-date information.

In summary, the research presented in this thesis contributes significantly to the field of rumour claim detection and claim veracity assessment on social media. The

proposed deep learning techniques and models demonstrate their effectiveness in addressing key challenges, outperforming existing methods on various benchmark datasets. These contributions have important implications for combating misinformation and promoting the dissemination of accurate information on online platforms.

List of Publications

Journal Papers

- **Tsun-Hin Cheung** and Kin-Man Lam, "Crossmodal Bipolar Attention for Multimodal Classification on Social Media," *Neurocomputing*, vol. 514, pp. 1-12, 2022.
- **Tsun-Hin Cheung** and Kin-Man Lam, "Causal Diffused Graph-Transformer Network with Stacked Early Classification Loss for Efficient Stream Classification of Rumours," *Knowledge-Based Systems*, vol. 277, pp. 11807, 2023.

Conference Papers

- **Tsun-Hin Cheung** and Kin-Man Lam, "Simultaneous Fake News and Topic Classification via Auxiliary Task Learning," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Auckland, New Zealand, 2020, pp. 376-380.
- **Tsun-Hin Cheung** and Kin-Man Lam, "FactLLaMA: Optimizing Instruction-Following Language Models with External Knowledge for Automated Fact-Checking." *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Taipei, Taiwan, 2023, pp. 846-853.

Under Review Papers

- **Tsun-Hin Cheung** and Kin-Man Lam, "Unifying Multimodal Source and Propagation Graph for Rumour Detection on Social Media with Missing Features." (Submitted to *IEEE Access*)
- **Tsun-Hin Cheung** and Kin-Man Lam, "Author-Aware Rumour Detection with Layer-Wise Parameter-Efficient Tuning and Incomplete Feature Learning." (Submitted to *IEEE Access)*
- **Tsun-Hin Cheung** and Kin-Man Lam, "Dual-Stream Cross-Attention Network for Claim Veracity Assessment with Social and External Evidence." (Submitted to *Engineering Applications of Artificial Intelligence*)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background of Social Media

In today's digital age, social media platforms have become an integral part of people's lives, transforming the way we communicate, interact, and access information. Platforms such as Facebook, Twitter[1], and Weibo have revolutionized the landscape of news consumption, offering unprecedented speed and accessibility to vast amounts of information. With just a few clicks, users can access news articles, images, and opinions from a wide range of sources around the world. The real-time nature of social media allows for immediate updates and discussions, creating a dynamic and interactive environment for news dissemination.

However, the rise of social media as a primary source of news has also given rise to new challenges. The democratization of content creation and distribution means that anyone can publish and share information, blurring the lines between reliable journalism and unreliable sources [1]. This has led to the emergence of fake

[1] Twitter has been rebranded to X in July 2023. As the study of this thesis was conducting from Sep 2019 to Jun 2023, we use the term Twitter throughout this thesis to affiliate X.

news and rumours, which are intentionally misleading or false information presented as factual news. The widespread dissemination of such misinformation poses significant risks, as it can deceive and mislead the public, impact public opinion, and even influence important events such as elections.

To address this pressing issue, researchers and technologists have been actively exploring effective methods to detect and combat fake news and rumours on social media. Various approaches have been employed, ranging from traditional manual fact-checking by journalists [2] to more advanced machine learning and deep learning techniques [3]. These methods aim to identify and verify the veracity of information circulated on social media, providing users with accurate and reliable news sources.

Given the speed and scale at which information spreads on social media, it has become crucial to develop automated systems that can quickly and accurately detect and classify fake news and rumours. These systems leverage the power of artificial intelligence and natural language processing to analyse the content, context, and propagation patterns of information on social media platforms. By employing sophisticated algorithms and models, they aim to differentiate between trustworthy news and deceptive content, enabling users to make informed decisions and navigate the complex landscape of online information.

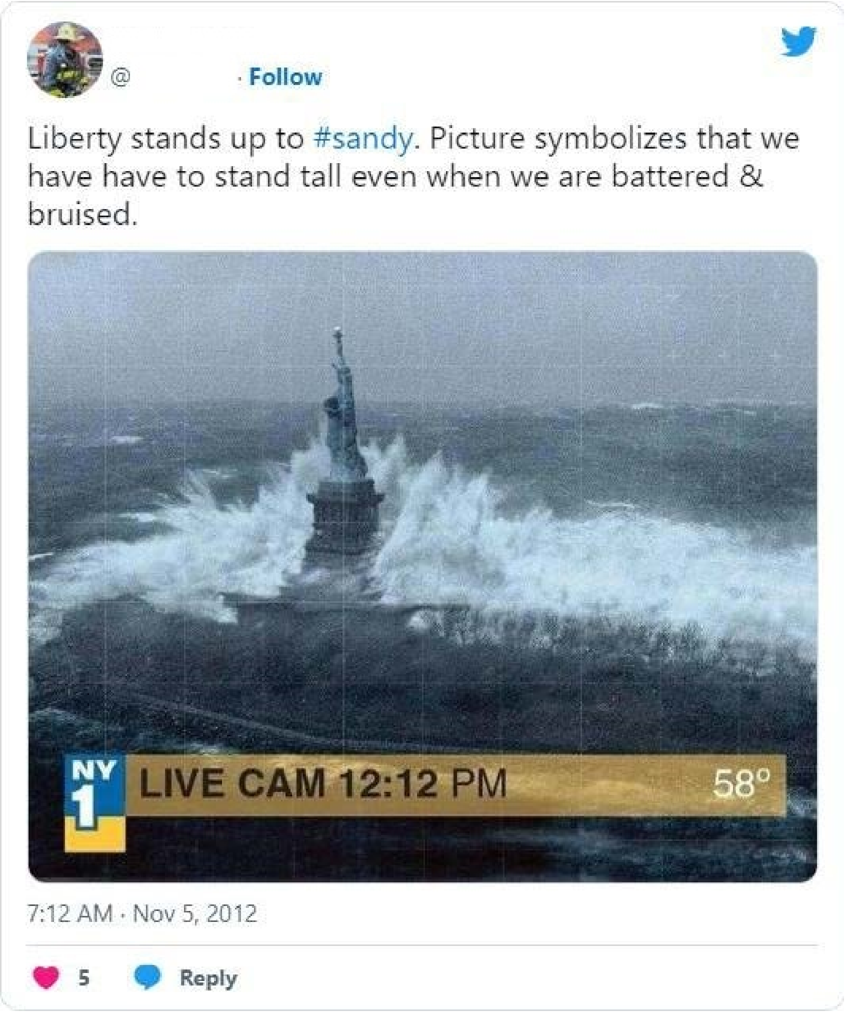
1.2 Misinformation Detection on social Media

Manual fact-checking has long been employed as a means of verifying the accuracy of news and claims. However, given the massive volume of information shared on social media platforms, manual fact-checking alone is insufficient to address the scale and speed at which misinformation spreads. As a result, researchers have turned to computational methods to assist in the detection of fake news and rumours.

Machine learning techniques have been widely adopted for automated misinformation detection. These methods leverage features, such as textual content, user profiles, and social network properties, to train models that can classify information as either true or false. By learning patterns from large datasets, machine learning algorithms can identify suspicious or misleading content and flag it for further investigation.

More recently, deep learning models have shown promise in enhancing the accuracy of misinformation detection. Deep learning algorithms, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), can capture complex patterns and dependencies in textual data, allowing for a more nuanced analysis of fake news and rumours. These models can learn representations that capture semantic meaning, context, and even subtle linguistic cues indicative of misinformation.

Nowadays, online news usually contains multimodal content with text and images. To identify rumours on social media, both the visual and textual features of a source message and its replies in the conversations are important. In general, multimodal rumours usually contain forged or computer-generated imagery together with a text description. For example, as shown in Fig. 1.1, the multimodal source is a misleading message accompanied by a fake picture purporting to be Hurricane Sandy in 2012, which was a rumour spreading on social media. The attached image is computer-generated imagery, which was captured from a movie. This combination of textual and visual features is an important cue for the model to accurately distinguish rumours from real news.

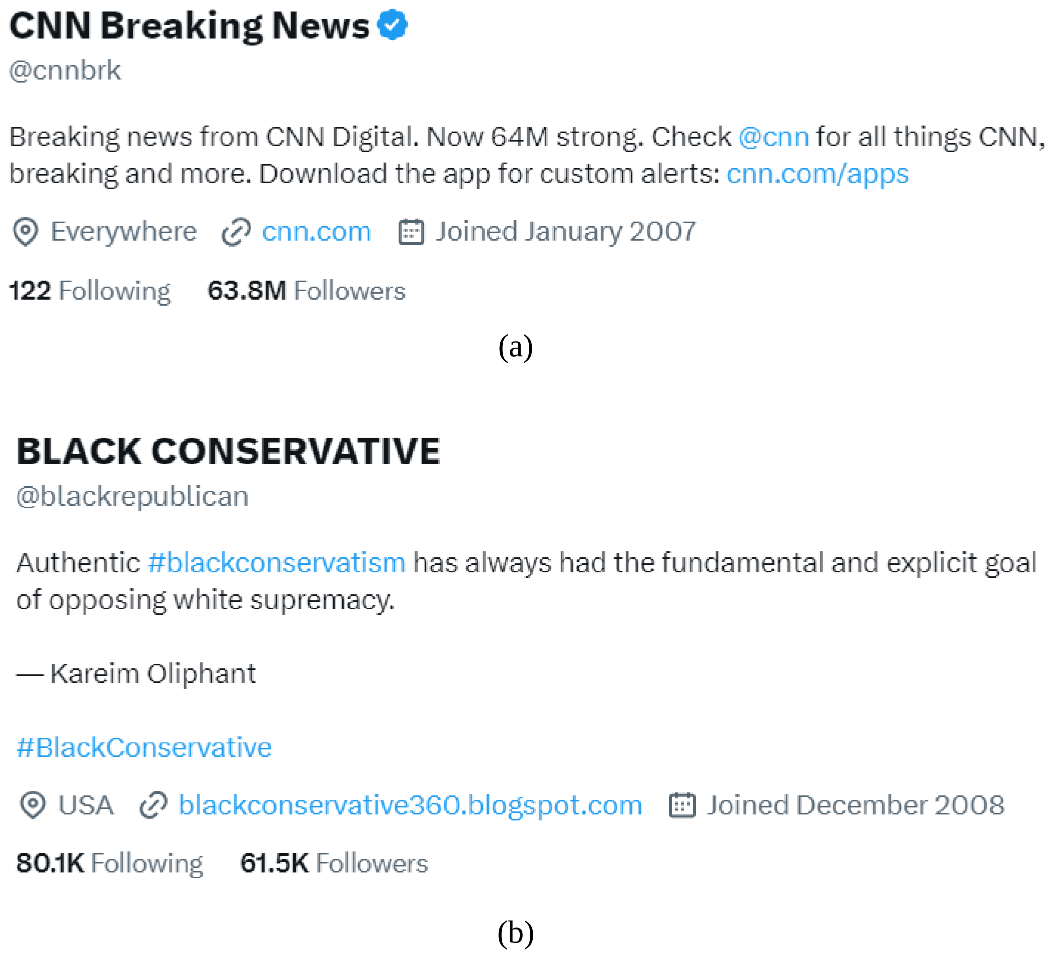


Liberty stands up to #sandy. Picture symbolizes that we have have to stand tall even when we are battered & bruised.

7:12 AM · Nov 5, 2012

Figure 1.1 Example of a multimodal rumour, i.e., image-text pair, on social media.

In addition to the multimodal source, replies usually contain opinions and judgements about the veracity of the source information, which are very useful indicators of unconfirmed and false information. As shown in Fig. 1.2, the replies contain some important phrases, such as “really?” and “fake”, expressing doubt and disagreement with the source information. These linguistic features are very important cues for learning models to classify rumours, especially those appearing on social media.



CNN Breaking News
@cnnbrk

Breaking news from CNN Digital. Now 64M strong. Check @cnn for all things CNN, breaking and more. Download the app for custom alerts: cnn.com/apps

Everywhere cnn.com Joined January 2007

122 Following 63.8M Followers

(a)

BLACK CONSERVATIVE
@blackrepublican

Authentic #blackconservatism has always had the fundamental and explicit goal of opposing white supremacy.

— Kareim Oliphant

#BlackConservative

USA blackconservative360.blogspot.com Joined December 2008

80.1K Following 61.5K Followers

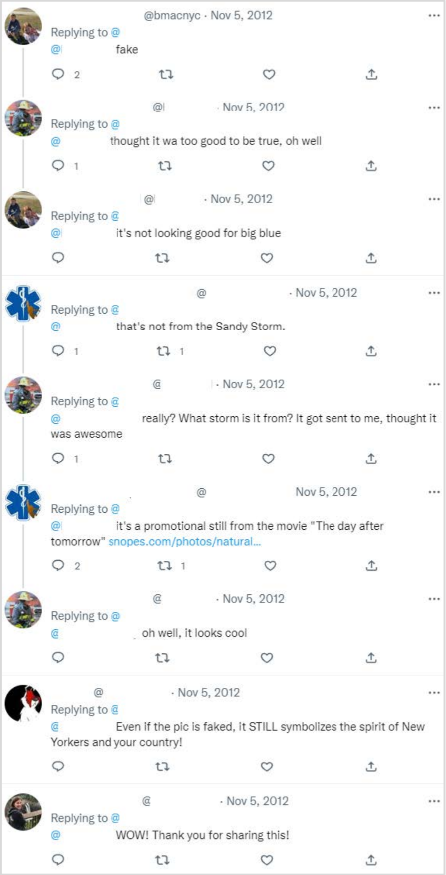
(b)

Figure 1.2 User credibility between (a) real news and (b) fake news spreaders.

Multimodal source content, considering the credibility of social media users plays a crucial role in effective rumour detection. The credibility of a user is an important factor as it reflects their trustworthiness and reliability in disseminating information. In the context of rumour propagation, credible users are more likely to share accurate and verified information, while less credible users may unknowingly or intentionally spread false or misleading claims.

To illustrate the impact of user credibility on rumour propagation, two examples of user descriptions are shown in Figure 1.2. Figure 1.2 (a) represents a real news spreader, who is a user with a high level of credibility. This user is known for sharing verified information from reliable sources and has built a reputation for being accurate in their posts. As a result, when this user shares a piece of news or debunks a rumour, their followers are more likely to trust and share the information, contributing to the propagation of accurate information.

On the other hand, Figure 1.2 (b) depicts a fake news spreader, who is a user with low credibility. This user may intentionally or unintentionally disseminate false information, relying on sensationalism or personal biases. When such a user shares a rumour or false claim, it can quickly spread among their followers who may not critically evaluate the information. Consequently, the dissemination of inaccurate or misleading information can lead to the rapid propagation of rumours on social media platforms.



@bmacnyc · Nov 5, 2012
Replying to @
@ fake

@ · Nov 5, 2012
Replying to @
@ thought it wa too good to be true, oh well

@ · Nov 5, 2012
Replying to @
@ it's not looking good for big blue

@ · Nov 5, 2012
Replying to @
@ that's not from the Sandy Storm.

@ · Nov 5, 2012
Replying to @
@ really? What storm is it from? It got sent to me, thought it was awesome

@ Nov 5, 2012
Replying to @
@ it's a promotional still from the movie "The day after tomorrow" snopes.com/photos/natural...

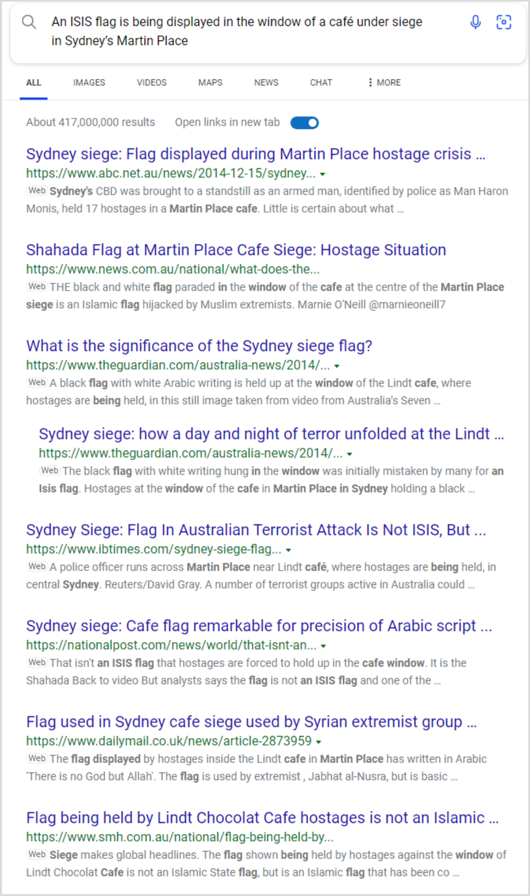
@ · Nov 5, 2012
Replying to @
@ oh well, it looks cool

@ · Nov 5, 2012
Replying to @
@ Even if the pic is faked, it STILL symbolizes the spirit of New Yorkers and your country!

@ · Nov 5, 2012
Replying to @
@ WOW! Thank you for sharing this!

Figure 1.3 A real example of propagation graph, i.e., replies, of a rumour on social media.

Propagation-graph refers to the community response diffusion patterns in the social media threads. As shown in Fig. 1.3, the replies include phrases such as "really?" and "fake," expressing doubt and disagreement with the source information. These linguistic features serve as crucial indicators for learning models to classify rumours, especially within the context of social media. As time goes, the veracity of information could be identified as true or false, based on later authoritative sources or crowdsourced fact-checking, or remains unverified if no evidence supports or rejects the claim.



An ISIS flag is being displayed in the window of a café under siege in Sydney's Martin Place

About 417,000,000 results

Sydney siege: Flag displayed during Martin Place hostage crisis ...
https://www.abc.net.au/news/2014-12-15/sydney...
Web **Sydney's** CBD was brought to a standstill as an armed man, identified by police as Man Haron Monis, held 17 hostages in a **Martin Place cafe**. Little is certain about what ...

Shahada Flag at Martin Place Cafe Siege: Hostage Situation
https://www.news.com.au/national/what-does-the...
Web THE black and white **flag** paraded **in** the **window** of the **cafe** at the centre of the **Martin Place siege** is an Islamic **flag** hijacked by Muslim extremists. Marnie O'Neill @marnieoneill7

What is the significance of the Sydney siege flag?
https://www.theguardian.com/australia-news/2014/...
Web A black **flag** with white Arabic writing is held up at the **window** of the Lindt **cafe**, where hostages are **being** held, in this still image taken from video from Australia's Seven ...

Sydney siege: how a day and night of terror unfolded at the Lindt ...
https://www.theguardian.com/australia-news/2014/...
Web The black **flag** with white writing hung **in** the **window** was initially mistaken by many for **an Isis flag**. Hostages at the **window** of the **cafe** in **Martin Place in Sydney** holding a black ...

Sydney Siege: Flag In Australian Terrorist Attack Is Not ISIS, But ...
https://www.ibtimes.com/sydney-siege-flag...
Web A police officer runs across **Martin Place** near Lindt **café**, where hostages are **being** held, in central **Sydney**. Reuters/David Gray. A number of terrorist groups active in Australia could ...

Sydney siege: Cafe flag remarkable for precision of Arabic script ...
https://nationalpost.com/news/world/that-isnt-an...
Web That isn't **an ISIS flag** that hostages are forced to hold up in the **cafe window**. It is the Shahada Back to video But analysts says the **flag** is not **an ISIS flag** and one of the ...

Flag used in Sydney cafe siege used by Syrian extremist group ...
https://www.dailymail.co.uk/news/article-2873959
Web The **flag displayed** by hostages inside the Lindt **cafe** in **Martin Place** has written in Arabic 'There is no God but Allah'. The **flag** is used by extremist , Jabhat al-Nusra, but is basic ...

Flag being held by Lindt Chocolat Cafe hostages is not an Islamic ...
https://www.smh.com.au/national/flag-being-held-by...
Web **Siege** makes global headlines. The **flag** shown **being** held by hostages against the **window** of Lindt Chocolat **Cafe** is not an Islamic State **flag**, but is an Islamic **flag** that has been co ...

Figure 1.4 A real example of external evidence retrieved from a search engine.

Apart from propagation-based rumour verification, external evidence-based fact-checking has become another paradigm attracting the research community in recent years. It is the task of evaluating the veracity of claims, which can be made in written or spoken language [4]. Evidence retrieval is the first step in a fact-checking process, and is used to find relevant sources that support or refute the claim. As shown in Fig. 1.4, the websites contain addition context that is useful for the model to determine the veracity of the input claim. The deep learning model predicts the veracity of the input claim with external evidence.

1.3 Scope of the Thesis

Building upon the advancements in machine learning and deep learning, this thesis aims to contribute to the field of misinformation detection by proposing novel methods to detect and combat fake news and rumours on social media platforms. The focus is on developing advanced deep learning techniques that can effectively analyse textual content, user interactions, and network properties to discern between genuine and misleading information. By harnessing the power of deep learning models, the objective is to enhance the accuracy and efficiency of misinformation detection, ultimately enabling users to make more informed decisions and promoting a more trustworthy information ecosystem. An overview of rumour claim detection and claim veracity assessment is shown in Fig. 1.5.

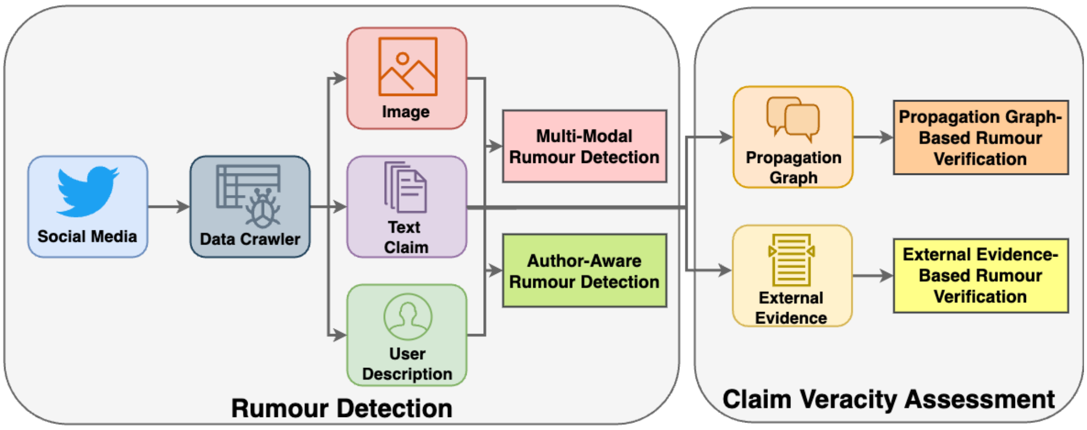

Social Media
Data Crawler
Image
Text Claim
User Description
Multi-Modal Rumour Detection
Author-Aware Rumour Detection
Rumour Detection
Propagation Graph
External Evidence
Propagation Graph-Based Rumour Verification
External Evidence-Based Rumour Verification
Claim Veracity Assessment

Figure 1.5 An overview of the rumour claim detection and claim veracity assessment system on social media.

1.4 Objectives and Research Questions

The research conducted in this thesis aims to address the following objectives and research questions:

- Develop novel deep learning architectures specifically designed for the detection of fake news and rumours on social media platforms.
- Investigate the effectiveness of multimodal fusion techniques, combining textual, visual, and user interaction data, to improve the accuracy of misinformation detection.
- Explore the integration of external evidence, such as fact-checking databases or domain-specific knowledge, to enhance the veracity assessment of claims on social media.
- Evaluate the proposed methods on benchmark datasets and compare their performance against existing state-of-the-art methods.
- By addressing these objectives and research questions, this thesis aims to advance the field of misinformation detection and contribute to the development of more robust and accurate methods for combating fake news and rumours on social media platforms.

1.5 Overview of the Thesis Structure

This thesis is organized as follows:

In Chapter 2, we provide a comprehensive review of existing studies related to fake news and rumour detection. We discuss various methodologies, including manual fact-checking, machine learning-based approaches, deep learning models, multimodal source classification, propagation graph classification, the integration of external evidence, and comparative studies. By examining the strengths and weaknesses of these approaches, we identify research gaps and lay the foundation for our proposed methods.

In Chapter 3, we present our proposed multimodal rumour detection framework. By combining textual and visual information, we aim to leverage the complementary cues present in multiple modalities to enhance the accuracy and robustness of rumour detection. We describe the architecture, feature extraction techniques, and model training procedures in detail, along with experimental evaluations on benchmark datasets.

In Chapter 4, we focus on the role of author representation in the rumour source and develop an author-aware analysis approach for rumour detection. By considering author profiles, we aim to capture the influence and credibility of individuals in spreading misinformation. We present the methodology, feature engineering techniques, and model training process, followed by comprehensive evaluations of real-world datasets.

In Chapter 5, we address the challenge of detecting rumours in real-time by developing a stream classification approach. By considering the temporal dynamics and evolving nature of information spread, we aim to provide timely detection and response to emerging rumours. We present the stream classification framework, feature selection strategies, and model adaptation techniques, accompanied by evaluations of streaming datasets.

In Chapter 6, we explore the integration of multiple sources of evidence for claim veracity assessment. By leveraging both textual information and external knowledge sources, we aim to enhance the accuracy and reliability of evaluating the truthfulness of claims. We discuss the methodology, evidence fusion techniques, and evaluation results on diverse claim datasets.

In Chapter 7, we investigate the use of external knowledge to improve the performance of pretrained instruction-following models. By incorporating external knowledge sources, such as domain-specific ontologies or semantic databases, we aim to enhance the model's understanding and reasoning capabilities. We present the tuning framework, knowledge integration techniques, and empirical evaluations on benchmark datasets.

In Chapter 8, we summarize the contributions of this thesis and discuss the key findings and insights gained from the research. We also highlight potential avenues for future work and extensions of the proposed methods. The chapter concludes with a reflection on the impact and significance of the research conducted.

Through the exploration and development of these novel techniques, this thesis aims to advance the field of rumour detection and claim veracity assessment on social media platforms. By addressing the challenges posed by fake news and rumours, we strive to contribute to the creation of a more informed and trustworthy online environment.

Chapter 2

Literature Review

In this chapter, we delve into the extensive research related to the detection and mitigation of fake news and rumours on social media platforms. Our exploration encompasses a wide range of methodologies, including manual fact-checking, machine learning-based techniques, deep learning models, integration of external evidence for claim veracity assessment, multimodal source classification, propagation graph classification, and comparative studies. By examining a diverse array of approaches, this review aims to provide a comprehensive understanding of the current state of the field, identify research gaps, and lay the groundwork for our proposed methods.

2.1 Misinformation on Social Media

The proliferation of fake news and misinformation has become a pressing concern in today's digital age. This subsection provides an overview of the background, impact, and characteristics of fake news and misinformation.

Fake news refers to intentionally false or misleading information presented as legitimate news. It often aims to deceive readers and manipulate public opinion. Fake news can take various forms, including fabricated stories, misleading headlines, manipulated images, and distorted facts [5]. It exploits the viral nature of social media platforms to rapidly reach a wide audience. Misinformation encompasses a broader spectrum of false information, including rumours, hoaxes, conspiracy theories, and urban legends. Misinformation can spread unintentionally, often due to misunderstandings, cognitive biases, or lack of fact-checking. It can originate from various sources, such as individuals, media organizations, political groups, or foreign actors seeking to influence public opinion. The dissemination of fake news and misinformation can have profound societal implications. It can undermine trust in traditional media, erode democratic processes, and contribute to the polarization of society. False information can sway public opinion, incite social unrest, and harm individuals, organizations, and even economies. Recognizing the potential harm, researchers and policymakers have recognized the importance of developing effective strategies to combat fake news and misinformation. The timeline of misinformation is shown in Fig. 2.1.

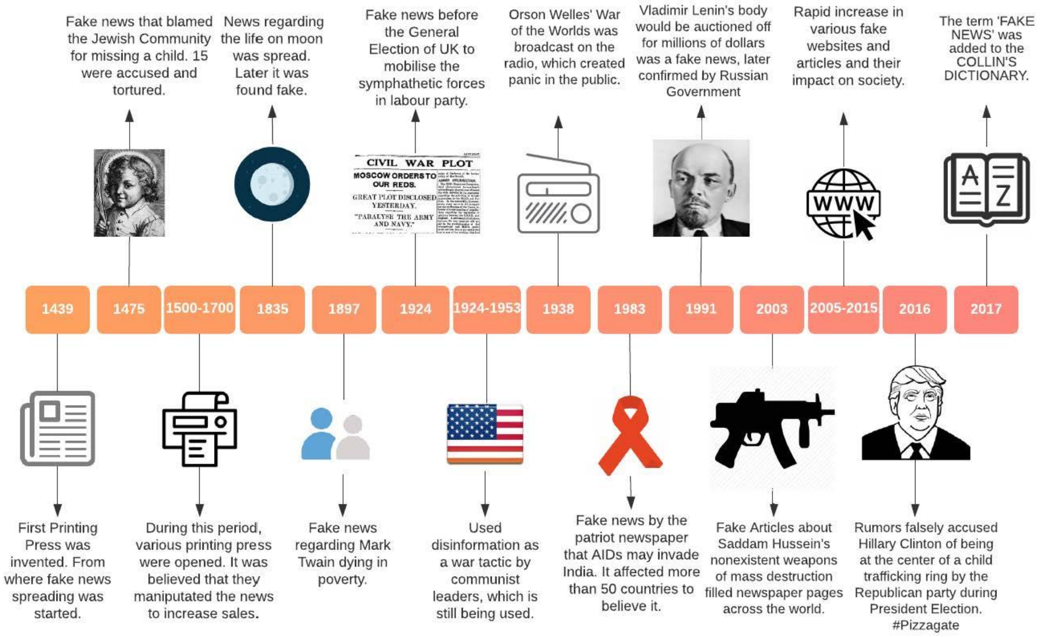

Fake news that blamed the Jewish Community for missing a child. 15 were accused and tortured.
News regarding the life on moon was spread. Later it was found fake.
Fake news before the General Election of UK to mobilise the symphathetic forces in labour party.
CIVIL WAR PLOT
MOSCOW ORDERS TO OUR REDS.
GREAT PLOT DISCLOSED YESTERDAY.
"PARALYSE THE ARMY AND NAVY."
Orson Welles' War of the Worlds was broadcast on the radio, which created panic in the public.
Vladimir Lenin's body would be auctioned off for millions of dollars was a fake news, later confirmed by Russian Government
Rapid increase in various fake websites and articles and their impact on society.
WWW
The term 'FAKE NEWS' was added to the COLLIN'S DICTIONARY.
A
Z
1439
1475
1500-1700
1835
1897
1924
1924-1953
1938
1983
1991
2003
2005-2015
2016
2017
First Printing Press was invented. From where fake news spreading was started.
During this period, various printing press were opened. It was believed that they maniputated the news to increase sales.
Fake news regarding Mark Twain dying in poverty.
Used disinformation as a war tactic by communist leaders, which is still being used.
Fake news by the patriot newspaper that AIDs may invade India. It affected more than 50 countries to believe it.
Fake Articles about Saddam Hussein's nonexistent weapons of mass destruction filled newspaper pages across the world.
Rumors falsely accused Hillary Clinton of being at the center of a child trafficking ring by the Republican party during President Election. #Pizzagate

Figure 2.1 Timeline for evolution of fake news [6].

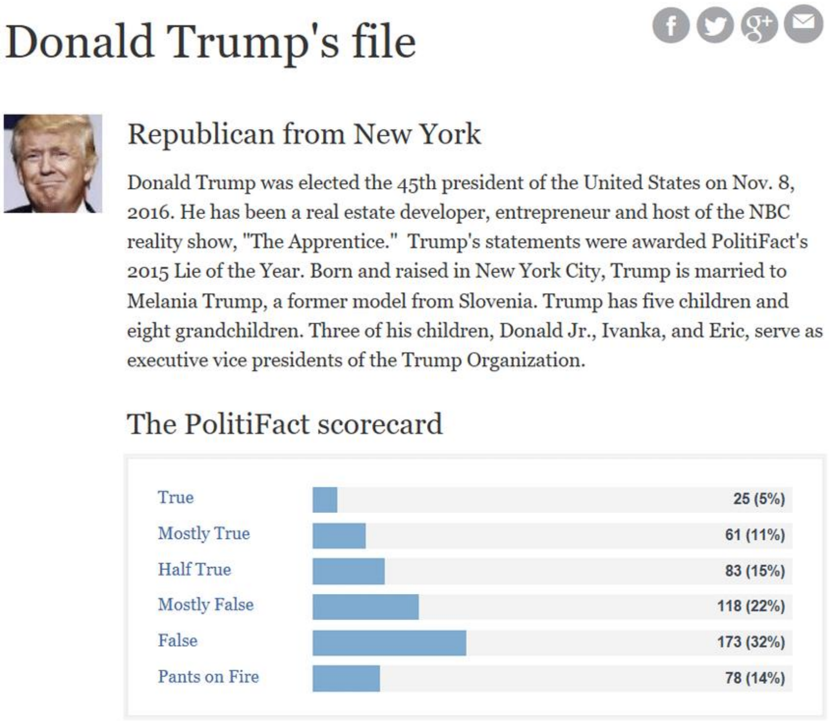
Detecting fake news and misinformation presents significant challenges due to several factors. First, the sheer volume of information circulating on social media platforms makes it challenging to identify false or misleading content [7]. Additionally, the rapid speed at which news spreads online requires prompt detection and response [8]. The evolving tactics used by purveyors of fake news, such as disinformation campaigns and deepfakes, further complicate detection efforts. Understanding the psychological factors and cognitive biases that influence individuals' susceptibility to fake news and misinformation is crucial. Confirmation bias, where individuals seek information that confirms their existing beliefs, can contribute to the spread and acceptance of false information [9]. Other cognitive biases, such as availability bias and the illusory truth effect, also play a role in the perpetuation of misinformation. Examining these biases can inform the development of effective detection and mitigation strategies.

Social media platforms serve as primary conduits for the dissemination of fake news and misinformation. Their algorithms, designed to optimize user engagement and content sharing, can unintentionally amplify false information. The lack of content moderation and the presence of echo chambers, where users are exposed to like-minded perspectives, further contribute to the spread of misinformation [10]. Understanding the dynamics of social media platforms is crucial for developing effective detection and intervention methods.

Overall, the background of fake news and misinformation highlights the importance of addressing these challenges to safeguard the integrity of information and promote informed decision-making. A comprehensive understanding of the nature and impact of fake news and misinformation is essential for the development of effective detection and mitigation strategies on social media platforms.

2.2 Manual Fact-Checking

Manual fact-checking involves human efforts to evaluate the veracity of news stories, claims, and social media content. It plays a crucial role in debunking fake news and misinformation. This subsection explores the process, methodologies, and prominent organizations involved in manual fact-checking. A manual fact-checking website is shown in Fig. 2.2.



Donald Trump's file

Republican from New York

Donald Trump was elected the 45th president of the United States on Nov. 8, 2016. He has been a real estate developer, entrepreneur and host of the NBC reality show, "The Apprentice." Trump's statements were awarded PolitiFact's 2015 Lie of the Year. Born and raised in New York City, Trump is married to Melania Trump, a former model from Slovenia. Trump has five children and eight grandchildren. Three of his children, Donald Jr., Ivanka, and Eric, serve as executive vice presidents of the Trump Organization.

The PolitiFact scorecard

True	25 (5%)
Mostly True	61 (11%)
Half True	83 (15%)
Mostly False	118 (22%)
False	173 (32%)
Pants on Fire	78 (14%)

Figure 2.2 Illustration of manual fact-Checking website [11].

Manual fact-checking typically follows a structured process to assess the accuracy of claims and news articles. Fact-checkers begin by selecting specific claims or stories to investigate based on their potential impact or widespread dissemination. They then conduct in-depth research, gathering relevant evidence, and consulting credible sources such as official records, experts, and data repositories. Fact-checkers scrutinize the content for accuracy, context, and reliability, aiming to determine whether the claim aligns with the available evidence. The findings are then presented in a transparent and accessible manner to inform the public about the veracity of the claim.

Fact-checking organizations employ various methodologies and techniques to assess the accuracy of claims. These may include:

- **Source verification**: Fact-checkers verify the credibility and reliability of the sources cited in news stories or claims. They cross-reference information with official records, scientific studies, expert opinions, and established news organizations to ensure accuracy.
- **Evidence-based analysis**: Fact-checkers employ evidence-based approaches, thoroughly examining the available evidence to support or refute a claim. This may involve analysing data, statistics, historical records, and eyewitness accounts to evaluate the validity of the information.
- **Expert consultation**: Fact-checkers consult subject matter experts to gain insights and verify claims in specialized fields. Experts provide valuable input based on their knowledge and experience, aiding in the accurate assessment of complex claims.
- **Contextual analysis**: Fact-checkers assess the context in which a claim is made to determine its accuracy. They consider factors such as the speaker's motivations, potential biases, timing, and the broader socio-political landscape to provide a comprehensive analysis.

Several prominent organizations and websites have emerged as key players in the field of manual fact-checking. These organizations employ teams of trained journalists, researchers, and subject matter experts to investigate claims and debunk fake news. Some well-known fact-checking organizations include:

- **Snopes.com**: Snopes.com is one of the oldest and most widely recognized fact-checking websites. It covers a broad range of topics, debunking urban legends, viral rumours, and misinformation across various domains.

- **Politifact.com**: Politifact.com focuses primarily on political claims made by politicians, pundits, and organizations. It rates claims on a truth scale, ranging from "True" to "Pants on Fire" based on their accuracy.
- **FactCheck.org**: FactCheck.org conducts independent fact-checking of claims made by political figures, media outlets, and advocacy groups. It provides comprehensive analysis and context to inform readers about the accuracy of the claims.
- **International Fact-Checking Network (IFCN)**: IFCN is a global network of fact-checkers committed to promoting accuracy in public discourse. It sets standards for fact-checking organizations, encourages collaboration, and provides resources and training to fact-checkers worldwide.

These organizations play a vital role in countering fake news and misinformation by providing reliable, evidence-based information to the public.

2.3 Machine Learning-Based Approaches

Machine learning (ML) techniques, including traditional algorithms, have been widely used for fake news detection. This subsection explores the application of traditional ML methods in fake news detection, focusing on various aspects of machine learning-based approaches.

The text-based analysis involves applying ML algorithms to analyse the textual content of news articles, social media posts and claims to identify patterns indicative of fake news [12]. Traditional ML algorithms, such as Support Vector Machines (SVM), Decision Trees, Random Forests, and Naive Bayes, have been widely used for text classification tasks in fake news detection. These algorithms leverage features extracted from the text, such as word frequencies, n-grams, or syntactic structures, to classify news articles or social media posts as fake or genuine.

Multimodal approaches leverage multiple modalities, such as text, images, videos, and metadata, to detect fake news [13]. Traditional ML algorithms can be applied to analyse features extracted from different modalities. For instance, features extracted from images or videos, such as visual cues, metadata, or image quality metrics, can be fed into traditional ML models like SVM or Decision Trees to detect manipulated or misleading content.

Machine learning models can leverage external evidence sources to enhance fact-checking and veracity assessment. Traditional ML algorithms can be applied to analyse the semantic connections between claims and entities in a knowledge graph, facilitating fact verification. Additionally, ML techniques can be used to develop models that utilize search engine results to retrieve relevant information and identify reputable sources to support or refute claims [14].

2.4 Deep Learning-Based Approaches

Deep learning models have emerged as powerful tools for improving the accuracy of misinformation detection [15]. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been extensively utilized to capture intricate patterns and dependencies in textual data. CNNs excel in analysing local patterns and identifying important textual features, while RNNs excel in modelling sequential dependencies and capturing long-range contextual information. These deep learning models can automatically learn and extract relevant features, enabling more accurate classification. For instance, Ma et al. proposed a deep learning framework that incorporated both text and visual features to identify fake news articles. By integrating information from multiple modalities, such as textual content and accompanying images, the model achieved a more comprehensive understanding of the information and its potential veracity. Zhang et al. introduced a graph-based convolutional network tailored for detecting rumours in propagation graphs, leveraging the network structure and information propagation patterns to enhance the discrimination between real and fake information.

2.4.1 Multimodal Content Classification

Text classification has been the most popular technique for social media analysis. However, with the increasing diversity of social media content, computer vision and image processing techniques have become valuable for analysis. Multimodal learning, which combines multiple modalities of information, has been applied to social media analysis, including fake news and rumour detection. Deep neural networks, capable of automatically learning deep representations for multimodal classification, have been commonly used to combine textual and visual information. For instance, Wang et al. [16] proposed an event-invariant adversarial neural network that learns multimodal domain-invariant features of a source post by using adversarial neural networks to remove event-specific features. Khattar et al. utilized a variational autoencoder, jointly learned with rumour classification, to extract a shared multimodal representation from textual and visual features. These approaches have shown improved performance in multimodal rumour classification. Recently, pretrained BERT has been adopted to replace RNNs for encoding text messages, significantly boosting the performance of multimodal rumour classification [17].

However, these methods often adopt simple fusion techniques, such as concatenation, to combine textual and visual representations, which may not fully exploit the intramodal relationship between the two modalities. To address this limitation, attention mechanisms have been applied to extract the deep correlation between text and image, resulting in more accurate rumour detection. For example, Jin et al. [18] proposed a multimodal recurrent neural network that combines visual and textual features using an attention mechanism for rumour detection. Zhou et al. [19] utilized image captioning with the LSTM network to explicitly learn the similarity and dissimilarity between the source text and image for fake news detection. Ying et al. [20] employed BERT with a cross-attention network to fuse visual and textual representations, forming a robust multimodal representation for misinformation detection. Chen et al. [21] proposed an ambiguity learning module that models the correlated and complementary relationship between textual and visual information. Furthermore, some researchers have considered the interactions between the source and replies from different people, which can enhance the accuracy and robustness of detection [22].

2.4.2 User Profile and Credibility for Rumour Detection

User profile-based approaches leverage information about users' behaviours, characteristics, and credibility indicators to identify fake news. By analysing user profiles, social connections, posting behaviour, and engagement patterns, these approaches aim to assess the trustworthiness and credibility of individuals in spreading or sharing false information. This subsection explores the use of user profile-based techniques in fake news detection.

Social network analysis techniques are employed to analyse the structure and dynamics of social connections among users. Measures such as centrality, community detection, and influence analysis can be used to identify users who are more likely to spread or engage with fake news. By examining the relationships between users and their roles within the network, these approaches provide insights into the potential influence and credibility of individuals in propagating misinformation.

2.4.3 Propagation Graph for Rumour Detection and Verification

Propagation-based approaches focus on analysing the propagation patterns and dynamics of information in social networks to detect and combat fake news. By studying how misinformation spreads through networks, these approaches aim to identify key nodes, influential users, and community responses that contribute to the dissemination and amplification of fake news. This subsection explores different aspects of propagation-based approaches in fake news detection.

To improve the robustness of rumour detection, the propagation graph of a conversation, including the replies to the source information, is considered. Zubiaga et al. [23] constructed the PHEME dataset and developed logistic regression with conditional random fields (CRF) for rumour detection, using linguistic features of the source and replies for classification. Ma et al. [24] constructed the first Chinese rumour dataset by collecting sources and replies of real and fake news from Weibo. They proposed to adopt tree-based recursive neural networks (RvNNs) to model the time-series linguistic features from the source and replies for rumour classification. Subsequently, different neural network architectures were explored for source-reply graph classification. These methods typically consider spatial and temporal features in the replies, where spatial features represent the semantic dependence between a message and its replies, while temporal features refer to the sequential relationship among all the replies in a time-series manner [25]. These two features have been proven effective for rumour detection on social media.

To model the spatial relationship between a message and its replies, CNNs and GNNs are commonly used to extract features from the replies. For instance, Yu et al. [26] proposed a CNN-based network that utilizes convolutional kernels to learn the spatial relationship among the replies by grouping relevant posts as a fixed-length representation. Bian et al. [27] employed a graph convolutional network to learn the propagation patterns of the source and replies and utilized convolutional kernels to learn the relationship between the replies. In temporal-based methods, RNNs and Transformers have been widely studied. Ma et al. [28] proposed an RNN to learn the long-term dependence among the replies to the source information by considering the replies as a variable-length time series of responses. The method was further improved by using Transformers [29] to enhance the temporal representation of the source-replies graph. Vu et al. [30] integrated spatial and temporal features by using GNN to extract spatial features in a propagation graph, followed by an RNN to aggregate the flattened node features generated by GNN. Song et al. [31] used the Temporal Graph Network [32] to incorporate temporal information into a graph attention network, generating a comprehensive representation graph for source-reply graph classification.

2.4.4 External Evidence-Based Automatic Fact-Checking

External evidence-based fact-checking approaches utilize external sources of information, such as knowledge graphs, search engines, and fact-checking databases, to verify the veracity of claims and detect fake news. By leveraging these external resources, these approaches aim to provide reliable and accurate assessments of information circulating on social media. This subsection explores different aspects of external evidence-based fact-checking in the context of fake news detection.

Knowledge graphs, such as Wikidata or DBpedia, provide structured and interconnected knowledge about various domains. These graphs contain factual information and relationships that can be leveraged for fact-checking purposes [33]. External evidence-based fact-checking approaches utilize knowledge graphs to validate claims against reliable sources and detect inconsistencies or contradictions. By cross-referencing information with the knowledge graph, these approaches enhance the accuracy of fact-checking and identify potentially fake news.

Search engine-based fact-checking approaches utilize search engines, such as Google or Bing, to retrieve relevant information and assess the veracity of claims [34]. These approaches involve querying search engines with specific keywords or claim-related terms to retrieve relevant documents, news articles, or authoritative sources. By analysing the search results and cross-referencing information, these approaches provide additional evidence to support or debunk claims and detect fake news.

External evidence-based fact-checking involves the task of Recognizing Textual Entailment (RTE), which aims to determine whether the evidence supports or refutes a given claim [35]. Various retrieval strategies have been employed in this context, such as utilizing commercial search APIs, performing entity linking with Named Entity Recognition (NER), or utilizing Lucene indices. Factual verification datasets are typically generated through artificial inputs, where annotators create claims based on online references, including authoritative news sources or Wikipedia. For example, Ferreira and Vlachos [36] proposed a dataset comprising 2,595 claims along with associated news articles, which were collected and labelled by journalists with an estimation of their veracity. Hanselowski et al. [37] focused on fact-checking natural language claims using articles from rumour-debunking sites. Jiang et al. [38] focused on fact-checking natural language claims using articles from rumour-debunking sites. Subsequently, Dougrez-Lewis et al. [39] explored rumour verification of social media claims using external evidence obtained from search engines. However, they did not consider the community response within propagation graphs, which has been proven to be an effective context for rumour verification. Consequently, this motivates us to simultaneously consider the community response and external evidence within a unified network architecture.

2.5 Summary

In summary, this comprehensive literature review provides an extensive overview of the research conducted on the detection and mitigation of fake news and rumours on social media platforms. The review encompasses manual fact-checking, machine learning-based approaches, deep learning models, multimodal source classification, propagation graph classification, the integration of external evidence, and comparative studies. By examining the strengths, weaknesses, and advancements of these approaches, we aim to identify research gaps and lay the groundwork for our proposed methods.

Chapter 3

Multimodal Rumour Detection

3.1 Motivation

In this chapter, we focus on multimodal rumour detection, which aims to classify a pair of image and text as rumour or non-rumour. In crossmodal fusion learning, an attention mechanism, which is called positive correlation, is used to aggregate feature vectors from two different modalities. The attention score is computed by matching the features between the two modalities. This relies on the assumption that the two modalities always match or are positively correlated. In practice, this assumption does not hold, due to the diverse characteristics of user-generated content on social media. However, most existing methods only consider correlated features and ignore inconsistent semantic meanings between modalities. Based on this motivation, we consider both positive and negative correlations between modalities during fusion learning, to explicitly model the consistent and contrary information between text and image for representation.

The contributions of our work are as follows:

- We propose a novel crossmodal bipolar attention mechanism to model the direct and inverse relationship between each textual feature and each visual feature and fuse them to form two sequences of features. We incorporate the bipolar attention mechanism with the dot-product and additive attention mechanisms.

- We employ the attentive pooling module to transform the fused features into the most informative features.
- We utilize the pretrained vision transformer (ViT) directly as an image feature extractor and incorporate the hidden representations of the image patches from ViT into the proposed bipolar attention mechanism for multimodal classification.

3.2 Crossmodal Bipolar Attention Network

Given an input image-text pair (I, T), our proposed model aims to classify into one of the two classes $\boldsymbol{y} \in \mathbb{R}^2$, where y_0 and y_1 represents non-rumour and rumour, respectively. The proposed crossmodal bipolar attention network (CBAN) contains four modules, as shown in Fig. 3.1. The first step in our framework is to extract features from the image and text separately, and represent them as two separate vector sequences. Then, our proposed crossmodal bipolar attention module is used to fuse the information between the modalities and form two feature sequences, where each feature sequence represents a feature in a modality and its correlation to the other modality. Positive attention aims to identify the most similar features between the modalities, while negative attention is to find out the contrary information. These two pieces of contrary information are important for representing multimodal data. After that, a unimodal attention module is applied to transform the two feature sequences into two fixed-length vectors by attentive pooling. Finally, the two vectors are concatenated and used for multimodal classification.

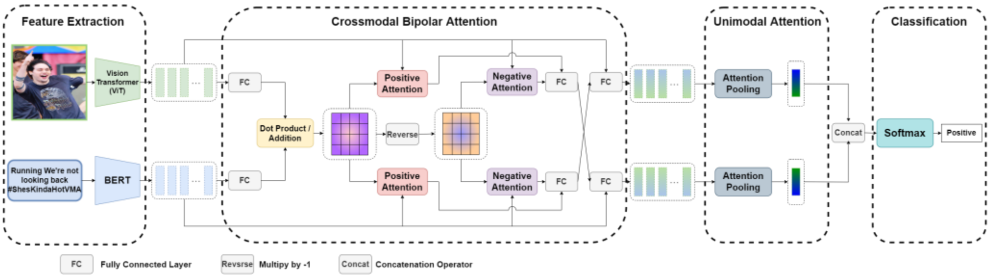


Figure 3.1 The proposed Crossmodal Bipolar Attention Network (CBAN).

3.2.1 Feature Extraction Modules

Our feature extraction module contains two branches, the visual branch and the textual branch. The visual branch is to extract features from an input image using a pretrained vision transformer (ViT) [40]. Specifically, an input image is divided into m patches of size 16 × 16, which are then fed to the vision transformer. We take the output of the ViT encoder layer as the extracted features of the input image. Mathematically, given an input image I, ViT extracts a sequence of m visual features $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \dots, \boldsymbol{v}_m]^{\mathrm{T}} \in \mathbb{R}^{m \times d_v}$, where d_v is the embedding dimension of ViT.

Pretrained language models have been widely used in many language processing tasks, such as text classification and machine translation. In the textual branch of our proposed model, we utilize the pretrained bidirectional transformer model, i.e., BERT [41], to represent the input sentence as a sequence of n real-valued vectors $= [\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3, \dots, \boldsymbol{t}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times d_t}$, where d_t is the embedding dimension of BERT and n is the number of words in the input sentence.

3.2.2 Crossmodal Bipolar Attention

Based on those previous attempts, which make use of a fine-grained attention matrix to compute the co-attention weights through the similarity between two modalities, our proposed Crossmodal bipolar attention module computes the positive and negative attention vectors between each feature of one modality and all features of the other modality. Positive attention aims to find the most similar features between the modalities, while negative attention is to compute the dissimilar or contrary information. We first non-linearly project the visual feature $\boldsymbol{V}$ and textual feature $\boldsymbol{T}$ into the same embedding space, $\boldsymbol{V}_{emb} \in \mathbb{R}^{m \times d_e}$ and $\boldsymbol{T}_{emb} \in \mathbb{R}^{n \times d_e}$, respectively, as follows:

$$\mathbf{T_{emb}} = \tanh(\mathbf{T}\mathbf{W}_{\mathrm{te}} + \mathbf{b}_{\mathrm{te}}), \tag{3.1}$$

$$\mathbf{V_{emb}} = \tanh\,(\mathbf{V}\mathbf{W}_{\mathrm{ve}} + \mathbf{b}_{\mathrm{ve}}), \tag{3.2}$$

where $\mathbf{W}_{\mathrm{te}} \in \mathbb{R}^{d_t \times d_e}$, $\mathbf{W}_{\mathrm{ve}} \in \mathbb{R}^{d_v \times d_e}$, $\mathbf{b}_{\mathrm{te}} \in \mathbb{R}^{1 \times d_e}$ and $\mathbf{b}_{\mathrm{ve}} \in \mathbb{R}^{1 \times d_e}$ are trainable parameters. Then, we compute the fine-grained similarity matrix $\mathbf{S} \in \mathbb{R}^{m \times n}$, which represents the similarity between every textual and visual feature. The similarity scores can be computed with the scaled dot-product attention mechanism, as follows:

$$\mathbf{S} = \frac{\mathbf{V_{emb}}\mathbf{T_{emb}^{\mathrm{T}}}}{\sqrt{d_e}}, \tag{3.3}$$

where $\boldsymbol{W}_s \in \mathbb{R}^{d_e \times 1}$ is a trainable weight vector. In Equations (3.1) and (3.2), the row i in $\boldsymbol{S}$ represents the similarity between the i-th visual feature and all textual features, while each column j represents the similarity between the j-th textual feature and all visual features.

Positive Attention Mechanism. Similar to the previous works, the most similar features between the modalities can be computed by the crossmodal attention vectors. The visually guided textual features $\boldsymbol{T^p} \in \mathbb{R}^{m \times d_t}$ and the textually guided visual features $\boldsymbol{V^p} \in \mathbb{R}^{n \times d_v}$ are computed, as follows:

$$\boldsymbol{T^p} = \mathrm{softmax}(\boldsymbol{S})\boldsymbol{T}, \tag{3.4}$$

$$\boldsymbol{V^p} = \mathrm{softmax}(\boldsymbol{S}^{\mathrm{T}})\boldsymbol{V}. \tag{3.5}$$

Negative Attention Mechanism. In the previous works, only the matched information between the modalities is highlighted by using positive attention. However, the most contrary information is ignored, which is also useful for representation. In addition to the positive-correlated attention vectors, we further compute the negatively correlated attention vectors, i.e., $\boldsymbol{T^n} \in \mathbb{R}^{m \times d_t}$ and $\boldsymbol{V^n} \in \mathbb{R}^{n \times d_v}$, by multiplying the similarity matrix $\boldsymbol{S}$ with a negative constant before applying Softmax, as follows:

$$\boldsymbol{T^n} = \mathrm{softmax}(-\boldsymbol{S})\boldsymbol{T}, \tag{3.6}$$

$$\boldsymbol{V^n} = \mathrm{softmax}(-\boldsymbol{S}^{\mathrm{T}})\boldsymbol{V}. \tag{3.7}$$

Similar to multi-head attention, we use a fully connected layer to aggregate the positive and negative attention vectors, as follows:

$$\boldsymbol{T}^* = \tanh\big((\boldsymbol{T}^p \oplus \boldsymbol{T}^n)\boldsymbol{W}_{tt} + \boldsymbol{b}_{tt}\big), \tag{3.8}$$

$$\boldsymbol{V}^* = \tanh\big((\boldsymbol{V}^p \oplus \boldsymbol{V}^n)\boldsymbol{W}_{vv} + \boldsymbol{b}_{vv}\big), \tag{3.9}$$

where $\boldsymbol{W}_{tt} \in \mathbb{R}^{2d_t \times d_t}$, $\boldsymbol{W}_{vv} \in \mathbb{R}^{2d_v \times d_v}$, $\boldsymbol{b}_{\mathrm{tt}} \in \mathbb{R}^{1 \times d_t}$ and $\boldsymbol{b}_{\mathrm{vv}} \in \mathbb{R}^{1 \times d_v}$ are trainable parameters. Having obtained the visually guided textual features $T^* \in \mathbb{R}^{m \times d_t}$ and the textually guided visual features $V^* \in \mathbb{R}^{n \times d_v}$, we use a fully connected layer to model the relationship between the modality features and the guided features to obtain the visual-fusion feature $\boldsymbol{V}_f = [\boldsymbol{v}_{f,1}, \boldsymbol{v}_{f,2}, \boldsymbol{v}_{f,3}, \dots, \boldsymbol{v}_{f,m}]^{\mathrm{T}} \in \mathbb{R}^{m \times d_v}$ and the textual-fusion feature $\boldsymbol{T}_f = [\boldsymbol{t}_{f,1}, \boldsymbol{t}_{f,2}, \boldsymbol{t}_{f,3}, \dots, \boldsymbol{t}_{f,n}]^{\mathrm{T}} \in \mathbb{R}^{n \times d_t}$, as follows:

$$\boldsymbol{V}_f = \tanh\big((\boldsymbol{V} \oplus \boldsymbol{T}^*)\boldsymbol{W}_v + \boldsymbol{b}_v\big), \tag{3.10}$$

$$\boldsymbol{T}_f = \tanh\big((\boldsymbol{T} \oplus \boldsymbol{V}^*)\boldsymbol{W}_t + \boldsymbol{b}_t\big), \tag{3.11}$$

where $\boldsymbol{W}_v \in \mathbb{R}^{(d_t + d_v) \times d_v}$, $\boldsymbol{W}_t \in \mathbb{R}^{(d_t + d_v) \times d_t}$, $\boldsymbol{b}_v \in \mathbb{R}^{1 \times d_v}$ and $\boldsymbol{b}_t \in \mathbb{R}^{1 \times d_t}$ are trainable parameters, and $\oplus$ represents the concatenation operator. These can generate more comprehensive features for multimodal classification.

3.2.3 Unimodal Attention Pooling

To obtain the most informative representation of the textual-fusion features and visual-fusion features, we employ an attentive pooling module to transform the sequences of fused features, $\boldsymbol{V}_f$ and $\boldsymbol{T}_f$. Similar to Equations (4) and (5), the attention vectors of values $\boldsymbol{\mathcal{V}}$ are computed. In contrast, the query $\boldsymbol{Q}$ is now learned during training. We compute the attention scores with the scaled dot-product attention mechanisms. Taking the visual-fusion features $\boldsymbol{V}_f$ as an example, we compute the attention score α_i^v for each feature $\boldsymbol{v}_{f,i}$ as follows:

$$\alpha_i^v = \frac{\boldsymbol{v}_{f,i}\boldsymbol{U}_d}{\sqrt{d_v}}, \tag{3.12}$$

where α_i^v represents the importance of each visual feature in the visual-fusion features, $\boldsymbol{U}_d \in \mathbb{R}^{d_v \times 1}$, $\boldsymbol{U}_a \in \mathbb{R}^{1 \times d_v}$ and $\boldsymbol{W}_a \in \mathbb{R}^{d_v \times 1}$ are trainable parameters. Then, we normalize the attention weights α_i^v by the Softmax function, as follows:

$$\tilde{\alpha}_i^v = \frac{\exp(\alpha_i^v)}{\sum_{i=1}^{m} \exp(\alpha_i^v)}. \tag{3.13}$$

Then, the final visual-fusion feature, $\boldsymbol{v}'_f \in \mathbb{R}^{1 \times d_v}$**,** can be calculated as follows:

$$\boldsymbol{v}'_{\boldsymbol{f}} = \sum_{i=1}^{m} \tilde{\alpha}_i^v \, \boldsymbol{v}_{f,i}. \tag{3.14}$$

Repeating Equations (3.12) to (3.14), with the visual-fusion feature $\boldsymbol{V}_f$ replaced by the textual-fusion feature $\boldsymbol{T}_f$, we can obtain the final fusion of the textual feature $\boldsymbol{t}'_f \in \mathbb{R}^{1 \times d_t}$.

3.2.4 Classification Layer

We simply concatenate the two final fused features, i.e., $\boldsymbol{v}_f^{\acute{}}$ and $\boldsymbol{t}_f^{\acute{}}$, to perform classification. The concatenated feature vector is classified by using a fully connected layer, as follows:

$$\hat{\boldsymbol{y}} = \tanh\left(\left(\boldsymbol{v}'_f \oplus \boldsymbol{t}'_f\right)\boldsymbol{W_c} + \boldsymbol{b_c}\right), \tag{3.15}$$

where $\boldsymbol{W}_c \in \mathbb{R}^{(d_t + d_v) \times c}$ and $\boldsymbol{b}_c \in \mathbb{R}^{1 \times c}$ are trainable parameters.

3.2.5 Loss Function

We employ the cross-entropy loss as the objective function in our proposed method. Given the predicted label $\hat{\boldsymbol{y}}$ and the ground-truth label $\boldsymbol{y}$, we minimize the negative log-likelihood with c classes after the Softmax function. Therefore, we have

$$Loss = -\sum_{i}^{c} \boldsymbol{y} \log\left(\mathrm{Softmax}(\hat{\boldsymbol{y}})\right). \tag{3.16}$$

3.3 Experiment

3.3.1 Datasets

Our experiments were conducted on two publicly available image-text classification data sets, widely used for multimodal rumour detection. PHEME [23] dataset contains 1972 rumours and 3830 non-rumour English conversations on Twitter, across five events, including Charlie Hebdo, Ferguson, Germanwings Crash, Ottawa Shooting, and Sydney Siege. The Weibo [18] dataset contains 6226 rumours and 9405 non-rumours Chinese conversations on Weibo. To ensure a fair comparison to previous work, we divide the two datasets in a ratio of 8:2, for training and testing, respectively.

3.3.2 Experiment Setup

Preprocessing: In all experiments, we process all input tweets by anonymizing the user mentions, as well as removing line breaks and website links. For the input images, we resize the short side of the images to 224 pixels, and crop the centre region of the images, so that the size of each cropped image is 224×224 pixels, which is the input size of the pretrained vision transformers.

Evaluation Metrics: To evaluate the performance of different classification models, we use the average F1 score and accuracy as performance metrics, which are the same in most of the literature.

Hyperparameters and pretrained models: For ViT and BERT, we use the pretrained models and the dimensions of both d_v and d_t are 768. In the proposed Crossmodal attention module, we set the embedding dimension to d_e =768. For all experiments, the models were trained with a mini-batch size of 64 for 10 epochs. We use the Adam Optimizer with a fixed learning rate of 0.00002. To avoid overfitting, we use dropout with a rate of 0.3. The best model obtained in the validation set is selected for testing. Our models are tuned on the validation sets, and we report the results on the test sets.

3.3.3 Comparison with State-of-the-Art Methods

Tables 3.1 and 3.2 show the overall results compared to the following state-of-the-art methods.

(1) *SVM-TS* [42] applies a linear classifier based on SVM along with heuristic rules to classify the claim.

(2) *GRU* [28] is a type of Recurrent Neural Network (RNN) used to use to model the sequential information among messages from a conversation for rumour detection.

(3) *CNN* [43] uses a convolutional neural network with fixed-length windows on posts to capture the features.

(4) *TextGCN* [44] models the whole corpus as a heterogeneous graph and feeds it into the GCN to obtain the textual semantic features.

(5) *Att-RNN* [18] produces embeddings of text and relevant social context via the LSTM module and then integrates the joint features with image features by neural attention. To a fair comparison, we eliminate the part addressing social context information.

(6) *EANN* [16] learns event-invariant multimodal features of each post for fake news detection by employing an adversarial network to eliminate event-specific components from the post features based on the concatenation of extracted textual and visual features.

(7) *MVAE* [45] employs a variational autoencoder with an encoder and decoder for each modality to obtain a shared multimodal representation between text and image, which is trained jointly with the subsequent classifier for fake news detection.

(8) *SAFE* [46] adopts neural networks to gain the latent representations of both texts and images and then takes the relationship (similarity) between modalities as feature combined with the concatenation of feature and visual feature to conduct fake news detection.

(9) *MMCN* [47] is a Multi-level Multimodal Cross-attention Network (MMCN) that exploits the multi-level semantics of textual content and jointly integrates the relationships of duplicate and different modalities of social multimedia posts.

Table 3.1 Results of multimodal rumour detection on PHEME dataset.

Methods	Accuracy	Precision	Recall	F1
SVM [42]	0.639	0.638	0.641	0.639
GRU [28]	0.832	0.819	0.804	0.805
CNN [43]	0.779	0.766	0.741	0.749
TextGCN [44]	0.828	0.801	0.782	0.783
Att-RNN [18]	0.850	0.834	0.824	0.829
EANN [16]	0.681	0.693	0.707	0.721
MVAE [45]	0.852	0.839	0.818	0.827
SAFE [46]	0.811	0.817	0.750	0.767
MMCN [47]	0.872	0.863	0.850	0.856
CBAN (Ours)	**0.894**	**0.868**	**0.878**	**0.894**

Table 3.2 Results of multimodal rumour detection on Weibo dataset.

Methods	Accuracy	Precision	Recall	F1
SVM [42]	0.640	0.696	0.686	0.679
GRU [28]	0.720	0.709	0.702	0.699
CNN [43]	0.740	0.742	0.740	0.740
TextGCN [44]	0.787	0.844	0.863	0.777
Att-RNN [18]	0.772	0.787	0.838	0.769
EANN [16]	0.782	0.790	0.818	0.780
MVAE [45]	0.824	0.828	0.829	0.823
SAFE [46]	0.763	0.775	0.846	0.761
MMCN [47]	0.879	0.880	0.880	0.880
CBAN (Ours)	**0.934**	**0.935**	**0.934**	**0.934**

The results on the PHEME dataset demonstrate the effectiveness of the various multimodal rumour detection methods. Notably, CBAN (the proposed method) achieves the highest accuracy of 0.890, surpassing all other methods. It also shows superior precision, recall, and F1 score in classifying both rumours and non-rumours. This indicates the efficacy of CBAN in effectively leveraging multimodal features for accurate rumour detection on the PHEME dataset.

Other methods, such as MMCN, MVAE, and TextGCN, also demonstrate strong performance, achieving high accuracy and F1 scores. They exhibit a balance between precision and recall, indicating their capability to accurately detect rumours while minimizing false positives and false negatives. These results highlight the effectiveness of leveraging multimodal information for rumour detection on the PHEME dataset.

The results on the Weibo dataset provide insights into the performance of the multimodal rumour detection methods in a different cultural and linguistic contexts. CBAN continues to exhibit superior performance, achieving an impressive accuracy of 0.922, along with high precision, recall, and F1 score. This indicates its robustness and adaptability across different datasets and contexts.

Similarly, MMCN, GRU, and TextGCN perform well on the Weibo dataset, showcasing their effectiveness in detecting rumours in a Chinese social media environment. These methods achieve high accuracy and F1 scores, emphasizing their potential for cross-cultural rumour detection.

Comparing the results from the two tables, it is evident that CBAN consistently performs well on both datasets, demonstrating its robustness and effectiveness in multimodal rumour detection tasks. MMCN also exhibits strong performance on both datasets, further emphasizing its reliability across different domains.

The variations in results between the two datasets can be attributed to the differences in the characteristics of the PHEME and Weibo datasets. The Weibo dataset presents unique challenges such as language differences and cultural nuances, which impact the performance of the methods. Despite these challenges, the multimodal rumour detection methods show promising results, highlighting their potential for addressing rumours in diverse contexts.

3.3.4 Ablation Study

The table presents the performance metrics of different models on two datasets: PHEME and Weibo. The models evaluated include CBAN with and without crossmodal attention, as well as CBAN with and without unimodal attention.

Table 3.3 Ablation study on the CBAN using PHEME and Weibo datasets.

Dataset	Model	Accuracy	Precision	Recall	F1
PHEME	CBAN w/o crossmodal attention	0.885	0.854	0.883	0.866
	CBAN w/o unimodal attention	0.883	0.854	0.868	0.860
	CBAN (Full)	**0.894**	**0.868**	**0.878**	**0.894**
Weibo	CBAN w/o crossmodal attention	0.881	0.882	0.881	0.881
	CBAN w/o unimodal attention	0.918	0.921	0.918	0.918
	CBAN (Full)	**0.934**	**0.935**	**0.934**	**0.934**

On the PHEME dataset, CBAN with crossmodal attention and unimodal attention achieves the highest accuracy of 0.894, followed by CBAN without unimodal attention with an accuracy of 0.883. CBAN without crossmodal attention performs slightly lower with an accuracy of 0.885. This suggests that incorporating crossmodal attention improves the overall performance of CBAN on the PHEME dataset.

In terms of precision, CBAN (Full) obtains the highest value of 0.868, indicating its ability to accurately classify rumour instances. CBAN without unimodal attention follows closely with a precision of 0.854. CBAN without crossmodal attention exhibits a similar precision score of 0.854.

Regarding recall, CBAN (Full) achieves a value of 0.878, demonstrating its capability to effectively capture true positive instances. CBAN without crossmodal attention shows the highest recall score of 0.883, indicating its strength in correctly identifying rumour instances. CBAN without unimodal attention achieves a recall score of 0.868.

The F1 score, which balances precision and recall, reflects the overall performance of the models. CBAN (Full) achieves the highest F1 score of 0.894, indicating its effectiveness in achieving a balance between precision and recall. CBAN without unimodal attention and CBAN without crossmodal attention exhibit F1 scores of 0.918 and 0.866, respectively.

Moving to the Weibo dataset, CBAN performs remarkably well across all metrics. It achieves the highest accuracy of 0.934, precision of 0.935, recall of 0.934, and F1 score of 0.934. CBAN without unimodal attention also demonstrates strong performance, achieving an accuracy of 0.918 and F1 score of 0.918. CBAN without crossmodal attention achieves an accuracy of 0.881, indicating a slightly lower performance compared to the other two models.

These results suggest that CBAN, particularly when incorporating both crossmodal and unimodal attention, exhibits superior performance on both the PHEME and Weibo datasets. The inclusion of attention mechanisms enhances the model's ability to effectively capture and utilize information from different modalities, leading to improved accuracy and balanced performance in rumour detection.

3.4 Applications of CBAN to other Multimodal Classification Tasks

We apply the proposed CBAN other multimodal classification tasks on image-text pairs generated on social media used for sentiment analysis, sarcasm detection, crisis categorization, and hate-speech detection. These data sets consist of real data generated by users on Twitter.

3.4.1 Datasets

Sentiment Analysis: MVSA-Single and MVSA-Multiple are two sentiment analysis data sets. Both data sets have the same three possible labels, which are positive, negative, and neutral. MVSA-Single contains 5,129 image-text pairs, while MVSA-Multiple contains 19,600 image-text pairs. The former is labeled by a single annotator, while the latter is labeled by three annotators. For a fair comparison, we process the two MVSA data sets as described in [48], to filter inconsistent labels and combine multiple labels into a single label. Same as [48], we divide the two data sets in a ratio of 80:10:10 to form the training set, validation set and test set.

Sarcasm Detection: We perform multimodal image-text sarcasm detection on the Twitter data set proposed in 2019 [49]. This data set is used for a binary classification task to detect whether an image-text pair is sarcasm or not. The data set has been divided into three parts, including 19,816 for training, 2,410 for validation, and 2,409 for testing

Crisis Categorization: CrisisMMDv2 [50] is a data set for categorizing crisis events on Twitter data. It is an updated version of CrisisMMDv1 constructed by the same authors. CrisisMMDv2 was manually checked, with the duplicated entries in CrisisMMDv1 being removed. Therefore, we conducted experiments on CrisisMMDv2 only, which has two sub-tasks, namely informativeness and humanitarian. The informativeness task is a binary classification task to classify whether the image-text pairs are informative about the concerns of humanitarian organizations. The humanitarian task contains five categories, which are affected individuals, infrastructure or utility damage, rescue volunteering or donation effort, other relevant information, and not-humanitarian. The informativeness task contains 16,058 image-text pairs, while the humanitarian task contains 18,082 image-text pairs. The data set has already been divided into a ratio of 70:15:15, for training, validation, and testing.

Hate-Speech Detection: MMHS150K [51] is the first and largest multimodal hate-speech detection data set. Each image-text pair was labeled by three annotators. A majority vote was employed to filter inconsistent entries. The data set contains 112,845 not-hate tweets and 36,978 hate tweets. 5,000 and 1,000 of them are separated into a validation set and a test set, respectively, and the rest is used for training.

3.4.2 Comparison of CBAN to Other State-of-the-Art Methods

Since a variety of models have been proposed for the different data sets, it is difficult to compare different methods on the different tasks. Thus, we first compare our models with those published results task by task. The detailed evaluation of our models was conducted in an ablation study.

Sentiment Analysis: We compare our proposed CBAN with the following state-of-the-art methods for sentiment analysis. The qualitative results are shown in Table 3.4.

- *MultiSentiNet* [52]: This is a textual LSTM network with attention guided by scene and object CNN features. The final feature representation is an aggregation of the guided textual features and visual CNN features.
- *CoMN* [48]: This is an improved version of MultiSentiNet, which uses a co-memory network to simultaneously model visually guided textual features and textually guided visual features.

- *MVAN* [53]: This is a multi-view attention network with a co-memory mechanism to model the feature attention between textual and visual features.
- *FENet* [54]: This is a fusion-based feature extraction network, which adopts co-attention mechanism, to fuse visual features and textual features with a fine-grained similarity matrix.

Table 3.4 Comparison of sentiment analysis on the MVSA-Single and MVSA-Multiple datasets.

Method	MVSA-Single		MVSA-Multiple	
	F1 Score	Accuracy	F1 Score	Accuracy
MultiSentiNet [52]	0.6963	0.6984	0.6811	0.6886
CoMN [48]	0.7001	0.7051	0.6883	0.6892
MVAN [53]	0.7298	0.7298	0.7230	0.7236
FENet [54]	0.7406	0.7421	0.7121	0.7146
CBAN (Ours)	**0.7777**	**0.7751**	**0.7499**	**0.7558**

Compared with MultiSentiNet, CoMN, and MVAN, our proposed CBAN improves both F1 score and accuracy by 3-7%. This is mainly because the co-memory network is not sufficient to model the semantic relationship between visual and textual features, as their attention weights for a modality are calculated based on the aggregation of the features from the other modality. Moreover, our proposed method achieves both F1-score and accuracy 3% higher than FENet. Although FENet uses a fine-grained similarity matrix between the two modal features, it is not sufficient to use a single similarity matrix only. Different from FENet, our model further considers the inverse relationship between the visually guided textual features and the textually guided visual features. Thus, our proposed CBAN can obtain robust semantic relationships across modalities.

Sarcasm Detection: We compare our proposed CBAN with the following baseline models for sarcasm detection. The qualitative results are shown in Table 3.5.

- *MMHFM* [49]: This is a hierarchical fusion model, which fuses image features, attribute features and text features with early fusion and representation fusion.
- *LXMBERT* [55]: This is a crossmodal transformer network, based on pretrained image-text Q&A and matching tasks, to learn the semantic relationships across modalities. The model is also fine-tuned on multimodal detection data sets.
- *ViLBERT* [56]: This is a crossmodal transformer network, pretrained with a large amount of unlabelled image-text pairs to learn the semantic relationship across modalities. The model can be used as an encoder for multimodal classification.
- *2D-Intra-Attention + RoBERTa* [57]: This model utilizes a 2D-Intra-Attention module, based on the co-attention mechanism, to guide the BERT model with visual features.

Table 3.5 Results of sarcasm detection.

Method	F1 Score	Accuracy
MMHFM [49]	0.8018	0.8344
LXMBERT [55]	0.8014	0.8393
ViLBERT [56]	0.8171	0.8468
2D-Intra-Attention + RoBERTa [57]	0.8605	0.8851
CBAN (Ours)	**0.9264**	**0.9261**

Compared to MMHFM, our model achieves an improvement of 12% and 9%, in terms of F1 score and accuracy, respectively. Although attributes are used in MMHFM to align visual and textual features, this is insufficient because sarcasm is not determined solely by attributes. In addition, our proposed method is about 7-10% higher, in terms of F1 score and accuracy, than both LXMBERT and ViLBERT. These two methods use a pretrained transformer network to model the relationship between text and images. It is worth noting that this representation learning has limited performance, because text and images are sometimes not highly correlated. In our proposed crossmodal attention module, we have a fully connected layer to connect the unimodal features, so that the unattended features also contribute to the final sarcasm detection.

Our proposed model outperforms the state-of-the-art RoBERTa network with 2D-intra-attention, by 6% in terms of F1 score and 4% in terms of accuracy for sarcasm detection. Although this method uses a transformer to model the relationship between the modalities, its proposed 2D-intra-attention mechanism can learn a certain amount of semantic relationships between the modalities.

Crisis Categorization: We compare our proposed CBAN, which consists of two subtasks, i.e., informativeness and humanitarian, to the following baseline models for crisis categorization. The qualitative results are shown in Table 3.6.

- *VGG + CNN* [50]: VGGNet is used to extract image features, while CNN is adopted to extract text features. Then, early fusion is employed to fuse the two features.
- *Relation-attention* [58]: A relation network is used to compute the self-attention of textual and visual data. Then, factorized bilinear pooling is applied to fuse these two features.
- *Transformer-attention* [58]: A transformer network is used to compute the self-attention of textual and visual data. Then, factorized bilinear pooling is applied to fuse these two features.
- *CrisisFlow* [59]: An early fusion approach is adopted to combine textual and visual features for informativeness classification.

Compared to VGG+CNN, a self-attention network is much more powerful than VGG and CNN in extracting visual and textual features. It simply concatenates the two modalities, and does not use the semantic relationship between the modalities. Our proposed model, in terms of F1 score and accuracy, achieves an improvement of 8.3% and 8.2%, respectively, in the informativeness task, and an improvement of 2.9% and 2.6%, respectively, in the humanitarian task.

Table 3.6 Results of crisis classification.

Method	Informativeness		Humanitarian	
	F1 Score	Accuracy	F1 Score	Accuracy
VGG + CNN [50]	0.8420	0.8400	0.7830	0.7840
Relation-attention [58]	-	-	0.8110	-
Transformer-attention [58]	-	-	0.8550	-
CrisisFlow [59]	-	0.9100	-	-
CBAN (Ours)	0.9267	0.9263	0.8840	0.8838

In the humanitarian classification task, Raj et al. used the transformer and relation-attention modules, with factorized bilinear pooling to fuse the two modality features. Although the method can have a 1% F1-score improvement compared to VGG+CNN, our proposed method still outperforms it by 2.9%, in terms of F1 score. This is likely due to the fact that the factorized bilinear pooling is not sufficient to model the semantic relationship across modalities, compared to the attention mechanism.

In the informativeness task, our method is 1.5%, in terms of accuracy and F1 score, better than CrisisFlow. This is likely because the early fusion approach is insufficient to represent the semantic relationship between the two modalities, compared to the crossmodal attention mechanism.

Hate-Speech Detection: We compare the results of our proposed CBAN with the following models for hate-speech detection. The results are shown in Table 3.7.

- *Davidson* [60]: Davidson et al. adopted bigram, unigram, and trigram as textual features, followed by logistic regression.
- *FCM* [51]: This is a feature concatenation model, which concatenates textual features to the average of visual features.
- *SCM* [51]: This is a spatial concatenation model, which concatenates textual features to each visual feature, followed by averaging pooling.

Table 3.7 Results of hate-speech detection.

Method	F1 Score	Accuracy
Davidson [60]	0.7030	0.6840
FCM [51]	0.7040	0.6840
SCM [51]	0.7020	0.6850
CBAN (Ours)	**0.7085**	**0.7143**

While the LSTM and CNN-based method proposed by Gomez et al. can improve the F1 score by 0.01%, compared to the traditional machine learning algorithm in hate-speech detection, our proposed method outperforms it by 0.8% in F1 score and 3% in accuracy. This is because the self-attention network is more robust than RNNs and CNNs in extracting textual and visual features.

In Fig. 3.2(a), the image-text pair is misclassified as negative sentiment when only the positive attention module is used. It is likely that the positive module can only pay attention to “He won't make it easy” in the text. Interestingly, if only the negative attention module is used, the image-text pair is wrongly detected as positive sentiment. It is likely that the model can locate the enthusiastic baseball player in the image. If the bipolar attention module is employed, the module can consider both the correlated and inconsistent information. Thus, it successfully classifies the image-text pair as a neutral sentiment. In Fig. 3.2(b), the positive attention-only network can successfully classify the image-text pair as a positive sentiment. This is because both the visual and textual information are about the heart ring, and they have a strong positive correlation. If only the negative-only module is added, it is misclassified as neutral. This means that the inconsistent information between the modalities is not enough for representation. When both the positive and negative attention modules are added, the model can classify it as positive. Figs. 3.2(c) and 3.2(d) show the successful classification results when the negative attention module or bipolar attention module is added. These two examples show that the image and text contain inconsistent information. For example, Fig. 3.2(c) shows that the text is about Lehmann, who is a cricketer, to convince bewildered Australia. However, the associated image shows him in a yard.

This means that the image and text are not correlated. Moreover, the text in Fig. 3.2(d) contains curious and questionable context, but the image is a group photo of singers at a concert. These two examples show that the negative attention module is particularly important when the image and text do not contain the same or a similar semantic meaning, which is a crucial feature for multimodal classification.

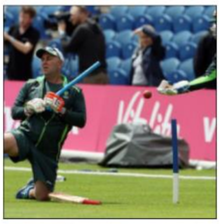
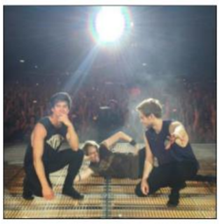
	Text	Image	Ground Truth	Positive Attention-only	Negative Attention-only	Bipolar Attention
(a)	#OldManHawkins sees the #Angels whippersnappers trying to chip away. He won't make it easy.		Neutral	Negative	Positive	Neutral
(b)	We heart this ring from @peterstorminc!!		Positive	Positive	Neutral	Positive
(c)	Lehmann must convince bewildered Australia it really was just a hiccup		Negative	Positive	Negative	Negative
(d)	WTH is this?!		Positive	Neutral	Positive	Positive

Figure 3.2 Examples of sentiment analysis of image-text pairs with the proposed module.

3.5 Summary

In this chapter, we propose a Crossmodal Bipolar Attention Network (CBAN), which effectively utilizes attention mechanisms for multimodal classification.

CBAN makes use of two self-attention networks, namely the vision transformer (ViT) and the bidirectional transformer (BERT), to extract features from images and text. To learn the semantic relationship across modalities, a novel crossmodal bipolar attention module, with both positive and negative attention mechanisms, is proposed. We inject this module with scaled dot-product attention mechanism. After that, attentive pooling is adopted to represent the fused features with the most informativeness. We evaluated our proposed method on two multimodal rumour detection datasets. The experimental results show that both consistency and inconsistency between the text and image is essential for accurately classify rumours on social media. Compared with the existing co-attention mechanism, the proposed CBAN achieve 1-3% accuracy improvement in multimodal rumour detection tasks. Furthermore, our method shows promising performance when applied to other multimodal classification tasks, including sentiment analysis, sarcasm detection, crisis classification, and hate-speech detection.

Chapter 4

Author-Aware Rumour Detection

4.1 Motivation

Existing methods for rumour detection often rely on user comments or other external information, causing delays in detection due to the lagging of crowd signals or authorized evidence, making it difficult to meet early detection requirements [61]. Furthermore, the credibility of the source author is a critical factor in determining the veracity of a social media post, but it is often ignored in current approaches. Although some previous studies have utilized user social network relationships [62]–[64], these features require significant consumption quotas of Twitter API access to retrieve author attributes and associated users involved in the conversation, making them impractical for real-world applications. Additionally, most author-aware rumour detection methods do not make effective use of pretrained language models [64], relying on newly initialized parameters or new modules. This limits the efficiency of pretrained models, as the size of current benchmark datasets for rumour detection is small and hard to optimize these additional modules [65].

To address these challenges, this chapter proposes an approach for the early detection of rumours that leverages textual claims and the credibility of the source author. The method converts an author-aware rumour detection into a language modelling problem using pretrained language models, fine-tuning them to recognize the relationship between post content and author profiles. A multi-task learning framework is used to simultaneously identify both rumour claims and malicious accounts, improving overall detection accuracy. Our approach is parameter-efficient and can be easily integrated into most transformer-based pretrained models, minimizing newly initialized parameters to be updated.

The chapter also proposes a Layer-Wise Parameter-Efficient Tuning (LWPET) strategy that optimizes the parameters of pretrained language models, reducing the computation and memory requirements for tuning these models for rumour detection. The proposed method is evaluated on three benchmark datasets, namely Twitter15 [28], Twitter16 [28], and CR-Twitter [66], and demonstrated to be effective in real-world rumour detection across English and Chinese datasets. In summary, this chapter presents an approach that addresses the key challenges in automatic rumour detection and demonstrates the potential for using pretrained language models for the early detection of rumours with improved accuracy.

4.2 Tuning Language Models for Author-Aware Rumour Detection

In this section, we describe our methodology for tuning language models for author-aware rumour detection. Our goal is to classify a pair of a source claim and an author profile for author-aware rumour detection. We consider it a text pair classification problem with two input streams. By doing so, it allows us to solve the problem directly with the pretrained language models without additional feature extraction or encoding modules. This minimizes the number of newly initialised parameters during fine-tuning.

Given a pair of user representation U and text representation T, our goal is to fine-tune a pretrained language model M, parametrized by θ, to maximize the probabilities of predicted class $\boldsymbol{y} = \{y_0, y_1\}$, where y_0 and y_1 denote non-rumour and rumour, respectively.

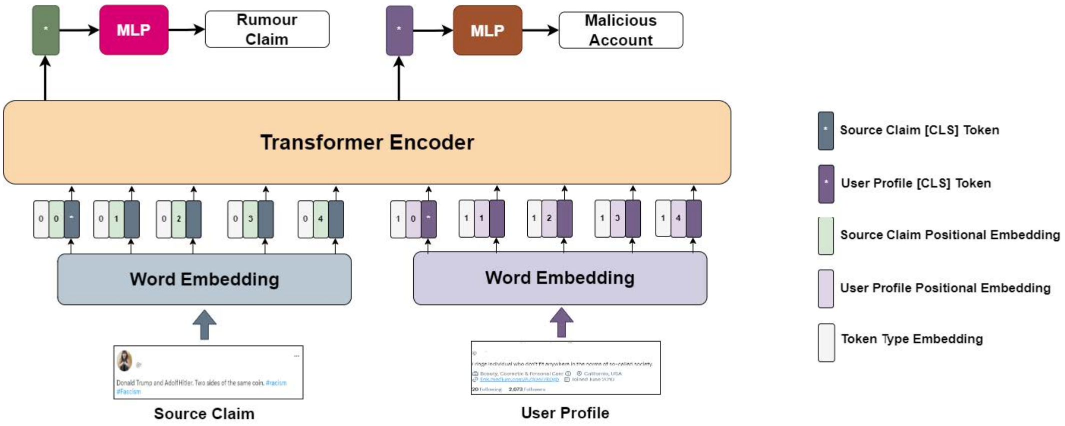


Figure 4.1 Illustration of rumour detection with author profile injection to pretrained language models.

4.2.1 Author Profile Injection for Rumour Detection with Pretrained Language Models

Motivated by multimodal Transformers [67], we consider the source claim T and user profile U as two modalities and then are inputted to a Transformer model for learning the semantic information and long-range dependencies between the two inputs, as shown in Fig. 4.1. We first send the source claim T and user profile U to the word embedding layers, resulting in $\boldsymbol{T_{emb}} \in \mathbb{R}^{m \times d_e}$ and $\boldsymbol{U_{emb}} \in \mathbb{R}^{n \times d_e}$, where d_e is the embedding dimension of the pretrained language models. m and n denote the number of tokens of the source claim and user description, respectively. Then, we add the position encoding and the type encoding to the input embedding, to obtain input features before sending it to the transformer encoder. We first compute the source claim features $\overline{\boldsymbol{T}} \in \mathbb{R}^{m \times d_e}$, as follows:

$$\overline{\boldsymbol{T}} = \boldsymbol{T_{emb}} + \boldsymbol{T_{pos}} + \boldsymbol{T_{type}}. \tag{4.1}$$

where $\boldsymbol{T_{pos}}$ and $\boldsymbol{T_{type}}$ represents the position embedding and the type embedding, respectively. Similarly, we obtain the user features $\overline{\boldsymbol{U}} \in \mathbb{R}^{n \times d_e}$ by adding the position embedding $\boldsymbol{U_{pos}}$ and the type of embedding $\boldsymbol{U_{type}}$, as follows:

$$\overline{\boldsymbol{U}} = \boldsymbol{U_{emb}} + \boldsymbol{U_{pos}} + \boldsymbol{U_{type}}. \tag{4.2}$$

To classify rumour claims without author profiles, we send the text features $\overline{\boldsymbol{T}}$ into the Transformer-based pretrained language models, to obtain the [CLS] token of output source claim representation $\mathbf{t_{cls}} \in \mathbb{R}^{1 \times d_e}$, as follows:

$$\mathbf{t_{cls}} = \text{Transformer}(\overline{\boldsymbol{T}}). \tag{4.3}$$

To classify malicious users, we send the user features $\overline{\boldsymbol{U}}$ into the same model, to obtain the [CLS] token of output user representation $\mathbf{u_{cls}} \in \mathbb{R}^{1 \times d_e}$, as follows:

$$\mathbf{u_{cls}} = \text{Transformer}(\overline{\boldsymbol{U}}). \tag{4.4}$$

For the author-aware rumour task, we first concatenate the text and user features to obtain the fine-grained features $\overline{\boldsymbol{F}}$, before sending to the Transformer models, as follows:

$$\overline{\boldsymbol{F}} = [\overline{\boldsymbol{T}}, \overline{\boldsymbol{U}}]. \tag{4.5}$$

Similarly, we obtain the output representation of the hybrid features $\mathrm{f}_{cls} \in \mathbb{R}^{1 \times d_e}$, as follows:

$$\boldsymbol{f_{cls}} = \text{Transformer}(\overline{\boldsymbol{F}}). \tag{4.6}$$

Then, the text features $\mathbf{t_{cls}}$, user features $\mathbf{u_{cls}}$, and hybrid features $\boldsymbol{f_{cls}}$ are used for classifying rumour claims, malicious accounts, and author-aware rumour claims, respectively.

4.2.2 Multi-Task Learning for Rumour Claim and Malicious Account Detection

Having obtained the feature representation f_{cls} of an author-aware claim, we use the Softmax classifier to predict whether it is a rumour or not, as follows:

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{W_c f_{cls}} + \boldsymbol{b_c}), \tag{4.7}$$

where $\mathbf{W}_{\boldsymbol{c}} \in \mathbb{R}^{2\times d}$ and $\boldsymbol{b_c} \in \mathbb{R}^2$ are trainable parameters. We employ the cross-entropy loss as the objective function in our proposed method. Given the predicted label $\hat{\boldsymbol{y}}$ and the ground-truth label $\boldsymbol{y}$, the negative log-likelihood is minimized. Thus, we have

$$\mathrm{loss} = -\big(\boldsymbol{y}\log(\hat{\boldsymbol{y}}) + (\boldsymbol{1} - \boldsymbol{y})\log(1 - \hat{\boldsymbol{y}})\big). \tag{4.8}$$

To improve the robustness of rumour detection, we apply multitask learning. Our goal is to train a neural network that can classify rumour claims and spam users. We simultaneously train our deep model on data with full and missing features, with a multitask loss. An illustration of multitasking learning on data with missing features is shown in Fig. 4.2.

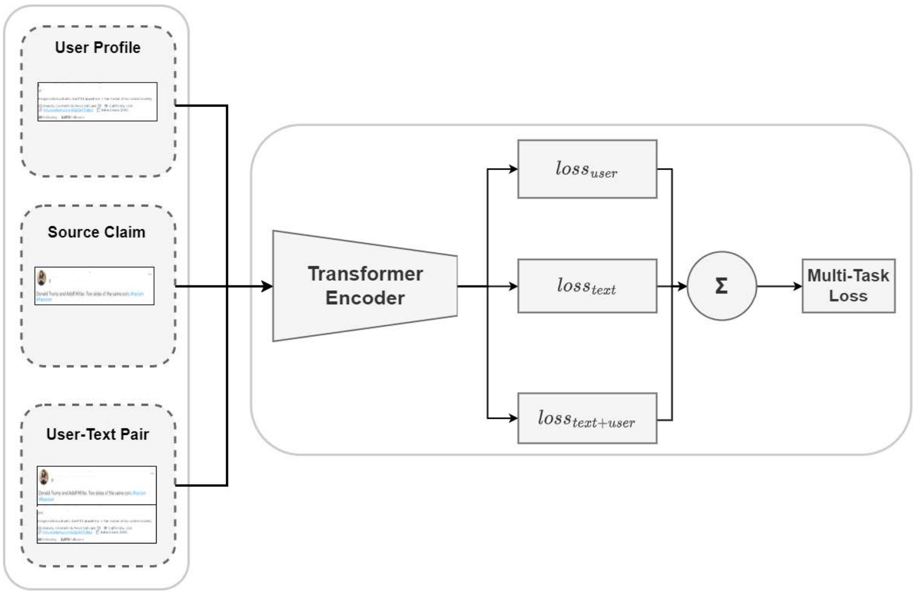


Figure 4.2 multi-task learning framework for rumour claim and malicious account detection.

We compute the multitask learning loss as follows:

$$\text{loss}_{Total} = \text{loss}_{text} + \text{loss}_{user} + \text{loss}_{multi}, \tag{4.9}$$

where loss_{text}, loss_{user}, and loss_{multi} represent the rumour claim classification, malicious account classification, and author-aware rumour classification, respectively.

4.2.3 Layer-Wise Parameter-Efficient Tuning (LWPET)

We propose a Layer-Wise Parameter-Efficient-Tuning (LWPET) method for tuning the pretrained language models, as shown in Fig. 4.3. We divide a Transformer-based language model into three parts, i.e., bottom layers, intermediate layers, and upper layers. We freeze the layers from layer 1 to layer b, so that the parameters of these layers can be shared but not updated during tuning. Then, we share the layers from layer b to layer $c - 1$. The parameters of these layers will be updated and shared across different tasks during tuning. Finally, we duplicate the pretrained parameters from layer c to the last layer, so that these task-specific layers will not be shared during multi-task learning, i.e., the parameters will be updated according to the errors made by the corresponding tasks, including rumour claim detection, malicious account detection, and author-aware rumour detection.

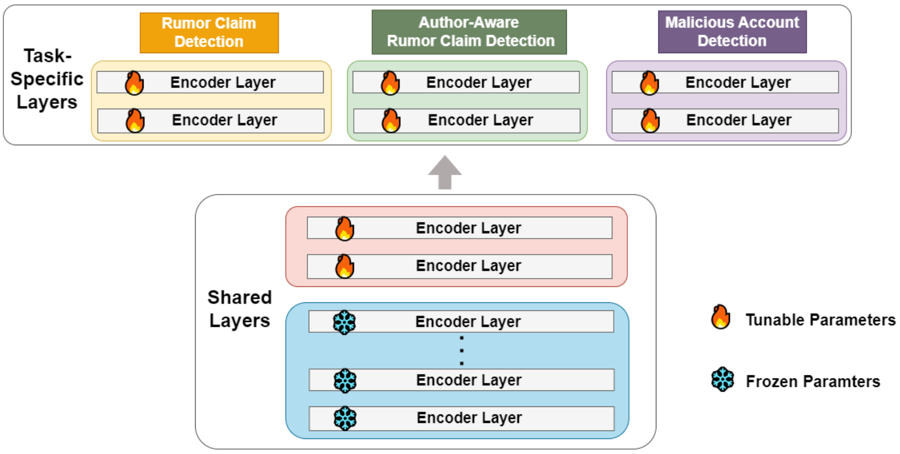


Figure 4.3 Illustration of Layer-Wise Parameter Efficient Tuning (LWPET).

4.3 Experimental Setup

4.3.1 Datasets

We evaluate our proposed method using three datasets for fine-grained rumour classification: Twitter15 [28], Twitter16 [28], and CR-Twitter [66]. As the author profiles are not provided in the previous datasets, we collect the author descriptions ourselves. In our experiments, Twitter15 contains 374 rumours, and 1116 non-rumours, while Twitter16 contains 205 rumours and 613 non-rumours. The CR-Twitter dataset contains 3616 rumours and 6295 non-rumours. We retrieved the author descriptions using the academic Twitter API. It is worth noting that comment tweets are not required in our approach. This greatly reduces the financial budget in industrial applications, as there is a tweet consumption cap that limits the number of tweets retrieved for a paid account for business use. Since some of the tweets and users have been removed or disabled from the social media website, the number of tweets in our experiment is smaller than that in the previous works. Therefore, we run all Transformer-based text classification baselines using our own sets of available data for a fair comparison. We divided all datasets into a ratio of 80:20 for training and testing, respectively, and used 10% of the training set as a validation set to tune the hyperparameters. After obtaining the optimal hyperparameters, we trained the model using all training samples and reported the testing results.

4.3.2 Experimental Setup

Evaluation Metrics: To evaluate the performance of different classification models, we use macro-F1 score, i.e., the mean of the F1 scores for all three classes, and accuracy to evaluate the performance of different classification methods for rumour detection.

Hyperparameter: We trained the models for 30 epochs with a mini-batch size of 16, using the Adam optimizer with a learning rate of 0.00002. To avoid overfitting, we applied L2 regularization with a rate of 0.001 and a dropout rate of

0.1. For the LWPET, we set the values $b = 8$, $c = 10$, so that the two Transformer layers would be task-specific layers, and another two layers would be tuneable shared layers, and the rest of the bottom layers are frozen during fine-tuning, to achieve the goal of parameter-efficient tuning.

4.4 Results

4.4.1 Results of Author-Profile Injection

Table 4.1 presents the results of experiments conducted to evaluate the performance of different language models for detecting rumours in social media. As the goal of this work is to investigate the effects of author profiles for rumour detection, instead of establishing state-of-the-art performance for rumour detection, we compare our approach with the Transformer-based baselines for text classification. We experiment our approach with the state-of-the-art pretrained language models, including BERT [41], RoBERTa [68], DistilBERT [69], and DeBERTa [70]. The results show that all the models perform better when author profile injection is used, indicating the importance of user information in detecting rumours. Among the models evaluated, DeBERTa with author profile injection achieved the highest Macro-F1 score of 0.9366 and the highest Accuracy of 0.9486 on the Twitter15 dataset. Similarly, on the Twitter16 dataset, BERT with author profile injection achieved the highest Macro-F1 score of 0.9270 and the highest Accuracy of 0.9389. On the CR-Twitter dataset, RoBERTa with author profile injection achieved the highest Macro-F1 score of 0.8174 and the highest Accuracy of 0.8282.

Table 4.1 Results of author-aware rumour detection on Twitter15, Twitter16, and CR-Twitter datasets.

Dataset	Method	Author Profile Injection	Macro-F1	Accuracy
Twitter15	BERT	✗	0.8505	0.8814
		✓	**0.9413**	**0.9526**
	RoBERTa	✗	0.8985	0.9170
		✓	0.9366	0.9486
	DistilBERT	✗	0.8472	0.8735
		✓	0.9372	0.9486
	DeBERTa	✗	0.8958	0.9130
		✓	0.9016	0.9170
Twitter16	BERT	✗	0.8328	0.8702
		✓	**0.9270**	**0.9389**
	RoBERTa	✗	0.8358	0.8702
		✓	0.8888	0.9084
	DistilBERT	✗	0.8440	0.8779
		✓	0.9073	0.9237
	DeBERTa	✗	0.8085	0.8473
		✓	0.8786	0.9008
CR-Twitter	BERT	✗	0.7996	0.8134
		✓	**0.8302**	**0.8411**
	RoBERTa	✗	0.7927	0.8103
		✓	0.8174	0.8282

4.4.2 Ablation Study

Table 4.2 presents the results of experiments conducted to evaluate the effectiveness of different models for rumour detection on three datasets: Twitter-15, Twitter-16, and CR-Twitter. We use BERT [41] as the pretrained model in the experiments. The models evaluated in the experiments include the proposed methods without author profile injection or multi-task learning.

Table 4.2 Ablation studies on author-aware rumour detection.

Dataset	Model	Macro-F1	Accuracy
Twitter-15	Ours w/o user profiling	0.8505	0.8814
	Ours w/o multi-task learning	0.9236	0.9368
	Ours (Full)	**0.9413**	**0.9526**
Twitter-16	Ours w/o user profiling	0.8328	0.8702
	Ours w/o multi-task learning	0.8988	0.9160
	Ours (Full)	**0.9270**	**0.9389**
CR-Twitter	Ours w/o user profiling	0.7996	0.8134
	Ours w/o multi-task learning	0.8162	0.8288
	Ours (Full)	**0.8302**	**0.8411**

The results show that the full model outperforms the other two models on all three datasets in terms of both Macro-F1 and Accuracy. For example, on the Twitter-15 dataset, the full model achieved a Macro-F1 score of 0.9281 and an Accuracy of 0.9447, while the model without author profiling achieved a Macro-F1 score of 0.8505 and an Accuracy of 0.8814. Similarly, on the Twitter-16 dataset, the full model achieved a Macro-F1 score of 0.9270 and an Accuracy of 0.9389, while "Ours w/o user profiling" achieved a Macro-F1 score of 0.8328 and an Accuracy of 0.8702.

The results indicate that user profiling and multi-task learning contribute significantly to improving the performance of the model for rumour detection. The model without profiling performs worse than the full model on all three datasets, highlighting the importance of considering user information in rumour detection. Similarly, the model without multi-task learning performs worse than the full model on Twitter-15 and Twitter-16 datasets, indicating the benefit of using multi-task learning to improve the performance of the model.

4.4.3 Comparison to Other Parameter-Efficient Methods

Table 4.3 shows the performance of different tuning strategies on three different datasets for rumour detection. We use BERT [41] as the pretrained model in the experiments. The strategies include fine-tuning [41], adapter-tuning with bottleneck [71], adapter-tuning with LORA [72], prefix-tuning [73], and the proposed LWPET.

Table 4.3 Overall results with different tuning strategies.

Dataset	Tuning Strategy	Macro-F1	Accuracy
Twitter-15	Fine-Tuning [41]	0.9040	0.9249
	Adapter-Tuning with Bottleneck [71]	0.8753	0.8933
	Adapter-Tuning with LORA [72]	0.8376	0.8656
	Prefix-Tuning [73]	0.8623	0.8933
	LWPET (Ours)	**0.9413**	**0.9526**
Twitter-16	Fine-Tuning [41]	0.9270	0.9389
	Adapter-Tuning with Bottleneck	0.8641	0.8855
	Adapter-Tuning with LORA [72]	0.7694	0.8244
	Prompt-Tuning [73]	0.7585	0.8092
	LWPET (Ours)	**0.9270**	**0.9389**
CR-Twitter	Fine-Tuning [41]	0.8050	0.8153
	Adapter-Tuning with Bottleneck [71]	0.8017	0.8147
	Adapter-Tuning with LORA [72]	0.7858	0.8023
	Prompt-Tuning [73]	0.7912	0.8073
	LWPET (Ours)	**0.8302**	**0.8411**

From table 4.3, it can be observed that LWPET consistently outperforms the other tuning strategies across all datasets. On the Twitter-15 dataset, LWPET achieves a Macro-F1 of 0.9413 and an accuracy of 0.9526, which is the highest among all strategies. On the Twitter-16 and CR-Twitter datasets, LWPET achieves a Macro-F1 of 0.9270 and 0.8302, respectively, which is either the highest or very close to the highest among all strategies. On the other hand, adapter tuning with

LORA consistently performs the worst among all tuning strategies. For instance, on the Twitter-15 dataset, adapter-tuning with LORA achieves a Macro-F1 of only 0.8376, which is significantly lower than other strategies. It is worth noting that adapter-tuning and prompt-tuning require additional parameters to be injected into the pretrained language models. This requires extra computation for inference. On the other hand, the proposed LWPET does not addition modules. This makes the inference speed of the LWPET the same as the original fine-tuning methods.

4.4.4 Comparison of Different Fusion Strategies

Table 4.4 presents the performance of various fusion strategies for rumour detection on three datasets: Twitter-15, Twitter-16, and CR-Twitter. We use BERT [41] as the pretrained model in the experiments. The fusion strategies include feature concatenation in the output feature space, score aggregation in the output logit, and token concatenation in embedding.

Table 4.4 Classification results with different fusion strategies.

Dataset	**Fusion Strategies**	**Macro-F1**	**Accuracy**
Twitter-15	Feature concatenation in output feature	0.8968	0.9170
	Score aggregation in output logit	0.8737	0.9012
	Token concatenation in input embedding	**0.9236**	**0.9368**
Twitter-16	Feature concatenation in output feature	0.8085	0.8473
	Score aggregation in output logit	0.8468	0.8779
	Token concatenation in input embedding	**0.8988**	**0.9160**
CR-Twitter	Feature concatenation in output feature	0.7964	0.8159
	Score aggregation in output logit	0.7948	0.8159
	Token concatenation in input embedding	**0.8162**	**0.8288**

For Twitter-15, the token concatenation in embedding achieves the highest Macro-F1 (0.9236) and accuracy (0.9368), outperforming the other two fusion strategies. For Twitter-16, token concatenation also performs the best with a Macro-

F1 of 0.8988 and an accuracy of 0.9160. On the other hand, feature concatenation in the output feature performs the worst on Twitter-16, with a Macro-F1 of 0.8085 and an accuracy of 0.8473.

For CR-Twitter, the token concatenation in embedding again performs the best with a Macro-F1 of 0.8162 and an accuracy of 0.8288. Feature concatenation and score aggregation both achieve similar results with Macro-F1 scores around 0.79 and accuracy scores around 0.81.

In summary, the token concatenation in the embedding strategy consistently performs the best among the three fusion strategies on all three datasets. The results suggest that integrating contextual information by embedding token sequences may be more effective than combining features or scores in the output. However, further investigation and experimentation are needed to determine the optimal fusion strategy for rumour detection in different contexts and applications.

4.4.5 Visualization of Classification Results

Table 4.5 shows examples of rumour and non-rumour claims with the author's descriptions. To classify the statements as rumours or non-rumours, we need to For the English Tweets, Table 4.5 (a) suggests that Hillary Clinton is lying about being the first woman nominated for president, and the source of the information is not provided. Therefore, this statement cannot be verified, and it is classified as a rumour. Table 4.5 (b) reports a recommendation from a credible source, CNN Digital, that doctors should screen all adults for depression at least once. This statement can be verified, and it is classified as a non-rumour.

For the Chinese Tweets, Table 4.5 (c) reports an incident in which a person in Jinjiang, Fujian, concealed their travel history to Wuhan, leading to 3-4 thousand people needing to be monitored. The source of the information is RFA Radio Free Asia political cartoonist/personal remarks, which may not be a reliable source for such news. Therefore, this statement is classified as a rumour. Table 4.5 (d) reports the movement of Hong Kong Chief Executive Carrie Lam from Shanghai to Nanjing to attend a summit and meet with leaders and businessmen. The source of the information is the Hong Kong SAR Government Online News Platform, which is a credible source. Therefore, this statement can be verified, and it is classified as a non-rumour.

Table 4.5 Examples of rumour and non-rumour claims with author's descriptions.

	Text	User	Label
(a)	Hillary Clinton is lying. Clinton is NOT the first woman nominated for president.	Just google me. It's pretty funny.	Rumour
(b)	Doctors should screen all adults for depression at least once, the task force recommends.	Breaking news from CNN Digital.	Non-Rumour
(c)	福建晉江某人隱瞞武漢旅行史，過年期間活躍出席當地各種公開活動和宴席，導致 3-4 千人需要被監控。 (A person in Jinjiang, Fujian concealed his travel history to Wuhan, and actively attended various local public events and banquets during the Chinese New Year, resulting in 3-4 thousand people needing to be monitored.)	RFA 自由亞洲電台政治漫畫家/個人言論与 RFA 公司立場無關 (RFA Radio Free Asia political cartoonist/personal remarks have nothing to do with RFA's position)	Rumour
(d)	林鄭月娥由上海轉往南京 ，出席第二屆蘇港融合發展峰會，並與江蘇省領導會面，又與港商交流。 (Carrie Lam transferred from Shanghai to Nanjing to attend the 2nd Suzhou-Hong Kong Integration Development Summit, met with leaders of Jiangsu Province, and communicated with Hong Kong businessmen.)	香港特區政府網上新聞平台 (Hong Kong SAR Government Online News Platform)	Non-Rumour

4.5 Summary

In conclusion, this chapter introduces an early detection approach for identifying rumours on social media using pretrained language models and author profiles. The proposed method employs a multi-task learning framework to simultaneously identify rumour claims and malicious accounts, resulting in improved detection accuracy. This chapter also introduces a Layer-Wise Parameter-Efficient Tuning (LWPET) strategy that optimizes pretrained language models' parameters, reducing the computation and memory requirements. The proposed approach outperforms state-of-the-art baselines on three benchmark datasets, namely Twitter15, Twitter16, and CR-Twitter, demonstrating its effectiveness in real-world rumour detection across English and Chinese datasets. This illustrates the importance of considering author profile in rumour detection tasks, as the profile may contain the credibility of the authors which is a useful feature for classifying rumours on social media.

Chapter 5

Stream Classification of Rumours

5.1 Motivation

In recent years, Graph Neural Networks (GNNs) and Transformers are both commonly used neural networks for rumour analysis. The former is used to model the graphical propagation pattern of social media conversations, while the latter is for capturing the sequential relationship between the source and replies. Both neural networks have shown promising performance in rumour classification. However, these two neural networks are non-causal. This means that the features extracted from a reply depend on the features from future replies, resulting in low efficiency in stream mining [74], because the features of existing nodes will be updated if a new node is linked to the graph.

Stream classification of rumours means that the rumours can be verified instantly whenever a reply is posted, as shown in Fig. 5.1. This makes early rumour verification feasible, which is particularly important, especially when the number of replies is small in the early stage of propagation [75]. Therefore, the main challenges of rumour analysis include increasing the efficiency of stream classification and improving the accuracy of early rumour verification.

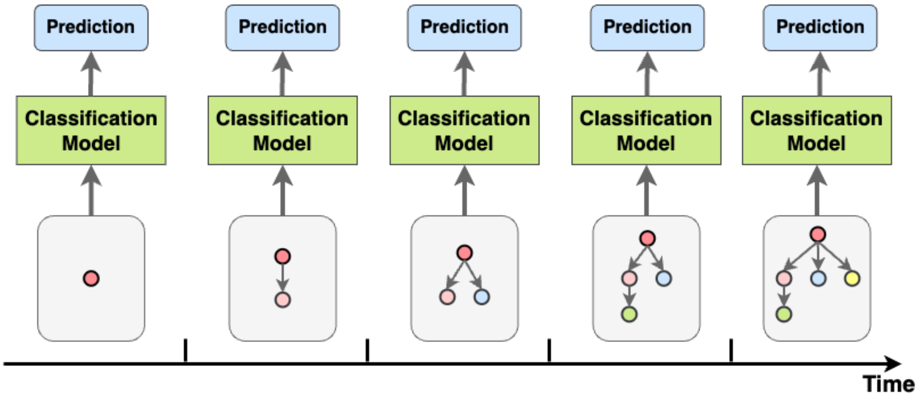


Figure 5.1 Stream verification of rumour. A circle represents a message posted at different time instances.

To address the abovementioned issues, this work makes the following contributions:

- We propose a Causal Diffused Graph-Transformer Network (CDGTN) to encode conversations on social media for rumour verification. The proposed network consists of a graph-aware causal-masked Transformer network. The hidden representations of a node sequence are independent of the future nodes, making it particularly suitable for the continuous classification of streaming posts.
- To fuse the encoded node embedding sequences, we propose a Source-Guided Incremental Attention Pooling (SGIAP) to aggregate the sequence of node features into a fixed-length representation at different timestamps. We consider the attention-pooled features from different timestamps for early rumour classification.
- To enhance the early classification of rumours, we propose a Stacked Early Classification Loss (SecLoss) that aims to minimize the overall classification loss of the node sequence predicted at every timestamp during training.
- An efficient and scalable framework is proposed for streaming rumour classification, which aims to continuously verify rumours in the streaming

of social media posts, by efficiently re-utilizing the extracted features propagated from the previous timestamp, using a cumulative average.

- To facilitate research in low-resource language rumour verification, we annotated a fine-grained Chinese rumour classification dataset based on the CR-Twitter dataset. The extended CR-Twitter dataset contains 143 unverified rumours, 303 true rumours, 405 false rumours, and 1334 non-rumours.
- We conducted experiments on the existing Twitter15, Twitter16, PHEME, Weibo, and extended CR-Twitter datasets. Our proposed method consistently improves the classification performance when the number of replies is reduced in early rumour classification experiments.

5.2 The Proposed Framework for early rumour verification

In this section, we present our proposed framework for early rumour verification, which comprises four main components: a Causal Diffused Graph-Transformer Network (CDGTN), a Source-Guided Incremental Attention Pooling (SGIAP), a Stacked Early Classification Loss (SecLoss), and a continued inference algorithm for stream classification. For binary rumour detection, our goal is to classify a source-reply graph G into rumour and non-rumour. For fine-grained rumour classification, our objective is to classify the propagation graph into one of four categories: non-rumours, false rumours, true rumours, or unverified rumours.

5.2.1 Causal Diffused Graph-Transformer Network (CDGTN)

Local features and global information are crucial for predicting the veracity of rumours. Local features refer to the aggregation of information between a reply and its linked reply, while global features refer to the flow of information from the source post to a reply. In Fig. 5.2, we propose integrating a graph neural network and Transformer to form a model called Causal Diffused Graph-Transformer

Network (CDGTN). Here, 'causal' means that the network generates an output embedding sequence that depends only on the inputs at the current and previous time steps. 'Diffused' means that the encoder takes into account the information diffusion process from the source reply to the latest reply. This model combines the top-down graph neural network and the unidirectional Transformer encoder, allowing us to simultaneously model the local and global information. Similar to previous research on hierarchical Transformer networks [76], the graph-Transformer network can process both local features and global features in a conversation. However, the hierarchical Transformer network may ignore the diffused graphical propagation patterns, which are important features for rumour analysis. Thus, we propose a diffused graph-Transformer network that can extract local and global propagation patterns in our rumour verification framework.

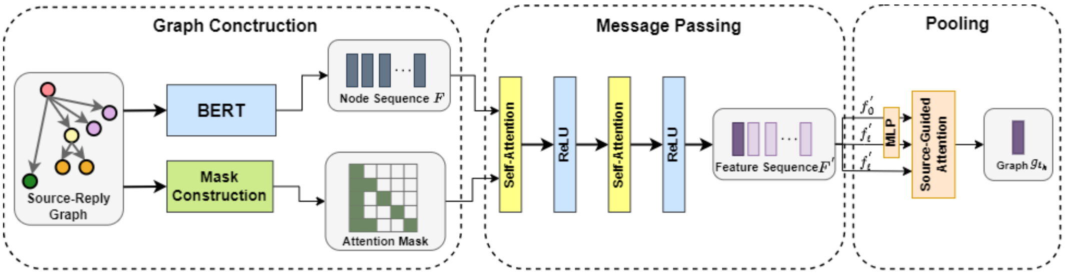


Figure 5.2 The proposed Causal Diffused Graph-Transformer Network (CDGTN).

To obtain a sequence of features F from a conversation, we first transform the message, which is either a source or a reply, into a fixed-length representation. Each message is encoded by a pretrained bidirectional Transformer (BERT) [41] into a d-dimensional vector. Specifically, we use the hidden representation of the first token, i.e., adding a special learnable embedding vector at the beginning of every sentence, before sending it to the pretrained BERT, to form the overall representation of the sentence. For each message m_i, we encode it into an embedding vector $\boldsymbol{f}_i \in \mathbb{R}^d$, as follows:

$$\boldsymbol{f}_i = \mathrm{BERT}(\boldsymbol{m}_i). \tag{5.1}$$

After processing the source post and the replies with sentence embeddings, the feature sequence $\boldsymbol{F} = \{\boldsymbol{f_s}, \boldsymbol{f_1}, \boldsymbol{f_2}, \dots, \boldsymbol{f_{n-1}}\} \in \mathbb{R}^{d \times n}$ is obtained, where the first element in $\boldsymbol{F}$, i.e., $\boldsymbol{f_s}$, is the feature representation of the source information, and the other elements are the feature representation of the replies, i.e., n–1 replies.

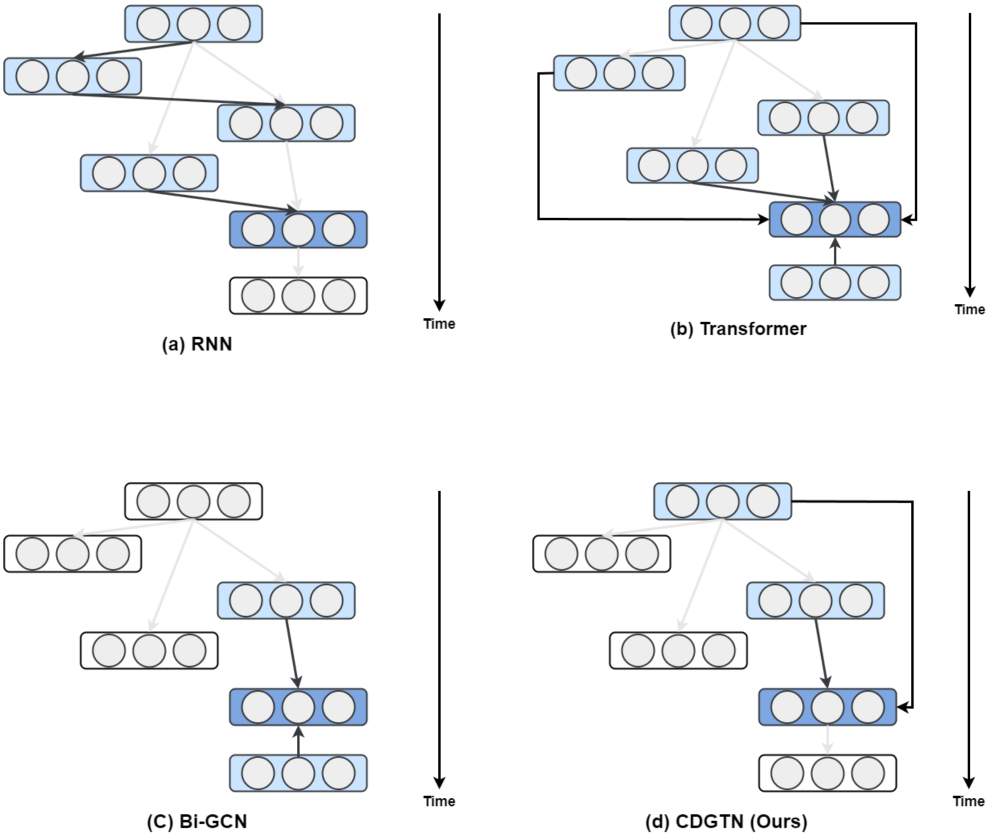


Figure 5.3 Message passing among different models.

Given a source tweet and its replies, we use a graph-aware causal masked Transformer to learn the diffusion process in the source-reply graph. The inputs of the graph-aware causal masked Transformer are the feature sequence matrix $\boldsymbol{F}$ and a graph-aware attention mask. Our goal is to enhance the representation power of the feature for each node feature, by considering the information flow from the source post to a reply.

Full self-attention in the standard Transformer network will result in non-causal behaviour in the transformed features, as shown in Fig. 5.2. Therefore, we consider the graph-aware causal mask, which forces the query not to attend to the key and value vectors in future positions. Furthermore, we only compute the attention scores that belong to the path of positions from the source post to the latest reply. This greatly reduces the computational complexity compared to the full attention mask. The Transformer model consists of several multi-head attention modules. A multi-head attention module consists of h heads of scaled dot-product attention blocks. Each scaled dot-product attention block accepts three inputs, i.e., query $\boldsymbol{\mathcal{Q}}$, key $\boldsymbol{\mathcal{K}}$, and value $\boldsymbol{\mathcal{V}}$. The query matrix $\boldsymbol{\mathcal{Q}}$ is used to compute the attention weights by measuring its similarity to the key matrix $\boldsymbol{\mathcal{K}}$. Then, the attention weights are multiplied by the value $\boldsymbol{\mathcal{V}}$ to obtain the attention vectors. Mathematically, the attention score matrix $\boldsymbol{S}$ is calculated, as follows:

$$\boldsymbol{S} = \mathrm{Softmax}(\frac{\mathcal{Q}\mathcal{K}^{\mathrm{T}}}{\sqrt{d_k}}), \tag{5.2}$$

where d_k is the embedding dimension of the key $\boldsymbol{\mathcal{K}}$. After calculating the attention score matrix $\boldsymbol{S}$, the attention value $\boldsymbol{\mathcal{V}'}$ can be computed as a linear combination of the attention score $\boldsymbol{S}$ and key $\boldsymbol{\mathcal{K}}$. Therefore, we have

$$\boldsymbol{\mathcal{V}'} = \boldsymbol{S\mathcal{V}}, \tag{5.3}$$

The multi-head attention mechanism repeats the scaled dot-product attention h times and aggregates the attention value $\boldsymbol{\mathcal{V}'}$ to obtain a more robust representation of the value $\boldsymbol{\mathcal{V}}$, as follows:

$$\mathrm{MultiHead}(\boldsymbol{\mathcal{Q}}, \boldsymbol{\mathcal{K}}, \boldsymbol{\mathcal{V}}) = \mathrm{Concat}(\boldsymbol{V'_1W'_1}, \boldsymbol{V'_2W'_2}, \dots, \boldsymbol{V'_hW'_h})\boldsymbol{W^o}, \tag{5.4}$$

where $\mathrm{MultiHead}(\,)$ and $\mathrm{Concat}(\,)$ represent the multi-head attention mechanism and the concatenation operation, respectively. It is worth noting that $\boldsymbol{W^o}$ and $\boldsymbol{W'_i}$ are trainable parameters jointly learned through backpropagation. Finally, the transformer uses skip connections to add the output of the multi-head attention with the value $\boldsymbol{\mathcal{V}}$.

Since our goal is to obtain a more comprehensive graph representation, we set the query $\boldsymbol{\mathcal{Q}}$, key $\boldsymbol{\mathcal{K}}$, and value $\boldsymbol{\mathcal{V}}$ equal to the sequence feature F, i.e., $\boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{K}} = \boldsymbol{\mathcal{V}} = \boldsymbol{F}$, in Equation (5.4), whose output is denoted as $\mathbf{F}' \in \mathbb{R}^{d \times n}$. In a real implementation, we can set the attention weights of those masked positions to negative infinity [77], such that the attention scores of the masked positions will become zero in Equation (5.2).

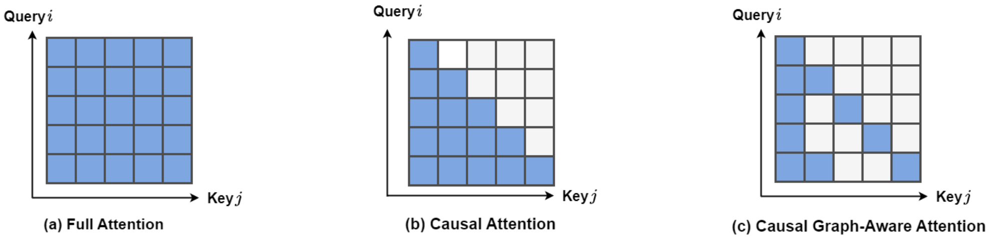


Figure 5.4 The attention mask in CDGTN.

5.2.2 Source-Guided Incremental Attention Pooling (SGIAP)

In the early stream verification of rumours, our goal is to predict the veracity of the source rumours as early as possible, ideally with fewer replies in the propagational graph. To achieve this, we propose source-guided incremental attention pooling. We consider the source post $\boldsymbol{f}_0'$ as the context vector, to compute attention scores, which measure the correlation of a reply to a source post, for every reply in a propagation graph. We denote t_k as the final time of a propagation graph. We use Multilayer Perceptron (MLP) to calculate the importance of the feature a_t at any timestamp t, given by:

$$a_t = \boldsymbol{W}_2 \text{Tanh}\,(\boldsymbol{W}_1(\,\boldsymbol{f}_0' \oplus \boldsymbol{f}_{\mathrm{t}}'\,)), \tag{5.5}$$

$$\widehat{a_t} = \frac{e^{a_{\mathrm{t}}}}{\sum_{t=1}^{t_k} e^{a_{\mathrm{t}}}}. \tag{5.6}$$

where $\oplus$ represents the concatenation operator, and $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ are learning parameters. The final representation of a conversation at the final time t_k is calculated by:

$$\boldsymbol{g_{t_k}} = \sum_{t=1}^{t_k} \widehat{a_t} \boldsymbol{f'_t}. \tag{5.7}$$

It is worth noting that Equations (5.6) and (5.7) are the standard equations of a typical attention mechanism. To adapt the attention mechanism for stream classification, we slightly modify it by moving the denominator in Equation (6.6) and placing it in Equation (5.7). Therefore, Equations (5.6) and (5.7) can be formulated, as follows:

$$\boldsymbol{g_{t_k}} = \frac{1}{\sum_{t=1}^{t_k} e^{a_t}} \sum_{t=1}^{t_k} e^{a_\mathrm{t}} \boldsymbol{f'_t}. \tag{5.8}$$

By doing so, it allows us to buffer the sum of unnormalized weighted features $\sum_{t=1}^{t_k} e^{a_\mathrm{t}} \boldsymbol{f'_t}$, denoted as $\boldsymbol{c_{t_k}}$, and the sum of the unnormalized weights $\sum_{t=1}^{t_k} e^{a_t}$, denoted as d_{t_k}, as follows:

$$\boldsymbol{g_{t_k}} = \frac{\boldsymbol{c_{t_k}}}{d_{t_k}}. \tag{5.9}$$

This is particularly useful for the proposed stacked early classification loss and continued inference, as these two terms are independent of the future nodes.

5.2.3 Stacked Early Classification Loss (SecLoss)

In the trivial training of rumour classification algorithms, the graph representation at the final timestamp $\boldsymbol{g_{t_k}}$ of a conversation is used for classification, as follows:

$$\boldsymbol{\hat{y}_{t_k}} = \mathrm{softmax}\left(\boldsymbol{W_c g_{t_k}} + \boldsymbol{b_c}\right), \tag{5.10}$$

$$\mathrm{loss}_{\boldsymbol{t_k}} = -\sum_{i=1}^{\mathrm{c}} \mathbf{y} \log\,(\boldsymbol{\hat{y}_t}). \tag{5.11}$$

where $\mathbf{W}_c \in \mathbb{R}^{c \times d}$ and $\boldsymbol{b}_c \in \mathbb{R}^{c}$ are trainable parameters. Here $\hat{\boldsymbol{y}}_t$ denotes the prediction at time $\boldsymbol{t}$. C is the number of categories in the datasets.

However, simply minimizing the classification loss at the final timestamp does not ensure that the loss is minimized at earlier times during graph propagation. To address this issue, we propose minimising the summarisation of predictions at all timestamps $t \in \{0,1,2, \dots t_k\}$, instead of just the final time t_k, as shown in Fig. 6.5. To obtain the predictions at all timestamps, we first calculate the graph representations $\boldsymbol{g}_t$ at any time t. Due to the causality of the proposed encoder, we can accelerate the calculation of graph representations in a parallel manner.

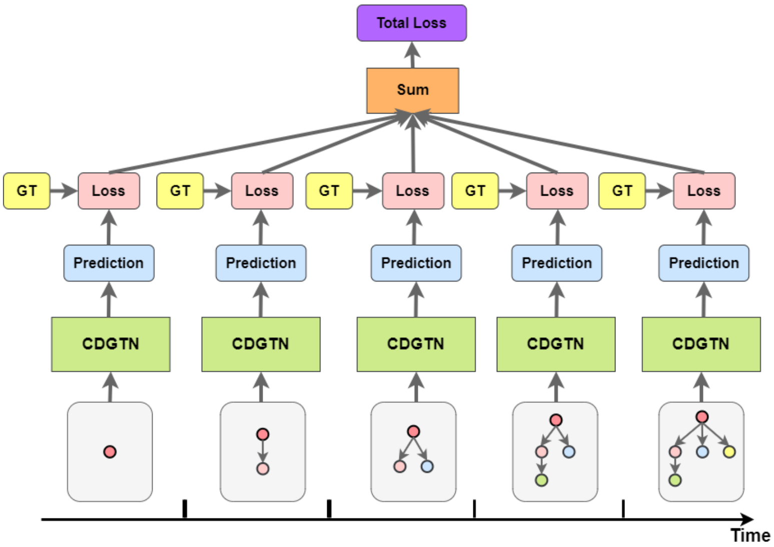


Figure 5.5 Illustration of the proposed Stacked Early Classification Loss (SecLoss).

Mathematically, we obtain the sum of unnormalized weighted features $\{\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3, \dots \boldsymbol{c}_{t_k}\}$ at any time t using a lower triangular matrix filled with ones, as follows:

$$\begin{Bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{Bmatrix} \times \begin{pmatrix} e^{a_1} f_1' \\ e^{a_2} f_2' \\ e^{a_3} f_3' \\ \vdots \\ e^{a_{t_k}} f_k' \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{1} e^{a_{\mathrm{t}}} \boldsymbol{f_t'} \\ \sum_{t=1}^{2} e^{a_{\mathrm{t}}} \boldsymbol{f_t'} \\ \sum_{t=1}^{3} e^{a_t} \boldsymbol{f_t'} \\ \vdots \\ \sum_{t=1}^{t_k} e^{a_t} \boldsymbol{f_t'} \end{pmatrix} = \begin{pmatrix} \boldsymbol{c_1} \\ \boldsymbol{c_2} \\ \boldsymbol{c_3} \\ \vdots \\ \boldsymbol{c_{t_k}} \end{pmatrix}. \quad (5.12)$$

Similarly, the sum of unnormalized weights $\{d_1, d_2, d_3, \dots d_{t_k}\}$ can be computed as follows:

$$\begin{Bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{Bmatrix} \times \begin{pmatrix} e^{a_1} \\ e^{a_2} \\ e^{a_3} \\ \vdots \\ e^{a_{t_k}} \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{1} e^{a_{\mathrm{t}}} \\ \sum_{t=1}^{2} e^{a_{\mathrm{t}}} \\ \sum_{t=1}^{3} e^{a_t} \\ \vdots \\ \sum_{t=1}^{t_k} e^{a_t} \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{t_k} \end{pmatrix}. \quad (5.13)$$

Therefore, we can obtain the graph representations $\{\boldsymbol{g_1}, \boldsymbol{g_2}, \boldsymbol{g_3} \dots, \boldsymbol{g_{t_k}}\}$ at all timestamps, as follows:

$$\begin{pmatrix} \frac{1}{d_1} \\ \frac{1}{d_2} \\ \frac{1}{d_3} \\ \vdots \\ \frac{1}{d_{t_k}} \end{pmatrix} \odot \begin{pmatrix} \boldsymbol{c_1} \\ \boldsymbol{c_2} \\ \boldsymbol{c_3} \\ \vdots \\ \boldsymbol{c_{t_k}} \end{pmatrix} = \begin{pmatrix} \frac{\boldsymbol{c_1}}{d_1} \\ \frac{\boldsymbol{c_2}}{d_2} \\ \frac{\boldsymbol{c_3}}{d_3} \\ \vdots \\ \frac{\boldsymbol{c_k}}{d_{t_k}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{g_1} \\ \boldsymbol{g_2} \\ \boldsymbol{g_3} \\ \vdots \\ \boldsymbol{g_{t_k}} \end{pmatrix}, \quad (5.14)$$

where $\odot$ represents Hadamard product operation. With the graph representation obtained at all timestamps $\{\boldsymbol{g_1}, \boldsymbol{g_2}, \boldsymbol{g_3} \dots, \boldsymbol{g_{t_k}}\}$, we use the Softmax classification in Equation (9) and minimize the summation of predictions at all timestamps $t \in \{0,1,2, \dots t_k\}$, as follows:

$$\mathrm{loss}_{stacked} = -\sum_{t=1}^{\boldsymbol{t_k}} \sum_{i=1}^{\mathrm{c}} \mathbf{y}_t \log(\hat{\boldsymbol{y}}_t). \quad (5.15)$$

Therefore, the proposed stacked early classification loss $\mathrm{loss}_{stacked}$ tries to minimize the total loss calculated by predictions at all timestamps $t \in \{0,1,2, \dots t_k\}$, which is particularly useful for the early classification of rumours.

5.2.4 Continued Inference in a Streaming Graph

Due to the causality of the graph-Transformer encoder, our encoded post-embedding is independent of future nodes. This means that the hidden representations are static for future nodes. Although it is possible to store the hidden representations of all nodes, we avoid redundant memory and computation for stream classification of rumours.

Based on the prefix-sum algorithm, which maintains the running total of the data as it grows with the time sequence, we keep a running sum of the unnormalized weighted features up to the previous timestamp t_{k-1}. The running total $\boldsymbol{c}_{t_k-1}$ up to the previous timestamp t_{k-1} is defined as follows:

$$\boldsymbol{c}_{t_k-1} = \sum_{t=1}^{t_{k-1}} e^{a_t} \boldsymbol{f}'_t. \tag{5.16}$$

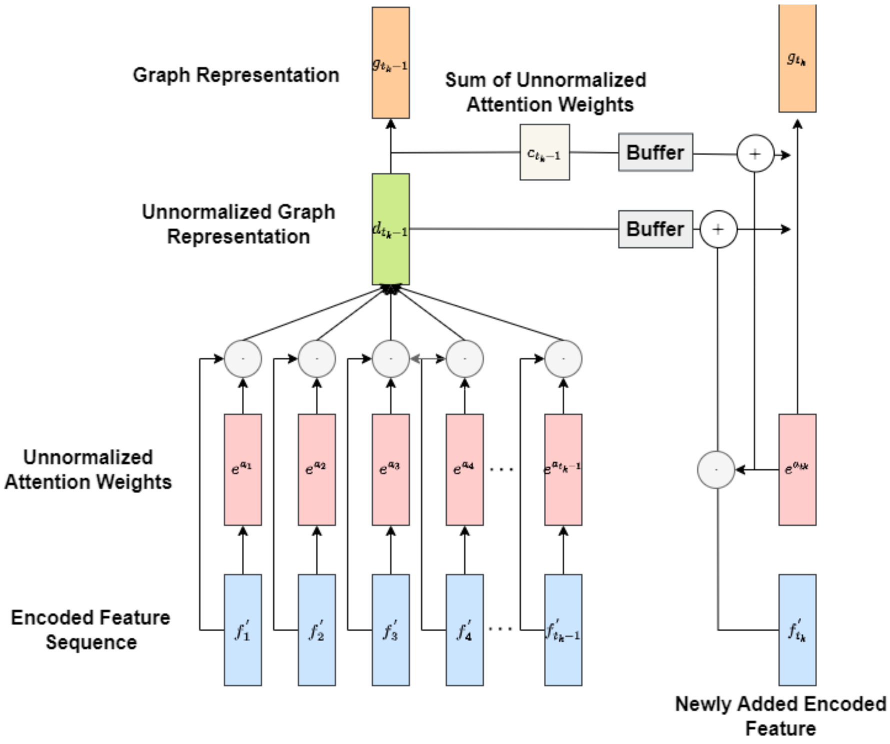


Figure 5.6 The proposed continue inference framework with buffering.

By buffering the running total of the unnormalized weighted features c_{t_k-1} at time t_{k-1} during stream classification, we compute the total sum of the feature vectors at the time t_k as follows:

$$\boldsymbol{c}_{\boldsymbol{t_k}} = \boldsymbol{c}_{\boldsymbol{t_{k-1}}} + e^{a_{t_k}} \boldsymbol{f'}_{\boldsymbol{t_k}}, \tag{5.17}$$

Similarly, by buffering the running total of the unnormalized weights at the time t_{k-1} during stream classification, the total sum of the unnormalized weights at the time t_k is computed as follows:

$$d_{t_k-1} = \sum_{t=1}^{t_{k-1}} e^{a_t}, \tag{5.18}$$

$$d_{t_k} = d_{t_k-1} + e^{a_{t_k}}. \tag{5.19}$$

After that, we obtain the graph representation and perform classification using Equations (6.9)–(6.11). Therefore, by buffering the sum of the unnormalized hidden features and the sum of the unnormalized weights from the last timestamp, we reduce the computational complexity from linear complexity O(n) to constant complexity O(1) in Equation (5.8). In other words, we avoid redundant storage and calculation of hidden node features in the holistic propagational graph during stream classification.

5.3 Experimental Setup and Results

In this section, we first describe the datasets and experimental setup. Then, we evaluate our proposed method and compare it with other state-of-the-art methods. After that, we show the results of early rumour verification including classification performance and complexity analysis. Finally, we visualize some examples of fine-grained rumour classification.

5.3.1 Datasets

We evaluate our proposed method using five datasets for rumour classification: Twitter15 [78], Twitter16 [78], and CR-Twitter [79]. Twitter15 contains 374 unverified rumours, 372 true rumours, 370 false rumours, and 373 non-rumours, while Twitter16 contains 201 unverified rumours, 207 true rumours, 205 false rumours, and 205 non-rumours. The CR-Twitter dataset is a collection of Chinese rumour and non-rumour tweets without veracity labels. We extend CR-Twitter by annotating the veracity labels of the rumours. To obtain reliable annotations, we collected the existing source tweets and replies using the Twitter API and manually collected rumours from Chinese rumour debunking websites. We annotated the rumours based on the veracity tags provided on these fact-checking websites, resulting in 143 unverified rumours, 303 true rumours, 405 false rumours, and 1334 non-rumours. Three annotators labelled all messages in the dataset, and we considered a valid label if two or three annotations agreed upon it. We removed tweets with three different annotations to ensure reliable annotation for rumour classification. We divided all datasets into a ratio of 80:20 for training and testing, respectively, and used 10% of the training set as a validation set to tune the hyperparameters. After obtaining the optimal hyperparameters, we trained the model five times with these optimal hyperparameters using all training samples and reported the testing results.

5.3.2 Experimental Setup

Evaluation Metrics: To evaluate the performance of different classification models, we use precision, recall, macro-F1 score, i.e., the mean of the F1 scores for all three classes, and accuracy to evaluate the performance of different classification methods for rumour verification.

Hyperparameter: We use the pretrained English [41] and Chinese [80] BERT models, with an embedding dimension of 768 for the Twitter15, Twitter16, and CR-Twitter datasets, respectively. We trained the models for {30,50,100} epochs with a mini-batch size of 16, which is the largest multiple of 2 that fits the GPU memory in the experiment setup. We employ Adam optimizer with an initial learning rate of {1e-5, 2e-5, 5e-5} and a linear learning decay from the initial value to 0. To avoid overfitting, we applied L2 regularization with a rate of 0.0001 and a dropout rate of {0.1,0.2,0.5}. Our models were implemented in PyTorch and trained on two GeForce RTX 2080 Ti GPUs. The number of Graph Transformer Layers is set to 2, this ensures a fair comparison to other graph-based methods. The number of multi-head is empirically set to 2 for all the experiments as we observed a consistently good performance across all datasets.

5.3.3 Comparison with State-of-the-Art Methods

We compare the performance of our proposed CDGTN model, which is designed for fine-grained rumour classification, with the following state-of-the-art rumour classification methods. The quantitative results of the different methods are shown in Table 5.1 and Table 5.2. We report the mean and the standard deviation of the matrices by experimenting five times.

(1) *DTC* [81]: A decision tree-based method that ranks enquiry phrases and clusters claims for rumour detection.

(2**)** *SVM-TS* [42]: An SVM-based machine-learning algorithm that utilizes handcrafted time-series features.

(3) *RNN-GUU* [28]: An RNN-based architecture that models the sequential relationship among all claims.

(4) *TD-TvNN* [24]: A recursive neural network model based on a top-down tree structure.

(5) *BU-TvNN* [24]: A recursive neural network model based on a bottom-up tree structure.

(6) *STS-NN* [25]: A temporal-temporal neural network that models the message propagation for rumour detection.

(7) *PLAN* [82]: A self-attention network that encodes each conversation thread using a standard Transformer.

(8) *StA-PLAN* [82]: A Structure-Aware transformer network that utilizes structural information in the attention mechanism.

(9) *Bi-GCN* [27]: A bi-directional graph convolutional network that fuses the top-down and bottom-up graph structure for rumour classification.

(10) *PPA-WAE* [83]: A bi-directional graph convolutional network that fuses the top-down and bottom-up graph structure for rumour classification.

(11) *DA-GCN* [84]: An attentive graph neural network that aims to capture informative semantic and propagation features using dual-attention networks.

(12) *GACL* [85]: A method that uses the contrastive loss function to explicitly perceive the difference between conversational threads of the same class and different classes, and an Adversarial Feature Transformation module to produce conflicting samples for mining event-invariant features.

In comparison to the sequence models, including RNN-GRU [28], PLAN [82], and StA-PLAN [82], our proposed CDGTN model shows an improvement of 7% F1-score for both Twitter15 and Twitter16 datasets. his improvement can be attributed to the fact that relying solely on sequential information is not sufficient for accurate rumour verification. Additionally, CDGTN outperforms Bi-GCN [27] in terms of F1-score and accuracy by 4%, as local aggregation in a graphical pattern does not account for the overall context from the source post to a reply. With the proposed graph-Transformer network that combines local and source-guided aggregation, CDGTN achieves the best performance in terms of both accuracy and F1-score. Compared to state-of-the-art methods such as DA-GCN [84] and GACL [85], our method exhibits a 1-2% improvement in accuracy. The stacked early classification loss in our method is equivalent to node dropping during training, leading to better generalization performance by learning to classify graphs with fewer nodes.

For rumour detection on PHEME and Weibo datasets, we compare our method with the following methods.

(1) *BERT* is a pretrained language model that is used to obtain the representation of the source post for classification.

(2) *EANN* is a GAN-based model to extract event invariant features to facilitate detecting newly arrived events.

(3) *QSAN* integrates quantum-driven text encoding and a signed attention mechanism to model complex semantics between source posts and responsive posts.

(4) *RumourGAN* generates uncertain or conflicting voices to enhance the discriminator to learn stronger rumour representations.

(5) *KMGCN* uses a graph convolution network to incorporate visual information and KG to enhance the semantic representation.

(6) *DDGCN* is a dual-dynamic graph convolutional network used to model the dynamics of messages in propagation as well as the dynamics of the background knowledge from Knowledge graphs in one unified framework.

(7) *DGNF* is a dynamic news propagation network-based dynamic news propagation network for misinformation detection on social media.

Table 5.1 Streaming Graph Rumour Classification on Twitter15 Dataset.

Model	Accuracy	F1 (Macro)	F1 (NR)	F1 (FR)	F1 (TR)	F1 (UR)
DTC [81]	0.454	0.455	0.733	0.355	0.317	0.415
SVM-TS [42]	0.544	0.539	0.796	0.472	0.404	0.483
GRU-RNN [28]	0.641	0.644	0.684	0.634	0.688	0.571
BU-RvNN [24]	0.708	0.709	0.695	0.728	0.759	0.653
TD-RvNN [24]	0.723	0.729	0.682	0.758	0.821	0.654
STS-NN [25]	0.809	0.809	0.797	0.811	0.856	0.773
PLAN [82]	0.845	0.845	0.823	0.858	0.895	0.802
StA-PLAN [82]	0.852	0.852	0.840	0.846	0.884	0.837
Bi-GCN [27]	0.886	0.886	0.891	0.860	0.93	0.864
PPA-WAE[83]	0.873	0.873	0.899	0.881	0.869	0.843
DA-GCN [84]	0.905	0.905	**0.959**	0.895	0.914	0.852
GACL [85]	0.901	0.897	0.958	0.851	0.903	**0.876**
CDGTN	**0.916 (0.00302)**	**0.915 (0.00266)**	0.947 (0.00898)	**0.951 (0.00973)**	**0.912 (0.00880)**	0.859 (0.00668)

Table 5.2 Streaming Graph Rumour Classification on Twitter16 Dataset.

Model	Accuracy	F1 (Macro)	F1 (NR)	F1 (FR)	F1 (TR)	F1 (UR)
DTC [81]	0.465	0.465	0.643	0.393	0.419	0.403
SVM-TS [42]	0.574	0.568	0.755	0.420	0.571	0.526
GRU-RNN [28]	0.633	0.609	0.617	0.715	0.577	0.527
BU-RvNN [24]	0.718	0.718	0.723	0.712	0.779	0.659
TD-RvNN [24]	0.737	0.737	0.662	0.743	0.835	0.708
STS-NN [25]	0.809	0.809	0.797	0.811	0.856	0.773
PLAN [82]	0.874	0.874	0.853	0.839	0.917	0.888
StA-PLAN [82]	0.868	0.869	0.826	0.833	0.927	0.888
Bi-GCN [27]	0.880	0.880	0.847	0.869	0.937	0.865
PPA-WAE[83]	0.887	0.887	0.882	0.903	0.921	0.842
DA-GCN [84]	0.902	0.902	0.894	0.872	0.928	0.913
GACL [85]	0.920	0.917	**0.934**	0.869	**0.959**	0.907
CDGTN	**0.929** **(0.0177)**	**0.927** **(0.0181)**	0.874 (0.0250)	**0.949** **(0.0201)**	0.956 (0.0225)	**0.931** **(0.112)**

The consistently high performance of CDGTN in both datasets suggests its robustness and effectiveness in rumour detection and classification tasks. Its superior results demonstrate that CDGTN can effectively leverage the underlying temporal and textual information to capture the dynamics and context of rumours, leading to improved accuracy and precision in identifying and debunking rumours. These findings highlight the potential of CDGTN as a reliable method for rumour detection and management in online social networks.

In the results for the CR-Twitter dataset, we compared our CDGTN model with RNN, Transformer, GAT, GCN, Bi-GCN, GACL and DCNF models, which are open-sourced methods, allowing for reproducibility of results. The RNN-based network and Bi-GCN-based methods achieved the lowest accuracy and F1 score. Unlike PLAN, which is a Transformer-based model, our method combines the diffusion graph from the source to a reply with a graph-aware masked Transformer, enabling it to utilize local and source information more effectively. This allows our proposed CDGTN to obtain robust semantic relationships across the entire graph, resulting in the best performance in terms of accuracy and Macro F1.

Table 5.3 Streaming Graph Rumour Classification on the CR-Twitter Dataset.

Model	Accuracy	F1 (Macro)	F1 (NR)	F1 (FR)	F1 (TR)	F1 (UR)
RNN-GRU [28]	0.841	0.778	0.903	0.744	0.726	0.739
PLAN [82]	0.855	0.790	0.916	0.766	0.702	0.778
GNN-LSTM [86]	0.855	0.721	0.928	0.765	0.790	0.400
GAT [87]	0.871	0.825	0.922	0.772	**0.815**	0.793
GCN [27]	0.841	0.757	0.911	0.770	0.667	0.682
Bi-GCN [27]	0.867	0.808	0.919	0.797	0.734	0.783
GACL [85]	0.880	0.825	**0.935**	0.813	0.760	0.792
DCNF [88]	0.880	0.826	0.930	**0.826**	0.739	0.809
CDGTN	**0.891**	**0.858**	**0.931**	0.824	0.805	**0.873**

Furthermore, our model achieved the best F1 scores in the non-rumour, true rumour, and unverified rumour classes, while the recurrent-based model performed best in false rumour classes. This is because the CR-Twitter dataset is highly imbalanced. We also found that non-rumours could be distinguished accurately from rumours, achieving an F1 score of 0.912, whereas the F1 scores of true rumours, false rumours, and unverified rumours ranged from 0.690 to 0.846. This is because all these samples are rumours, and therefore similar, as they are unverified at the time of posting. With the help of the community response, we were able to differentiate false rumours from true rumours.

5.3.4 Ablation Study

To further analyse the effectiveness of the graph-aware causal attention mask in the Transformer and the incremental attention pooling in the proposed network, we conducted experiments by removing these two components separately and then evaluating the model under the same settings. The result is shown in Table 5.4.

The experiment results demonstrate that the model's performance drops by about 2%, in terms of accuracy and F1-score, for Twitter15, Twitter16, CR-Twitter, PHEME, and Weibo datasets without the two components. This indicates that the graph-Transformer and attention pooling can enhance the feature representation of the source-reply graph diffusion process.

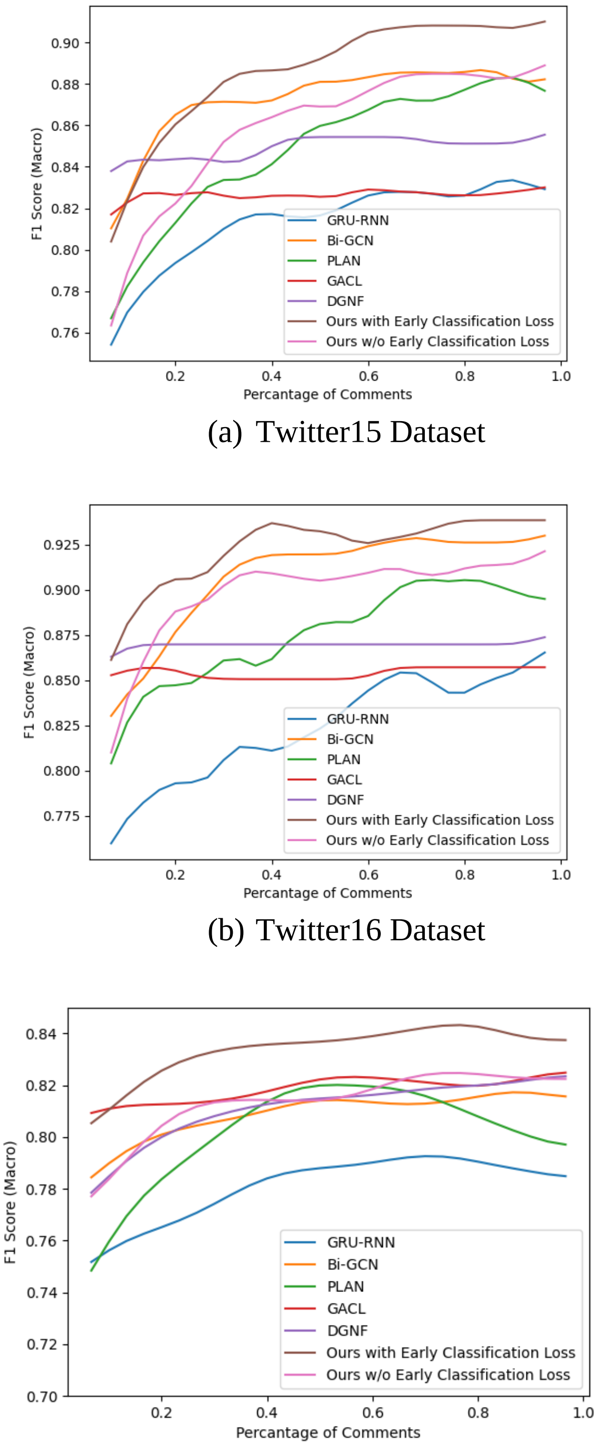
Table 5.4 Ablation study of CDGTN.

Dataset	Model	Accuracy	F1 (Macro)
Twitter15	CDGTN w/o attention mask	0.909	0.908
	CDGTN w/o SGIAP	0.882	0.881
	CDGTN (Full)	**0.912**	**0.912**
Twitter16	CDGTN w/o attention mask	0.933	0.932
	CDGTN w/o SGIAP	0.926	0.926
	CDGTN (Full)	**0.939**	**0.939**
PHEME	CDGTN w/o attention mask	0.820	0.780
	CDGTN w/o SGIAP	0.874	0.856
	CDGTN (Full)	**0.875**	**0.863**

Moreover, the results reveal that source-guided attention pooling is more crucial than the graph-transformer mask. This is because the source post is the most important claim during feature aggregation since the aim of a rumour classification system is to classify the veracity of the source information. Therefore, source-guided attention pooling can significantly improve the performance of the proposed method.

5.3.5 Results of Early Rumour Verification

Figures 5.7 (a), 5.7 (b), 5.7 (c) depict the early rumour verification results for the Twitter15, Twitter16, and CR-Twitter datasets proposed method significantly improves early detection results, with a 3% higher F1 score in the early stage of the rumour diffusion process when both components are utilized in the network. This improvement can be attributed to our method's stacked early classification loss, which enables the model to prioritize important replies during propagation as early as possible. Compared to state-of-the-art methods like GRU-RNN, PLAN, Bi-GCN, GACL, and DGNF, our model's early detection performance is notably superior.



(a) Twitter15 Dataset

(b) Twitter16 Dataset

(c) CR-Twitter Dataset

Figure 5.7 Early rumour classification results on Twitter15, Twitter16, and CR-Twitter datasets.

5.3.6 Time Complexity Analysis of Stream Verification

Fig. 5.8 illustrates the total runtime required for processing continued inference in the stream verification setting. The results demonstrate that by buffering the cumulative sum of conversations, the runtime for processing a graph in stream verification has a constant time complexity of O(1). These findings align with the theoretical analysis of our proposed CDGTN.

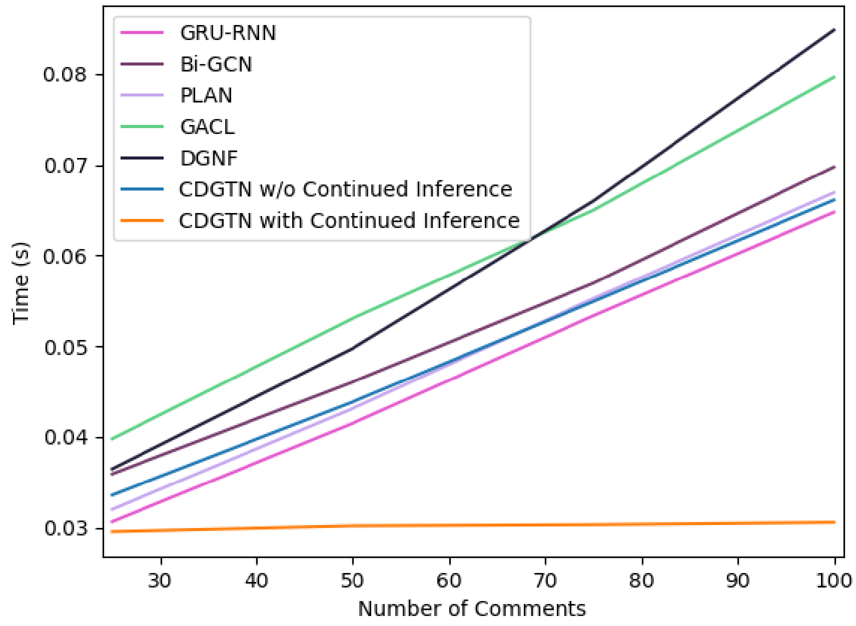


Figure 5.8 Average running time against number of replies during stream classification of rumours.

Furthermore, the runtime of the proposed CDGTN without continued inference is comparable to that of the Transformer (TNN), while Bi-GCN requires the most time. This is because Bi-GCN needs to process the node features twice using bottom-up and top-down graph neural networks.

5.3.7 Case Study of Rumour Verification

Table 5.5 presents four examples of social media conversations classified by the proposed model, including a false rumour, a true rumour, an unverified rumour, and a non-rumour. The source post and several replies are provided to demonstrate rumour classification using community responses.

In the case of the false rumour, the claim suggests a large-scale protest in Hunan, China, which has been verified as false on existing fact-checking sites. The replies express judgments and disagreement with the source information. Some replies say "It is impossible" and "See it clearly, don't spread fake news." These replies are crucial for the model to classify the veracity of a social media conversation.

Regarding the true rumour, the source claim pertains to a baseball match during a power failure, which has been proven true, as there was supporting news available after the rumour was posted on social media. The replies generally support the source message, and people did not disagree with the source statement. Therefore, our method can successfully classify it as a true rumour.

In the case of the unverified rumour, the claim is about a fire truck getting blocked during a fire in Sichuan. However, there is no supporting evidence about the fire truck getting blocked during the fire. The replies contain judgments about the source's veracity, such as "True or false?", "Is it true?", and "It is going to spread rumours again". These replies are particularly crucial for the model to classify the source claim as an unverified rumour.

Regarding the non-rumour, it is about a policy implemented by China's Group of Prevention and Control during COVID-19. The source post contains details of the implemented policy, which is reported by the news agency in China. The replies generally express positive messages, such as "Keep it up," "Excellent work in China," etc. These signals are important for the model to predict the source message as a non-rumour.

Table 5.5 Classification results on the extended CR-Twitter dataset.

Label	Message	
False Rumour	Source	Large-scale protest in Hunan, China
	Replies	Please give a news link, thank you.
		Impossible, absolutely impossible.
		You idiot. See it clearly, don't spread fake news.
		This kind of fake account is not serious.
True Rumour	Source	Playing baseball in a Power failure.
	Replies	This is so funny.
		Does everyone wear night vision goggles?
		The Uni-President Lions still played today, but Su Zhijie lost power halfway through the game.
		I laughed so hard yesterday when watching the live.
Unverified Rumour	Source	At 4 a.m., a fire broke out in a community in Yibin, Sichuan, and fire trucks came, but they couldn't get in, and the iron sheet of closure and control blocked it.
	Replies	No casualties, right?
		True or false?
		Is it true?
		It is going to spread rumours again.
Non-Rumour	Source	On August 22, Zheng Zhongwei, director of the Science and Technology Development Center of the National Health and Medical Commission and head of the Vaccine Research and Development Working Group of the Joint Prevention and Control Mechanism of the State Council, said in the CCTV "Dialogue" program that my country has officially launched the new crown vaccine on July 22. for emergency use.
	Replies	Give the vaccine to African brothers first, they need it more.
		Keep it up!
		In China, epidemic prevention and control and vaccine research and development are so excellent that people feel safe and peaceful.
		Pay tribute to the "rebels" all over the world and may the human disasters on the scene disappear.

5.4 Summary and Future Work

In this chapter, we propose a Causal Diffused Graph-Transformer Network (CDGTN), used to encode conversations on social media for rumour verification. The proposed network fuses a top-down graph neural network and a causal Transformer network, using a graph-aware attention mask. The output embedding of the node sequences is independent of future nodes and is particularly suitable for continuous verification in streaming posts. To fuse the encoded node embedding sequences, we propose a Source-Guided Incremental Attention Pooling (SGIAP), to aggregate the sequence of node features into a fixed-length representation for early rumour verification. To improve the performance of the early classification of rumours, we propose a Stacked Early Classification Loss (SecLoss), which aims to minimize the classification loss in all time instances. Then, we present an efficient and scalable streaming rumour verification framework, which aims to continuously verify rumours in streams of social media posts, by efficiently reusing the running totals of propagation features from previous timestamps. To facilitate research on low-resource language rumour verification, we annotated a Chinese rumour verification dataset based on the CR-Twitter dataset. We conducted experiments on the Twitter15, Twitter16, and CR-Twitter datasets. In the experiments of early rumour verification, our proposed method can consistently improve performance when the number of replies is reduced. Moreover, the computational complexity of stream classification is constant when a new reply is added to the graph. In our future work, we aim to perform evidence-based rumour verification based on authorized sources.

Chapter 6

Dual Evidence for Claim Veracity Assessment

6.1 Motivation

In propagation-based approaches, many efforts have been made to improve the effectiveness and robustness of propagation-based rumour classification algorithms. However, the performance of current propagation-based veracity assessment methods is still limited, especially when there are only a small number of replies available in a propagation graph. The community response is insufficient to identify the falsity of the claim. However, the web-retrieved results contain important evidence to refute the source claim. Therefore, seeking evidence outside social media is essential for trust-worthy claim veracity assessment.

In addition to propagation-based rumour verification, external evidence-based fact-checking has become another paradigm attracting the research community in recent years. It is the task of evaluating the veracity of claims, which can be made in written or spoken language [4]. Evidence retrieval is the first step in a fact-checking process, and is used to find relevant sources that support or refute the claim. To address this problem, Dougrez-Lewis et al. [39] proposed external evidence-based rumour verification on social media. However, they ignored the community response in the propagation graph, which has been proven to be an effective feature for rumour verification. To address this, this chapter aims to simultaneously utilize the community response and external evidence for effective claim veracity assessment. Inspired by multimodal neural networks, we propose a dual-stream cross-attention network to enhance the feature presentation, by leveraging the cross-correlation between the social response and external evidence, for claim veracity assessment.

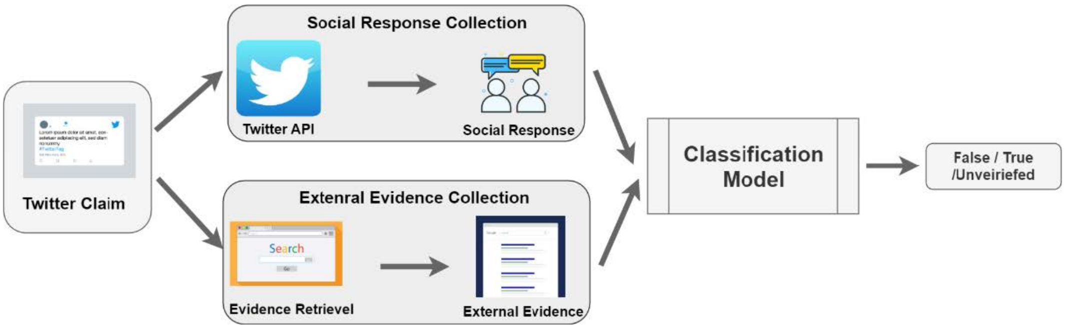


Figure 6.1 The pipeline of dual evidence approach for automatic claim veracity assessment.

The contributions of our work are summarized as follows:

- We propose integrating external evidence with the community response for effective rumour verification on social media claims. To achieve this, we propose a Dual-Stream Cross-Attention Network (DSCAN) to learn informative and correlative features extracted in social response and external evidence, through a dual attention mechanism.
- We extend two publicly available datasets, namely PHEME and RumourEval, for rumour verification on social media by collecting external evidence using three search engines, including Google Search, Bing Search, and DuckDuckGo Search, which are then used in our experiments.
- We demonstrate the effectiveness of using hybrid features of social response and external evidence for rumour verification. Experimental results show that our proposed DSCAN achieves state-of-the-art performance for rumour verification on social media, evaluated on the extended PHEME and RumourEval datasets.

6.2 Dual-Stream Cross-Attention Network (DSCAN)

In this section, we first introduce the proposed framework for rumour verification using social and external evidence. Our goal is to classify a source media claim into three categories (unverified, false, true). The overall architecture of the proposed network is shown in Fig. 6.1. The proposed rumour verification framework contains four components, a text embedding module, an intra-evidence attention module, a cross-evidence attention module, and a classification module. The text embedding module is to transform a source message, a set of social replies, and a set of external evidence into fixed-length representations. After that, we adopt a dual-stream intra-evidence attention network to model the community response and external evidence, respectively. Then, we propose using the cross-attention mechanism to learn the semantic relationship between social response and external evidence. Finally, the Softmax classifier is employed for evaluating the veracity of the input claims.

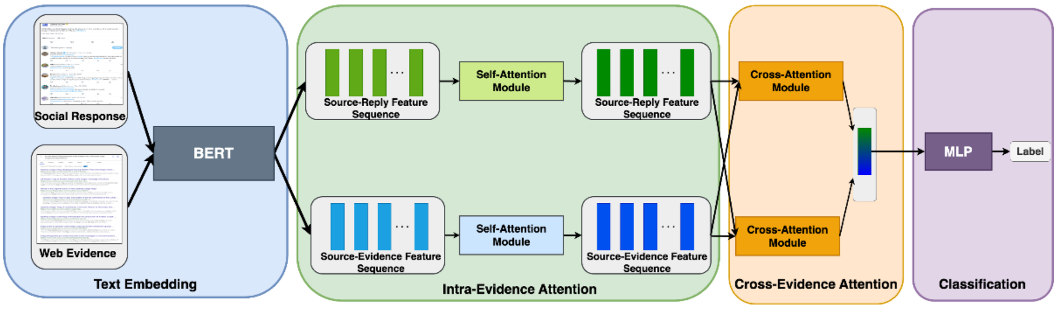


Figure 6.2 The proposed Dual-Stream Cross-Attention Network (DSCAN).

6.2.1 Text Embedding Module

The text embedding module is responsible for converting the input text, which can be a source claim, a social reply, or an external web result, into a fixed-length representation. Each text message is encoded using a pretrained bidirectional Transformer (BERT) to obtain a d-dimensional vector. Specifically, we utilize the hidden representation of the first token by adding a special learnable embedding vector at the beginning of each text before passing it to the pretrained BERT. This process forms the overall representation of the text. The source claim feature is concatenated with the reply sequence to create a source-reply sequence, denoted as R. Similarly, the source claim feature is concatenated with the external evidence sequence to create a source-external sequence, denoted as E.

6.2.2 Intra-Evidence Attention Module

The intra-evidence attention module is a crucial component of our framework, consisting of a two-stream attention network. Each stream utilizes the self-attention mechanism to focus on features within its respective stream, enhancing the representation power. To achieve this, we incorporate multi-head attention modules, which consist of h heads of scaled dot-product attention blocks. These attention blocks take three inputs: query $\boldsymbol{\mathcal{Q}}$, key $\boldsymbol{\mathcal{K}}$, and value $\boldsymbol{\mathcal{V}}$.

Once the attention score matrix $\boldsymbol{S}$ is obtained, we compute the attention value matrix $\boldsymbol{\mathcal{V}}'$ by taking the linear combination of the attention score $\boldsymbol{S}$ and the value matrix $\boldsymbol{V}$. Therefore, we have

$$\boldsymbol{S} = \mathrm{Softmax}(\frac{\boldsymbol{\mathcal{Q}}\boldsymbol{\mathcal{K}}^{\mathrm{T}}}{\sqrt{d_k}}), \tag{6.1}$$

where d_k is the embedding dimension of the key $\boldsymbol{\mathcal{K}}$. After calculating the attention score matrix $\boldsymbol{S}$, the attention value $\boldsymbol{\mathcal{V}}'$ can be computed as a linear combination of the attention score $\boldsymbol{S}$ and the key $\boldsymbol{\mathcal{K}}$. Therefore, we have

$$\boldsymbol{\mathcal{V}}' = \boldsymbol{S}\boldsymbol{\mathcal{V}}, \tag{6.2}$$

The multi-head attention mechanism repeats the scaled dot-product attention h times and aggregates the attention values $\boldsymbol{\mathcal{V}}'$ to obtain a more robust representation of the value $\boldsymbol{\mathcal{V}}$, as follows:

$$\mathrm{MultiHead}(\boldsymbol{\mathcal{Q}}, \boldsymbol{\mathcal{K}}, \boldsymbol{\mathcal{V}}) = \mathrm{Concat}(\boldsymbol{V}_1'\boldsymbol{W}_1', \boldsymbol{V}_2'\boldsymbol{W}_2', \dots, \boldsymbol{V}_h'\boldsymbol{W}_h')\boldsymbol{W}^o, \tag{6.3}$$

where $\mathrm{MultiHead}(\,)$ and $\mathrm{Concat}(\,)$ represent the multi-head attention mechanism and the concatenation operation, respectively. It is worth noting that $\boldsymbol{W^{o}}$ and $\boldsymbol{W_i'}$ are trainable parameters jointly learned through backpropagation. Finally, the intra-evidence attention module uses skip connections to add the output of the multi-head attention to the value $\boldsymbol{\mathcal{V}}$, as follows:

$$\mathrm{IntraAttention}(\boldsymbol{F}) = \mathrm{LayerNorm}(\mathrm{MultiHead}(\boldsymbol{F}, \boldsymbol{F}, \boldsymbol{F}) + \boldsymbol{F}), \tag{7.4}$$

where $\boldsymbol{F}$ denotes the input feature sequence, which can be either the source-reply feature or the source-evidence feature. We feed the source-reply feature $\boldsymbol{R}$ and the source-evidence feature $\boldsymbol{E}$ to the respective two streams of the intra-evidence attention module. We denote the output of the social response stream as $\widehat{\boldsymbol{R}} = \mathrm{IntraAttention}(\boldsymbol{R})$. Similarly, the output of the external evidence stream is denoted as $\widehat{\boldsymbol{E}} = \mathrm{IntraAttention}(\boldsymbol{E})$.

6.2.3 Cross-Evidence Attention Module

The Cross-Evidence Attention Module incorporates the cross-attention mechanism to compute attention vectors between the query feature in one stream and all key features in the other stream. This enables us to capture the semantic relationship between the social response and external evidence, facilitating a comprehensive assessment of claim veracity.

To measure the alignment between the social response and external evidence, we consider the first element in the social response stream, denoted as $\hat{\boldsymbol{r}}_{\mathbf{0}}$, as the query vector. The correlation scores $\boldsymbol{a_{re}}$ between $\hat{\boldsymbol{r}}_{\mathbf{0}}$ and the enhanced external evidence features $\widehat{\boldsymbol{E}}$ are calculated using the dot product, and divided by the square root of the embedding dimension d_k, as shown in Equation (5):

$$\boldsymbol{a_{re}} = \frac{\hat{\boldsymbol{r}}_{\mathbf{0}}\widehat{\boldsymbol{R}}^{\boldsymbol{T}}}{\sqrt{d_k}}, \tag{6.5}$$

$$\boldsymbol{f_{re}} = \mathrm{Softmax}(\boldsymbol{a_{re}})\boldsymbol{R} + \hat{\boldsymbol{r}}_{\mathbf{0}}, \tag{6.6}$$

The output of the social reply-guided external evidence feature is denoted as $\boldsymbol{f_{re}}$. Similarly, we can compute the external evidence-guided social feature as follows:

$$\boldsymbol{a_{er}} = \frac{\hat{e}_0 \widehat{E}^T}{\sqrt{d_k}}, \tag{6.7}$$

$$\boldsymbol{f_{er}} = \mathrm{Softmax}(\boldsymbol{a_{er}})\boldsymbol{E} + \boldsymbol{\hat{e}_0}, \tag{6.8}$$

where $\boldsymbol{\hat{e}_0}$ represents the first element in the enhanced external evidence features. We concatenate the output features of the two streams, i.e., $\boldsymbol{f_{re}}$ and $\boldsymbol{f_{er}}$, to form the final representation, denoted as $\boldsymbol{f_{cls}}$, which is forwarded to the classification layer.

6.2.4 Classification Layer and Loss Function

We use the Softmax classifier to classify the veracity of rumour claims, as follows:

$$\boldsymbol{\hat{y}} = \mathrm{softmax}(\boldsymbol{W_c f_{cls}} + \boldsymbol{b_c}), \tag{6.9}$$

where $\boldsymbol{W_c} \in \mathbb{R}^{3\times d}$ and $\boldsymbol{b_c} \in \mathbb{R}^3$ are trainable parameters. We employ the cross-entropy loss as the objective function in our proposed method. Given the predicted label $\boldsymbol{\hat{y}}$ and the ground-truth label $\boldsymbol{y}$, the negative log-likelihood is minimized. Thus, we have

$$\mathrm{loss} = -\sum_{i=1}^{3} \mathrm{y}_j \log(\hat{y}_j), \tag{6.10}$$

6.3 Experimental Setups and Results

6.3.1 Data Datasets and Evidence Collection

In the experiments, we use two benchmark datasets, including PHEME [89] and RumourEval [90], to evaluate the methods for rumour veracity assessment. The PHEME dataset contains 1067 true rumours, 638 false rumours, and 697 unverified rumours. The RumourEval dataset contains 145 true rumours, 74 false rumours, and 106 unverified rumours. The PHEME dataset contains conversations from 5 real events. Following previous works, we conduct leave-one-event-out crossvalidation on the PHEME dataset. RumourEval has already been split into training and test sets for a fair comparison.

For evidence collection, we use the replies collected from the original datasets as internal evidence. To collect external evidence, we use the Google Programmable Search API, Bing Search API, and DuckDuckGo Search engine to retrieve relevant web pages. To achieve this, we first process the source claims, by removing user mentions, punctuations, emojis, and external URLs. Then, we take the processed claims as query text to the search engines. Finally, we use the list of text snippets, returned by the search engines, as external evidence to verify the query-text claims.

6.3.2 Experimental Setup

Evaluation Metrics: In our experiments, we evaluate different rumour verification methods using Precision, Recall, Macro-F1 score, i.e., the mean of F1 scores for all three classes, and accuracy.

Hyperparameter: We use the pretrained English BERT [41] as the text embedding module. The embedding dimension of the BERT model is 768. For all experiments, the models were trained for 30 epochs with a mini-batch size of 32. We use the Adam optimiser with a learning rate of 0.00002. To avoid overfitting, we use L2 regularization with a rate of 0.001 and a dropout rate of 0.1. All experiments were conducted on two GeForce RTX 2080 Ti GPUs.

6.3.3 Comparison to Other Methods

We compare our proposed model, i.e., DSCAN, with the following state-of-the-art rumour-verification methods. The quantitative results of the different methods are shown in Table 6.1. BrnachLSTM [89] adopts an LSTM-based architecture to model the sequential branches in each thread. (2) TD-TvNN, proposed by Ma et al [24], is a recursive neural network model based on a top-down tree structure. (3) GCN-RNN [91] is a conversational GCN for veracity prediction. (4) HiTPLAN [82] is a recently proposed self-attention network, which encodes each conversation thread using a standard Transformer. (5) Hierarchical Transformer [92] combines local context with BERT and global context with Transformer. (6) VAED [93] is a Variational Autoencoder (VAE) for rumour verification using community response. (7) SAVED [93] is an enhanced version of VAED with stance classifiers and a topic-learning module. (8) VRoc [94] is an LSTM-based Variational Autoencoder with multi-task learning, which consists of a rumour detector, a rumour tracker, a stance classifier, and a veracity classifier. (9) Roberta+Evidence [39] is an external evidence-based rumour verification method, which adopts Natural Language Inference (NLI) to learn the relationship between a rumour and the retrieved evidence.

Table 6.1 shows that the proposed DSCAN significantly outperforms all previous propagation-based methods. Compared to the previous state-of-the-art rumour verification model *VRoC*, the proposed DSCAN achieves 5.90% and 5.80% improvement, in terms of macro-F1 and accuracy, respectively, when evaluated on the PHEME dataset. On the RumourEval dataset, our method outperforms the previous state-of-the-art methods correspondingly by 5.40% and 3.60%. This is mainly because the proposed DSCAN utilizes external evidence. After all, the community response is insufficient to classify the veracity of out-of-context claims. Compared to the Roberta with evidence method, which adopts external evidence for rumour verification, our method achieves a 14% higher F1 score. This is because the Roberta with evidence method does not utilize social context features, i.e., the community response in a propagation graph. The difference is that our method leverages the cross-correlation between social context and external evidence, which is particularly important for providing sufficient context for rumour verification on social media. Moreover, our method aggregates retrieved web search results with three different search engines, rather than using one search engine. We will explore the importance of cross-evidence attention and the choice of search engines in ablation studies.

Table 6.1 Comparison to other state-of-the-art methods on the RumourEval and PHEME datasets.

Dataset	Model	Macro-F1	Accuracy
RumourEval	BranchLSTM [89]	0.491	0.500
	TD-RVNN [24]	0.509	0.536
	Hierarchical GCN-RNN [91]	0.540	0.536
	HiTPLAN [82]	0.581	0.571
	Hierarchical Transformer [92]	0.592	0.607
	DSCAN (ours)	**0.646**	**0.643**
PHEME	BranchLSTM [89]	0.259	0.314
	TD-RVNN [24]	0.264	0.341
	Hierarchical GCN-RNN [91]	0.317	0.356
	HiTPLAN [82]	0.361	0.438
	Hierarchical Transformer [92]	0.372	0.441
	VAED [93]	0.362	0.380
	SAVED [93]	0.434	0.528
	VRoC [94]	0.484	0.521
	Roberta + Evidence [39]	0.405	-
	DSCAN (ours)	**0.543**	**0.579**

6.3.4 Ablation Studies

The results of the ablation studies are shown in Tables 6.2, 6.3, and 6.4. In Table 6.2, we compare the results using inputs without the propagation graph or external evidence. The results show that both the propagation graph and external evidence are important for our proposed method to achieve promising performance. For the proposed methods without the propagation graph, the Macro-F1 score drops by 10.7% and 4.9% on the RumourEval and PHEME datasets, respectively. After removing external evidence, the Macro-F1 score drops by 15.1% and 5.6% on the RumourEval and PHEME datasets, respectively. This reflects that external evidence is slightly more important than community response because using community response alone is insufficient to accurately classify the veracity of out-of-context claims.

Table 6.2 Classification results with different input features.

Dataset	Model	Macro-F1	Accuracy
RumourEval	DSCAN w/o Social Replies	0.571	0.582
	DSCAN w/o External Evidence	0.495	0.536
	DSCAN (full)	**0.646**	**0.643**
PHEME	DSCAN w/o Social Replies	0.494	0.518
	DSCAN w/o External Evidence	0.487	0.481
	DSCAN (full)	**0.543**	**0.579**

In Table 6.3, we test the performance of the proposed method using evidence retrieved from only one search engine each time. On the RumourEval dataset, the performance of using only the DuckDuckGo search engine is slightly better than using either Google Search or Bing Search. Alternatively, in terms of the Macro-F1 score, methods using the results from Google Search achieve the best performance when evaluated on the PHEME dataset. The overall results show that performance, in terms of Macro F1 score, is the best when all web search results retrieved by the three search engines are used. However, we also see that the performance does not drop significantly, i.e., the Macro-F1 score drops by only 1-3%, when the web search results from only one search engine are used for evaluation.

To understand the effectiveness of using the self-attention and cross-evidence attention modules, we remove the two fusion modules separately. When the self-attention modules are removed, the Marco F1 score drops by 6.8% and 4.5%, evaluated on the RumourEval and PHEME datasets, respectively. Furthermore, the Marco F1 score drops by 3.6% and 3.50%, evaluated on the RumourEval and PHEME datasets, respectively. This means that the self-attention modules are more important than the cross-evidence attention module. One possible reason is that the correlation between each item in the evidence is much higher than that of the cross-evidence, as the items come from the same sources.

Table 6.3 Results with different search engines.

Dataset	Search Engine	Macro-F1	Accuracy
RumourEval	Google Search only	0.613	0.607
	Bing Search only	0.611	0.607
	DuckDuckGo Search only	0.641	0.643
	Aggregated Search Engine Results	**0.646**	**0.643**
PHEME	Google Search	0.541	0.585
	Bing Search	0.537	0.568
	DuckDuckGo Search	0.531	0.565
	Aggregated Search Engine Results	**0.543**	**0.579**

Table 6.4 Results with different fusion strategies.

Dataset	Model	Macro-F1	Accuracy
RumourEval	DSCAN w/o intra-attention	0.578	0.571
	DSCAN w/o cross-attention	0.610	0.607
	DSCAN (full)	**0.646**	**0.643**
PHEME	DSCAN w/o intra-attention	0.498	0.497
	DSCAN w/o cross-attention	0.513	0.520
	DSCAN (full)	**0.543**	**0.579**

6.3.5 Case Study

Table 6.5 shows three examples of social media rumours, with veracity provided in the PHEME datasets. The source post, some replies, and external evidence are included to illustrate rumour classification using community responses and external evidence.

In Table 6.5, the false rumour claimed that an ISIS flag was displayed in the window of a besieged cafe in Sydney’s Martin Place during a siege event. The replies are not sufficient to indicate whether it is a true or a false claim. With the help of external evidence, some relevant news articles were retrieved and these articles pointed out that the flag captured during Sydney’s siege was initially mistaken by many for an Isis flag.

Table 6.5 Examples of false rumours, true rumours, and unverified rumours with the corresponding social replies and external evidence.

Label	Text	
False Rumour	Source	An ISIS flag is being displayed in the window of a café under siege in Sydney’s Martin Place.
	Social Response	holy lord! Prayers to all. This is too much.
		So Islamists are no threat to Australia, right? As you were…
		"Omg, I feel sick! “@USER: UPDATE: An ISIS flag is being displayed in the window.
	External Evidence	The flag displayed during the siege at a Sydney cafe is not the same one used by the Islamic State terrorist group.
		The black flag with white writing hung in the window was initially mistaken by many for an Isis flag.
		Siege makes global headlines. The flag shown being held by hostages against the window of Lindt Chocolat Cafe is not an Islamic State flag but an Islamic flag that has been co-opted.
True Rumour	Source	Germanwings Airbus A320 crashes in French Alps
	Social Response	Things like this make me scared to travel
		Very sad German plane crashed
		Ang scary :(I know plane crashes happen a lot in a year but with social media and easy info, we hear about it more often
	External Evidence	Germanwings A320 aircraft flying from Barcelona to Düsseldorf goes down in southern French Alps with 150 on board ... German Airbus A320 plane crashes in French Alps.
		A Germanwings plane carrying 150 people has crashed in the French Alps on its way from Barcelona to Duesseldorf. The Airbus A320 - flight 4U 9525 - went down between Digne and Barcelonnette.
		An Airbus A320 with 144 passengers and 6 crew members has crashed in Digne region.
Unverified Rumour	Source	The Charlie Hebdo attack was carefully planned. and then you leave your ID card in the getaway car?
	Social Response	do criminals go to commit their crimes with IDcard in their pockets
		Losing your ID card is very common during a terrorist attack, here is another example
		are they sure it's the RIGHT ID card?
		They need to get with the bum bags
	External Evidence	"Charlie Hebdo shooting, a series of terrorist attacks that shook France in January 2015, claiming the lives of 17 people, including 11 journalists and security personnel at the Paris offices of Charlie Hebdo, a satiric magazine.
		"AP. On the morning of Friday 9 January, the manhunt entered its final phase as police closed in on the Charlie Hebdo attack suspects at Dammartin-en-Goele
		7 Jan 2015. AFP/Getty Images. Gunmen burst into the offices of the satirical French magazine Charlie Hebdo today, killing 12 people and wounding at least 40 more.
		A survivor of the Charlie Hebdo killings has told a French court of the trauma she has suffered since she was forced at gunpoint to let two attackers into the magazine's office

For the true rumour, the source claim is about a Germanwings airbus A320 reportedly crashed in the Digne region. The social replies tend to trust the news, and external web search results include relevant news articles that support the social claim on social media. Therefore, it can be detected as a true rumour.

For the unverified rumour, it claimed that the Charlie Hebdo attack was obviously well-planned, and then the ID card was left. The social replies judge the veracity of the claim, but the external evidence does not show any evidence to support or refute the claim. Therefore, the rumour remains unverified.

6.4 Summary

In this chapter, we propose to integrate external evidence for propagation-based rumour verification. We propose the Dual-Stream Cross-Attention Network (DSCAN) to fuse social and web evidence. The proposed DSCAN consists of two streams of self-attention modules, followed by cross-attention modules. The two-stream self-attention module can learn the intra-relationship within the evidence, while the cross-attention is designed to learn the relationship between community responses and external evidence. To facilitate future research on evidence-based rumour verification on social media, we extend two publicly available datasets, namely PHEME and RumourEval. We collected web search results from three web search engines using social media claims as queries. Our experimental results show that the dual evidence-based approach outperforms existing single evidence-based methods, for assessing the veracity of claims on social media. In future work, we aim to study uncertainty-based rumour verification, which aims to measure the confidence scores of a rumour verification system, with both internal and external evidence.

Chapter 7

Instruction-Following Language Models with External Knowledge for Automatic Fact-Checking

7.1 Motivation

With the recent rise in popularity of large language models (LLMs) [77], [95] in natural language processing (NLP) tasks, such as machine translation, text classification, and data extraction, they have also been used for fake news detection [96].

Despite the impressive capabilities of these LLMs, a significant limitation of these language models is their reliance on pre-existing knowledge, which may not always be up-to-date. In the context of fact-checking, the reliance of language models solely on their internal knowledge raises concerns about their ability to accurately assess the veracity of claims, especially when faced with rapidly evolving information [34]. To address this limitation, it becomes imperative to consider external knowledge sources that provide updated and reliable information in recent fact-checking algorithms [34], [97]–[99].

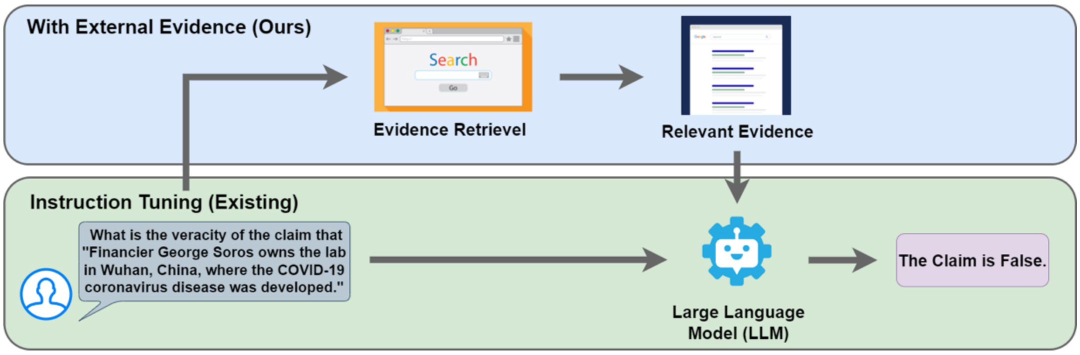

With External Evidence (Ours)
Search
Evidence Retrievel
Relevant Evidence
Instruction Tuning (Existing)
What is the veracity of the claim that "Financier George Soros owns the lab in Wuhan, China, where the COVID-19 coronavirus disease was developed."
Large Language Model (LLM)
The Claim is False.

Figure 7.1 The pipeline of the proposed fact-check framework using large language model with evidence retrieval.

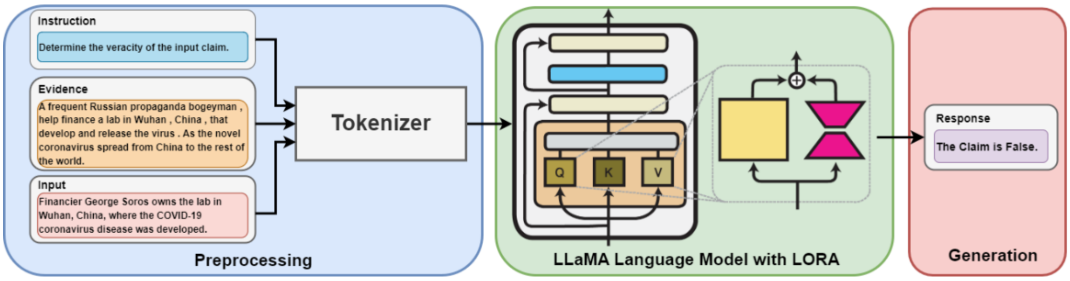
This chapter aims to enhance the fact-checking capabilities of instruction-following language models by leveraging external evidence. We propose a method that combines the power of pretrained language models with the retrieval of relevant external evidence from search engines. By integrating this external evidence during the instruct-tuning process, we aim to augment the knowledge of the language model, enabling it to make more accurate predictions The contributions of our work are summarized as follows:

- Introducing instruct-tuned language models for fact-checking: We propose the application of instruct-tuned language models, i.e., LLaMA, for automatic fact-checking tasks, expanding their scope beyond language generation.
- Addressing the limitation of instruct-tuned models: We identify the limitation of instruct-tuned language models in fact-checking due to outdated knowledge and propose the integration of external evidence to enhance their accuracy and reliability.
- Proposing a method for incorporating external knowledge: We present a novel approach that combines pretrained language models with external evidence retrieval from search engines, augmenting the knowledge base for fact-checking.

- Achieving state-of-the-art performance: Through experiments on the RAWFC and LIAR datasets, we demonstrate that our method achieves state-of-the-art performance in fact-checking tasks.

7.2 Instruction-Tuning Large Language Models with External Evidence for Automatic Fact-Checking

In this study, we instruct-tune the pretrained LLaMA [100] model using the LORA algorithm [72]. Our approach not only takes the text claim as input for factual classification but also the retrieval evidence.



Determine the veracity of the input claim.

A frequent Russian propaganda bogeyman , help finance a lab in Wuhan , China , that develop and release the virus . As the novel coronavirus spread from China to the rest of the world.

Financier George Soros owns the lab in Wuhan, China, where the COVID-19 coronavirus disease was developed.

The Claim is False.

Figure 7.2 Instruct-tuning with external evidence for automated fact-checking.

7.2.1 Optimization of Language Model

During instruct-tuning, we aim to optimize our LLaMA model's parameters θ to minimize a loss function that measures the difference between the predicted fact-check results and the ground truth of the training dataset. Suppose we have a set of input-output pairs (x, y) in the training set, where x is the instruction-evidence-input to be inputted to the LLaMA model, and y is the corresponding label for that claim. Suppose $f(x; \theta)$ denotes the output (i.e., predicted veracity of the input claim) of the LLaMA with parameters θ for input x. We define the loss function as follows:

$$L(\theta) = \sum(y - f(x; \theta))^2. \tag{7.1}$$

7.2.2 LORA Tuning

Our goal is to find the optimal values for the parameters of the LLaMA model θ that minimize the loss function $L(\theta)$. To achieve this, we leverage the LORA algorithm, which involves a low-rank approximation of the parameter matrix θ. As a result, it reduces the number of trainable parameters and improves generalization while preserving most of the information contained in the original parameters. Specifically, LORA compresses θ into a low-rank matrix product:

$$\Omega = UV, \tag{7.2}$$

where U and V are matrices of lower rank than θ. The LORA algorithm updates the parameters θ by adding a regularized factorization of θ to the existing parameters as follows:

$$\theta' = UV, \tag{7.3}$$

$$\theta' = \theta + (\Omega - \theta), \tag{7.4}$$

where θ' is the low-rank parameters to be updated through back-propagation.

7.3 Experimental Setups and Results

7.3.1 Datasets

We use two publicly available datasets, namely RAWFC [98], and LIAR [101]datasets for evaluating our proposed approaches. The RAWFC dataset is a collection of claims and evidence related to factual verification tasks, consisting of 2,012 claims with supporting evidence, labelled either true, false or half-true, from Snopes.com. The LIAR dataset is a dataset of political statements fact-checked by PolitiFact.com, consisting of 12,836 short statements, each labelled as true, mostly-true, half-true, barely true, false, or pants-on-fire. For fair comparison to other methods, we use the same split released in [98].

7.3.2 Experimental Setup

We evaluate our proposed approach for automatic fact-checking with explanations using three standard evaluation metrics: Precision Scores, Recall Scores, and F1 Scores. To instruct-tune the LLaMA model, we trained the models for 3 epochs with a mini-batch size of 32. We employ an Adam optimizer with an initial learning rate of 1e-4 and a linear learning decay from the initial value to 0. To avoid overfitting, we applied a dropout rate of 0.05. Our models were implemented in PyTorch [102] and HuggingFace [103].

7.3.3 Comparison to Other Methods

Table 7.1 and Table 7.2 provide the evaluation results of various methods on two different fact-checking datasets, i.e., RAWFC and LIAR. The methods are compared based on precision, recall, and F1-score, which are commonly used metrics to assess the performance of classification tasks. The baselines that we compared include *SVM* [104], which utilizes bag-of-words features for fake news detection. *CNN* [101] incorporates metadata features to enhance representation learning. *RNN* [105]learns representations from word sequences without relying on external resources. *DeClarE* [106] combines word embeddings from the claim, report, and source to assess the credibility of the claim. *dEFEND* [97]employs a GRU-based model for veracity prediction, providing explanations. *SentHAN* [34] represents each sentence based on coherence and semantic conflicts with the claim. SBERT-FC [34] utilizes *SentenceBERT* (SBERT) for encoding and identifying fake news based on the top-ranked sentences. *GenFE* [99] and *GenFE-MT* [99] detect fake news independently or jointly with explanations in a multi-task setup. *CofCED* [98] is a Coarse-to-fine Cascaded Evidence-Distillation neural network for explainable fake news detection based on such raw reports, alleviating the dependency on fact-checked ones.

Table 7.1 Results of external-evidence-based fact checking on RAWFC dataset.

Methods	Precision	Recall	F1-Score
SVM [104]	0.3233	0.3251	0.3171
CNN [101]	0.3880	0.3850	0.3859
RNN [105]	0.4135	0.4209	0.4039
DeClarE [106]	0.4339	0.4352	0.4218
dEFEND [97]	0.4493	0.4326	0.4407
sentHAN [107]	0.4566	0.4554	0.4425
SBERT-FC [34]	0.5106	0.4592	0.4551
GenFE [99]	0.4429	0.4474	0.4443
GenFE-MT [99]	0.4564	0.4527	0.4508
CofCED [98]	0.5299	0.5099	0.5107
FactLLaMA (Ours)	**0.5611**	**0.5550**	**0.5565**

From Table 7.1 for the RAWFC dataset, it can be observed that traditional machine learning methods like SVM, CNN, and RNN achieve moderate results in terms of precision, recall, and F1-score. However, more advanced models such as DeClarE, dEFEND, sentHAN, SBERT-FC, GenFE, GenFE-MT, and CofCED outperform the traditional methods, particularly CofCED, which achieves the highest F1-score of 0.5107.

Table 7.2 Results of external-evidence-based fact checking on LIAR dataset.

Methods	Precision	Recall	F1-Score
SVM [104]	0.1578	0.1592	0.1534
CNN [101]	0.2258	0.2239	0.2136
RNN [105]	0.2436	0.2120	0.2079
DeClarE [106]	0.2286	0.2055	0.1843
dEFEND [97]	0.2309	0.1856	0.1751
sentHAN [107]	0.2264	0.1996	0.1846
SBERT-FC [34]	0.2409	0.2207	0.2219
GenFE [99]	0.2801	0.2616	0.2649
GenFE-MT [99]	0.1855	0.1990	0.1515
CofCED [98]	0.2948	0.2955	0.2893
FactLLaMA (Ours)	**0.3246**	**0.3205**	**0.3044**

Interestingly, LLaMA without tuning, i.e., zero-shot prediction, performs relatively poorly compared to the other methods. However, when Instruct-Tuning is applied, there is a significant performance improvement, particularly when external knowledge is incorporated. Instruct Tuned LLaMA with External Knowledge achieves the highest F1-score of 0.5565, surpassing all other methods and demonstrating the effectiveness of leveraging external evidence.

On the evaluation of the LIAR dataset, as shown in Table 8.2, similar patterns can be observed. Traditional machine learning methods show relatively low performance, while more advanced models exhibit better results. CofCED achieves the highest F1-score of 0.2893, indicating its effectiveness in fact-checking on the LIAR dataset.

Once again, LLaMA without tuning performs poorly, but Instruct-Tuning leads to substantial improvements. Incorporating external knowledge in the Instruct-Tuning process further enhances the performance, with LLaMA with Instruct-Tuning and External Knowledge achieving the highest F1-score of 0.3044.

In summary, the evaluation results from both datasets highlight the superiority of advanced models over traditional machine learning methods in fact-checking tasks. The Instruct-Tuning approach, particularly when combined with external knowledge, consistently outperforms other methods, showcasing the value of leveraging external evidence for accurate fact-checking. These findings emphasize the importance of staying updated with the latest information and leveraging advanced techniques to combat the spread of misinformation effectively.

7.3.4 Ablation Study

Table 7.3 Ablation study on FactLLaMA using RAWFC and LIAR datasets.

Dataset	Model	Precision	Recall	F1
RAWFC	LLaMA (w/o tuning)	0.3350	0.3255	0.2643
	FactLLaMA (Instruct-tuning w/o external knowledge)	0.5376	05400	0.5376
	FactLLaMA (Instruct-tuning with external knowledge)	**0.5611**	**0.5550**	**0.5565**
LIAR	LLaMA (w/o tuning)	0.1587	0.2069	0.1224
	FactLLaMA (Instruct-tuning w/o external knowledge)	0.3232	0.3157	0.2998
	FactLLaMA (Instruct-tuning with external knowledge)	**0.3246**	**0.3205**	**0.3044**

The table presents the evaluation results of different models on two datasets, RAWFC and LIAR, in terms of precision, recall, and F1 score. These metrics are commonly used to assess the performance of classification models, indicating the model's ability to correctly classify instances of different classes.

For the RAWFC dataset, the LLaMA model without tuning achieved a precision of 0.3350, recall of 0.3255, and F1 score of 0.2643. These results demonstrate the baseline performance of the LLaMA model on the dataset without any tuning or additional external knowledge. However, the FactLLaMA model, which incorporates instruct-tuning without external knowledge, showed significant improvement, with a precision of 0.5376, recall of 0.5400, and F1 score of 0.5376. This indicates that instruct-tuning alone enhances the model's ability to classify instances more accurately.

Furthermore, when external knowledge is incorporated into the FactLLaMA model through instruct-tuning, the precision increased to 0.5611, recall to 0.5550, and F1 score to 0.5565. These results highlight the positive impact of incorporating external knowledge, such as information retrieved from search engines, in improving the model's performance and enhancing its ability to classify instances effectively.

On the LIAR dataset, the LLaMA model without tuning achieved a lower precision of 0.1587, recall of 0.2069, and F1 score of 0.1224. Similar to the RAWFC dataset, the FactLLaMA model showed improvements, with a precision of 0.3232, recall of 0.3157, and F1 score of 0.2998 when instruct-tuning was applied without external knowledge. Incorporating external knowledge in the instruct-tuning process resulted in a precision of 0.3246, recall of 0.3205, and F1 score of 0.3044. These results indicate that instruct-tuning, with or without external knowledge, enhances the performance of the model on the LIAR dataset.

Overall, the evaluation results demonstrate the effectiveness of the FactLLaMA model, particularly when instruct-tuning is applied and external knowledge is incorporated. These findings highlight the importance of leveraging external information and fine-tuning strategies to improve the accuracy and performance of fact-checking models in classifying and verifying the veracity of statements or claims.

7.3.5 Confusion Matrices Evaluated on RAWFC and LIAR Datasets

Figures 7.2 and 7.3 present the confusion matrices for the RAWFC and LIAR datasets, respectively. The rows and columns in the figures represent the ground truth and predictions, respectively.

In Figure 7.2, it is evident that the model can effectively distinguish between the TRUE and FALSE labels. However, classifying the HALF-TRUE label proves to be more challenging for the model. This difficulty arises because both the HALF-TRUE and FALSE labels contain misinformation, albeit with differing degrees of accuracy. Moving to Figure 8.3, we observe that the model shows clear classification performance for the TRUE and PLANT-FIRES classes compared to the other classes. However, it struggles to accurately classify the barely-true, half-true, and mostly-true classes. This difficulty arises from the fact that items in these classes contain a mixture of true and false information, making it a subjective task for both humans and machines to classify them accurately without specialized expertise.

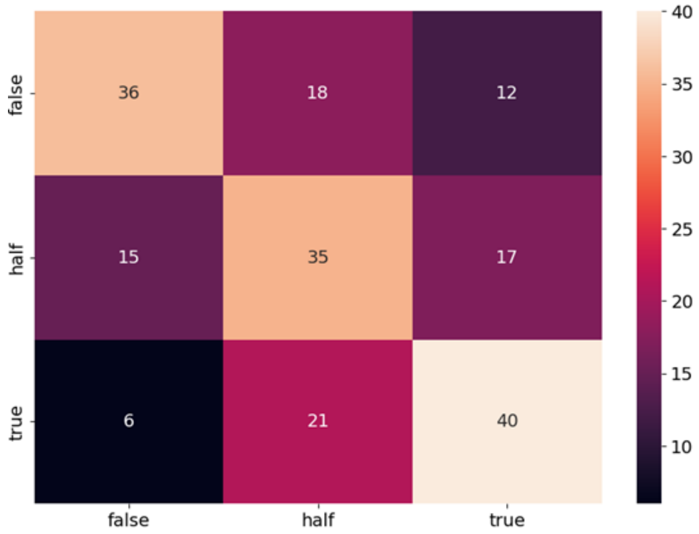


Figure 7.3 Confusion matrix on RAWFC dataset

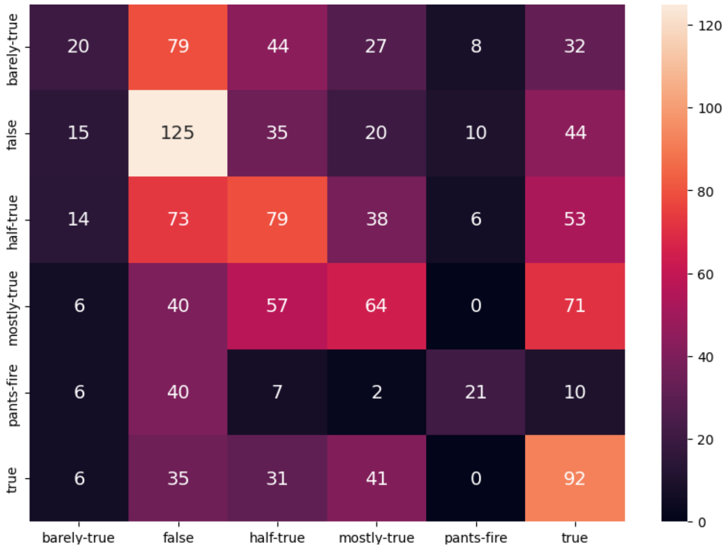
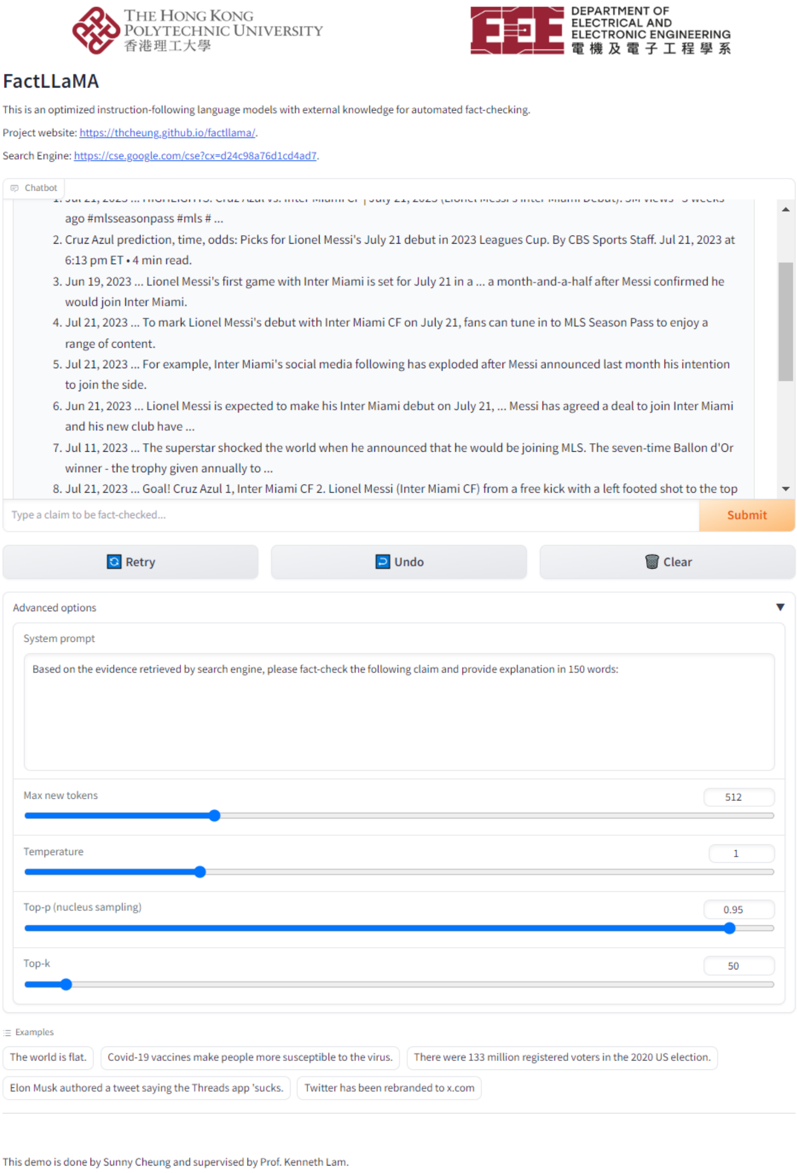


Figure 7.4 Confusion matrix on LIAR dataset.

7.4 A Web Chatbot System for Assisting Fact-Checking

A web chatbot system, based on the proposed model has been developed for general to perform fact-checking. A user may input a claim to be fact-checked for asking the veracity of it. Then, the system will automatically search through the Internet to find the up-to-date information for assessing the veracity of the claim. Fig. 7.5 shows the demonstration of the proposed web chatbot system.


The Hong Kong Polytechnic University
香港理工大學
Department of Electrical and Electronic Engineering
電機及電子工程學系

FactLLaMA

This is an optimized instruction-following language models with external knowledge for automated fact-checking.

Project website: https://thcheung.github.io/factllama/.

Search Engine: https://cse.google.com/cse?cx=d24c98a76d1cd4ad7.

ago #mlsseasonpass #mls # ...
2. Cruz Azul prediction, time, odds: Picks for Lionel Messi's July 21 debut in 2023 Leagues Cup. By CBS Sports Staff. Jul 21, 2023 at 6:13 pm ET • 4 min read.
3. Jun 19, 2023 ... Lionel Messi's first game with Inter Miami is set for July 21 in a ... a month-and-a-half after Messi confirmed he would join Inter Miami.
4. Jul 21, 2023 ... To mark Lionel Messi's debut with Inter Miami CF on July 21, fans can tune in to MLS Season Pass to enjoy a range of content.
5. Jul 21, 2023 ... For example, Inter Miami's social media following has exploded after Messi announced last month his intention to join the side.
6. Jun 21, 2023 ... Lionel Messi is expected to make his Inter Miami debut on July 21, ... Messi has agreed a deal to join Inter Miami and his new club have ...
7. Jul 11, 2023 ... The superstar shocked the world when he announced that he would be joining MLS. The seven-time Ballon d'Or winner - the trophy given annually to ...
8. Jul 21, 2023 ... Goal! Cruz Azul 1, Inter Miami CF 2. Lionel Messi (Inter Miami CF) from a free kick with a left footed shot to the top

System prompt

Based on the evidence retrieved by search engine, please fact-check the following claim and provide explanation in 150 words:

Max new tokens 512

Temperature 1

Top-p (nucleus sampling) 0.95

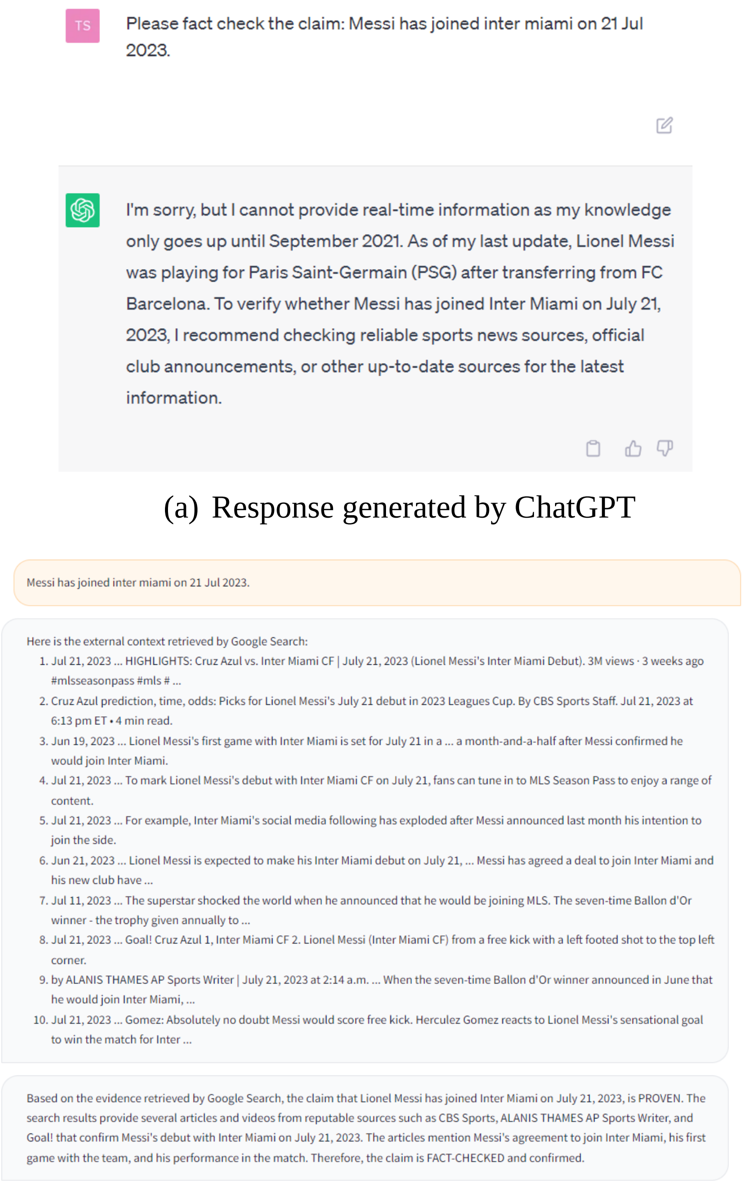
Top-k 50

Examples
- The world is flat.
- Covid-19 vaccines make people more susceptible to the virus.
- There were 133 million registered voters in the 2020 US election.
- Elon Musk authored a tweet saying the Threads app 'sucks.
- Twitter has been rebranded to x.com

This demo is done by Sunny Cheung and supervised by Prof. Kenneth Lam.

Figure 7.5 A Web chatbot system for automatic fact-checking.

Fig. 7.6 shows the comparison between ChatGPT and the proposed FactLLaMA. We compare the response generated with news appeared after the cutoff date of the training corpus of two models. This ensures that the training corpus of the models do not contain any information of the future event to be tested.



Please fact check the claim: Messi has joined inter miami on 21 Jul 2023.

I'm sorry, but I cannot provide real-time information as my knowledge only goes up until September 2021. As of my last update, Lionel Messi was playing for Paris Saint-Germain (PSG) after transferring from FC Barcelona. To verify whether Messi has joined Inter Miami on July 21, 2023, I recommend checking reliable sports news sources, official club announcements, or other up-to-date sources for the latest information.

(a) Response generated by ChatGPT

Messi has joined inter miami on 21 Jul 2023.

Here is the external context retrieved by Google Search:

1. Jul 21, 2023 ... HIGHLIGHTS: Cruz Azul vs. Inter Miami CF | July 21, 2023 (Lionel Messi's Inter Miami Debut). 3M views · 3 weeks ago #mlsseasonpass #mls # ...
2. Cruz Azul prediction, time, odds: Picks for Lionel Messi's July 21 debut in 2023 Leagues Cup. By CBS Sports Staff. Jul 21, 2023 at 6:13 pm ET • 4 min read.
3. Jun 19, 2023 ... Lionel Messi's first game with Inter Miami is set for July 21 in a ... a month-and-a-half after Messi confirmed he would join Inter Miami.
4. Jul 21, 2023 ... To mark Lionel Messi's debut with Inter Miami CF on July 21, fans can tune in to MLS Season Pass to enjoy a range of content.
5. Jul 21, 2023 ... For example, Inter Miami's social media following has exploded after Messi announced last month his intention to join the side.
6. Jun 21, 2023 ... Lionel Messi is expected to make his Inter Miami debut on July 21, ... Messi has agreed a deal to join Inter Miami and his new club have ...
7. Jul 11, 2023 ... The superstar shocked the world when he announced that he would be joining MLS. The seven-time Ballon d'Or winner - the trophy given annually to ...
8. Jul 21, 2023 ... Goal! Cruz Azul 1, Inter Miami CF 2. Lionel Messi (Inter Miami CF) from a free kick with a left footed shot to the top left corner.
9. by ALANIS THAMES AP Sports Writer | July 21, 2023 at 2:14 a.m. ... When the seven-time Ballon d'Or winner announced in June that he would join Inter Miami, ...
10. Jul 21, 2023 ... Gomez: Absolutely no doubt Messi would score free kick. Herculez Gomez reacts to Lionel Messi's sensational goal to win the match for Inter ...

Based on the evidence retrieved by Google Search, the claim that Lionel Messi has joined Inter Miami on July 21, 2023, is PROVEN. The search results provide several articles and videos from reputable sources such as CBS Sports, ALANIS THAMES AP Sports Writer, and Goal! that confirm Messi's debut with Inter Miami on July 21, 2023. The articles mention Messi's agreement to join Inter Miami, his first game with the team, and his performance in the match. Therefore, the claim is FACT-CHECKED and confirmed.

(b) Response generated by the proposed system

Figure 7.6 Comparison between ChatGPT and the proposed system.

The example shown in Figure 7.6 is the latest news about “Messi has joined Inter Miami on July 21, 2023”. The results show that ChatGPT does not have access to latest information while the proposed method does, by utilizing search engines to retrieve update-to-date information. These research findings show that the integration of external retrieval system is essential to the problem of automatic fact-checking.

7.5 Summary

In conclusion, this research highlights the crucial role of automatic fact-checking in combating the spread of misinformation online. While large language models (LLMs) and Instruction-Following variants like InstructGPT and Alpaca have demonstrated remarkable performance in various natural language processing tasks, their potential lack of up-to-date knowledge can lead to inaccuracies in fact-checking. To address this limitation, we proposed a method that combines pretrained language models with external evidence retrieval, resulting in enhanced fact-checking accuracy. By leveraging search engines to retrieve relevant evidence for a given claim, we successfully augmented the knowledge of the pretrained language model. Through instruct-tuning an open-source language model called LLaMA, with this external evidence, we achieved more accurate predictions regarding the veracity of input claims. Experimental evaluations on widely used fact-checking datasets, RAWFC and LIAR, showcased that our approach achieved state-of-the-art performance in fact-checking tasks. The integration of external evidence effectively bridged the knowledge gap between the model and the most up-to-date information available, leading to improved fact-checking outcomes. We believe our research has significant implications for combatting misinformation and promoting the dissemination of accurate information on online platforms. In our future work, we plan to generate explanations with these pretrained language models for more general use.

Chapter 8

Conclusion and Future Works

8.1 Summary of the Study

This thesis has presented a comprehensive investigation into rumour claim detection and claim veracity assessment on social media platforms. The study addressed the pressing need to combat misinformation, rumours, and fake news, by proposing novel approaches to enhance the accuracy and reliability of rumour detection and claim verification. The research considered various techniques, including multimodal image-text classification, author-aware rumour detection, propagation graph-based analysis, and external evidence-based claim veracity assessment.

8.2 Key Findings and Contributions

The findings of this study highlight several important contributions to the field of rumour detection and claim veracity assessment on social media platforms. First, the proposed CBAN demonstrated the effectiveness of leveraging both the correlative and complementary relationships between textual and visual information to improve the accuracy of rumour detection. By considering subtle cues and contextual information embedded in images, the proposed CBAN achieved superior performance compared to the state-of-the-art methods for multimodal classification on social media.

Second, the author-aware rumour detection approach sheds light on the significance of analysing user description and account features in identifying rumours. By differentiating between genuine users and spam accounts, the method greatly enhanced the precision of rumour detection and mitigated the influence of malicious actors spreading false information.

Third, the propagation graph-based analysis approach provided valuable insights into the structure and dynamics of rumour propagation. By modelling the community response in a streaming manner, the proposed CDGTN identified the rumours and facilitated the verification of claims, based on the propagation patterns observed in the graph more effectively and efficiently.

Lastly, the external evidence-based claim veracity assessment approach demonstrated the potential of leveraging external sources, such as search engines, to retrieve relevant evidence for assessing the veracity of claims. By comparing the information from social media with external sources, the method established a more comprehensive and reliable assessment of claim veracity.

8.3 Comparison of the Proposed Methods

The major advantages and disadvantages of the four methods related to early rumour detection and claim veracity assessment on social media are as follows:

Multimodal Image-Text Rumour Detection:

Advantages:

- Integration of textual and visual information allows for a more comprehensive understanding of rumours.
- Deep learning-based approach can capture complex patterns and features in text and images.

- Potential to detect rumours that may be more difficult to identify using only text-based analysis.

Disadvantages:

- Requires a reliable dataset with labeled textual and visual content for training the model.
- Dependency on the availability and quality of images associated with social media posts.
- Computational resources and processing time required for training and inference may be high due to the multimodal nature of the approach.

Simultaneous Rumour and Malicious Account Detection:

Advantages:

- Simultaneous detection of rumours and malicious accounts provides a more holistic approach to combating misinformation.
- Consideration of user behaviour patterns helps in identifying accounts involved in spreading rumours.
- Deep learning techniques can capture complex relationships between textual content, user behaviour, and malicious intent.

Disadvantages:

- Identification of malicious accounts may require access to additional metadata or account-level information, which may not always be available.
- Overreliance on behaviour patterns may result in false positives or false negatives.
- The challenge of keeping up with evolving techniques employed by malicious actors.

Propagation Graph-Based Claim Veracity on Social Media:

Advantages:

- Graph-based approach captures the spread and propagation patterns of claims, providing valuable insights into their veracity.
- Incorporation of network and source credibility information enhances the assessment of claim veracity.
- Can identify influential sources or clusters within the network that contribute to the spread of rumours.

Disadvantages:

- Constructing accurate and comprehensive propagation graphs can be challenging, particularly for large-scale social media datasets.
- Reliance on the availability and accuracy of network and source credibility information.
- Difficulty in accounting for dynamic network structures and real-time updates of information.

Claim Veracity Assessment using Social and External Evidence:

Advantages:

- Integration of social and external evidence provides a broader perspective for assessing claim veracity.
- Access to fact-checking reports, expert opinions, and additional sources of information improves the reliability of assessments.
- Can consider user engagement metrics, such as likes, shares, and comments, to gauge the credibility of claims.

Disadvantages:

- Reliability and trustworthiness of external sources and fact-checking reports may vary.
- The challenge of effectively integrating diverse sources of evidence and weighing their importance.
- Difficulty in real-time retrieval and analysis of external evidence.

Overall, these methods offer innovative approaches to early rumour detection and claim veracity assessment on social media. However, they also face challenges related to data availability, computational requirements, the dynamic nature of social media platforms, and the reliability of external sources. Addressing these limitations and continuously adapting the methods to evolving social media landscapes is crucial for their successful application.

8.4 Implications and Future Directions

The findings of this study have significant implications for researchers, practitioners, and policymakers working in the field of social media and misinformation. In order to enhance the effectiveness of identifying and combating rumours and misinformation, it is crucial to incorporate misinformation detection into decentralized social media platforms. The proposed approaches presented in this study contribute to the development of effective tools and techniques in this regard.

In future work, by implementing decentralized social media platforms with robust misinformation detection capabilities, we can improve the accuracy of rumour detection and claim verification. This, in turn, will help mitigate the harmful consequences of false information on individuals, communities, and society as a whole.

However, it is important to acknowledge the limitations of this study. The proposed approaches were evaluated on specific datasets and social media platforms, which may not fully capture the diversity and complexity of rumour detection and claim veracity assessment across different contexts. Therefore, future research should aim to expand the scope of evaluation and consider the generalizability of the proposed approaches across various platforms and contexts.

Additionally, given the evolving nature of social media platforms and the continuous development of new techniques for spreading rumours and misinformation, ongoing research and adaptation are crucial. Future studies should explore emerging trends, such as deep learning techniques, knowledge graph integration, and real-time monitoring, to further improve the effectiveness of rumour detection and claim veracity assessment.

Moreover, it is imperative to address ethical considerations in future research. Privacy protection and user profiling should be carefully taken into account to ensure the responsible use of data. The development of transparent and accountable algorithms is essential in order to adequately address the ethical implications associated with these approaches.

In conclusion, this study lays the foundation for integrating misinformation detection into decentralized social media platforms. Future research should strive to overcome the limitations, explore emerging trends, and address ethical considerations to advance the field and effectively combat the spread of rumours and misinformation on social media.

8.5 Summary

In conclusion, this study has made significant contributions to the field of rumour claim detection and claim veracity assessment on social media platforms. The proposed approaches, including multimodal image-text classification, author-aware rumour detection, propagation graph-based analysis, and external evidence-based claim veracity assessment, have demonstrated their effectiveness in enhancing the accuracy and reliability of detecting rumours and assessing claim veracity.

By addressing the challenges posed by misinformation and fake news, this research contributes to the advancement of techniques that can have a positive impact on individuals, society, and the information ecosystem as a whole. As the field continues to evolve, future research should build upon these findings, considering emerging trends and ethical considerations, to further improve the detection and mitigation of rumours and misinformation on social media platforms.

References

[1] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, and S. Liu, “Fake news on Social Media: the Impact on Society,” *Information Systems Frontiers,* 2022, doi: 10.1007/s10796-022-10242-z.

[2] X. Zhang and A. A. Ghorbani, “An overview of online fake news: Characterization, detection, and discussion,” *Inf Process Manag,* vol. 57, no. 2, 2020, doi: 10.1016/j.ipm.2019.03.004.

[3] S. Ghosh and C. Shah, “Towards automatic fake news classification,” *Proceedings of the Association for Information Science and Technology,* vol. 55, no. 1, 2018, doi: 10.1002/pra2.2018.14505501125.

[4] Z. Guo, M. Schlichtkrull, and A. Vlachos, “A Survey on Automated Fact-Checking,” *Trans Assoc Comput Linguist,* vol. 10, 2022, doi: 10.1162/tacl_a_00454.

[5] D. M. J. Lazer *et al.*, “The science of fake news: Addressing fake news requires a multidisciplinary effort,” *Science (1979),* vol. 359, no. 6380, 2018, doi: 10.1126/science.aao2998.

[6] D. Rohera *et al.*, “A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects,” *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3159651.

[7] F. Torabi Asr and M. Taboada, “Big Data and quality data for fake news and misinformation detection,” *Big Data Soc*, vol. 6, no. 1, 2019, doi: 10.1177/2053951719843310.

[8] U. Mertoğlu and B. Genç, “Automated fake news detection in the age of digital libraries,” *Information Technology and Libraries*, vol. 39, no. 4, 2020, doi: 10.6017/ITAL.V39I4.12483.

[9] Á. Escolà-Gascón, N. Dagnall, A. Denovan, K. Drinkwater, and M. Diez-Bosch, “Who falls for fake news? Psychological and clinical profiling evidence of fake news consumers,” *Pers Individ Dif*, vol. 200, 2023, doi: 10.1016/j.paid.2022.111893.

[10] S. C. Rhodes, “Filter Bubbles, Echo Chambers, and Fake News: How Social Media Conditions Individuals to Be Less Critical of Political Misinformation,” *Polit Commun*, vol. 39, no. 1, 2022, doi: 10.1080/10584609.2021.1910887.

[11] X. Zhou and R. Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities,” *ACM Comput Surv*, vol. 53, no. 5, 2020, doi: 10.1145/3395046.

[12] A. A. A. A. Et al., “Detecting Fake News using Machine Learning: A Systematic Literature Review,” *Psychology and Education Journal*, vol. 58, no. 1, 2021, doi: 10.17762/pae.v58i1.1046.

[13] V. K. Singh, I. Ghosh, and D. Sonagara, “Detecting fake news stories via multimodal analysis,” *J Assoc Inf Sci Technol*, vol. 72, no. 1, 2021, doi: 10.1002/asi.24359.

[14] W. Xu, J. Wu, Q. Liu, S. Wu, and L. Wang, “Evidence-aware Fake News Detection with Graph Neural Networks,” in *WWW 2022 - Proceedings of the ACM Web Conference 2022*, 2022. doi: 10.1145/3485447.3512122.

[15] P. Dhiman, A. Kaur, C. Iwendi, and S. K. Mohan, “A Scientometric Analysis of Deep Learning Approaches for Detecting Fake News,” *Electronics (Switzerland),* vol. 12, no. 4. 2023. doi: 10.3390/electronics12040948.

[16] Y. Wang *et al.*, “EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA: ACM, Jul. 2018, pp. 849–857. doi: 10.1145/3219819.3219903.

[17] T. Zhang *et al.*, “BDANN: BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jul. 2020, pp. 1–8. doi: 10.1109/IJCNN48605.2020.9206973.

[18] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, “Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs,” in *Proceedings of the 25th ACM international conference on Multimedia*, New York, NY, USA: ACM, Oct. 2017, pp. 795–816. doi: 10.1145/3123266.3123454.

[19] X. Zhou, J. Wu, and R. Zafarani, “SAFE: Similarity-Aware Multi-modal Fake News Detection,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2020. doi: 10.1007/978-3-030-47436-2_27.

[20] L. Ying, H. Yu, J. Wang, Y. Ji, and S. Qian, “Multi-Level Multi-Modal Cross-Attention Network for Fake News Detection,” *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3114093.

[21] Y. Chen *et al.*, “Cross-modal Ambiguity Learning for Multimodal Fake News Detection,” in *Proceedings of the ACM Web Conference 2022*, New

York, NY, USA: ACM, Apr. 2022, pp. 2897–2905. doi: 10.1145/3485447.3511968.

[22] Y. Liu and S. Xu, “Detecting Rumors Through Modeling Information Propagation Networks in a Social Media Environment,” *IEEE Trans Comput Soc Syst*, vol. 3, no. 2, pp. 46–62, Jun. 2016, doi: 10.1109/TCSS.2016.2612980.

[23] A. Zubiaga, M. Liakata, and R. Procter, “Exploiting Context for Rumour Detection in Social Media,” in *International Conference on Social Informatics*, 2017, pp. 109–123. doi: 10.1007/978-3-319-67217-5_8.

[24] J. Ma, W. Gao, and K.-F. Wong, “Rumor Detection on Twitter with Tree-structured Recursive Neural Networks,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 1980–1989. doi: 10.18653/v1/P18-1184.

[25] Q. Huang, C. Zhou, J. Wu, L. Liu, and B. Wang, “Deep spatial–temporal structure learning for rumor detection on Twitter,” *Neural Comput Appl*, Aug. 2020, doi: 10.1007/s00521-020-05236-4.

[26] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, “A Convolutional Approach for Misinformation Identification,” in *IJCAI International Joint Conference on Artificial Intelligence*, 2017. doi: 10.24963/ijcai.2017/545.

[27] T. Bian *et al.*, “Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, pp. 549–556, Apr. 2020, doi: 10.1609/aaai.v34i01.5393.

[28] J. Ma *et al.*, “Detecting rumors from microblogs with recurrent neural networks,” in *IJCAI International Joint Conference on Artificial Intelligence*, 2016.

[29] J. Ma, J. Li, W. Gao, Y. Yang, and K. F. Wong, “Improving Rumor Detection by Promoting Information Campaigns with Transformer-based Generative Adversarial Learning,” *IEEE Trans Knowl Data Eng*, 2021, doi: 10.1109/TKDE.2021.3112497.

[30] D. T. Vu and J. J. Jung, “Rumor detection by propagation embedding based on graph convolutional network,” *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, 2021, doi: 10.2991/ijcis.d.210304.002.

[31] C. Song, K. Shu, and B. Wu, “Temporally evolving graph neural network for fake news detection,” *Inf Process Manag*, vol. 58, no. 6, 2021, doi: 10.1016/j.ipm.2021.102712.

[32] R. Trivedi, M. Farajtabar, P. Biswal, and H. Zha, “Dyrep: Learning representations over dynamic graphs,” in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[33] Y. Wang, S. Qian, J. Hu, Q. Fang, and C. Xu, “Fake News Detection via Knowledge-driven Multimodal Graph Convolutional Networks,” in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, New York, NY, USA: ACM, Jun. 2020, pp. 540–547. doi: 10.1145/3372278.3390713.

[34] N. Kotonya and F. Toni, “Explainable automated fact-checking for public health claims,” in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020. doi: 10.18653/v1/2020.emnlp-main.623.

[35] I. DAGAN, B. DOLAN, B. MAGNINI, and D. ROTH, “Recognizing textual entailment: Rational, evaluation and approaches,” *Nat Lang Eng*, vol. 15, no. 4, pp. i–xvii, Oct. 2009, doi: 10.1017/S1351324909990209.

[36] W. Ferreira and A. Vlachos, “Emergent: a novel data-set for stance classification,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 1163–1168. doi: 10.18653/v1/N16-1138.

[37] A. Hanselowski, C. Stab, C. Schulz, Z. Li, and I. Gurevych, “A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking,” in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 493–503. doi: 10.18653/v1/K19-1046.

[38] 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*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 3441–3460. doi: 10.18653/v1/2020.findings-emnlp.309.

[39] J. Dougrez-Lewis, E. Kochkina, M. Arana-Catania, M. Liakata, and Y. He, “PHEMEPlus: Enriching Social Media Rumour Verification with External Evidence,” in *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 49–58. doi: 10.18653/v1/2022.fever-1.6.

[40] A. Dosovitskiy *et al.*, “AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE,” in *ICLR 2021 - 9th International Conference on Learning Representations*, 2021.

[41] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference,* 2019.

[42] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect Rumors Using Time Series of Social Context Information on Microblogging Websites," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management,* New York, NY, USA: ACM, Oct. 2015, pp. 1751–1754. doi: 10.1145/2806416.2806607.

[43] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A Convolutional Approach for Misinformation Identification," in *IJCAI International Joint Conference on Artificial Intelligence,* 2017. doi: 10.24963/ijcai.2017/545.

[44] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019,* 2019. doi: 10.4000/books.aaccademia.4577.

[45] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal Variational Autoencoder for Fake News Detection," in *The World Wide Web Conference,* New York, NY, USA: ACM, May 2019, pp. 2915–2921. doi: 10.1145/3308558.3313552.

[46] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-Aware Multi-modal Fake News Detection," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2020. doi: 10.1007/978-3-030-47436-2_27.

[47] L. Ying, H. Yu, J. Wang, Y. Ji, and S. Qian, "Multi-Level Multi-Modal Cross-Attention Network for Fake News Detection," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3114093.

[48] N. Xu, W. Mao, and G. Chen, "A co-memory network for multimodal sentiment analysis," in *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, 2018. doi: 10.1145/3209978.3210093.

[49] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in Twitter with hierarchical fusion model," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020. doi: 10.18653/v1/p19-1239.

[50] F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal twitter datasets from natural disasters," in *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, 2018. doi: 10.1609/icwsm.v12i1.14983.

[51] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," in *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, 2020. doi: 10.1109/WACV45572.2020.9093414.

[52] N. Xu and W. Mao, "MultiSentiNet: A deep semantic network for multimodal sentiment analysis," in *International Conference on Information and Knowledge Management, Proceedings*, 2017. doi: 10.1145/3132847.3133142.

[53] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Trans Multimedia*, vol. 23, 2021, doi: 10.1109/TMM.2020.3035277.

[54] T. Jiang, J. Wang, Z. Liu, and Y. Ling, “Fusion-Extraction Network for Multimodal Sentiment Analysis,” 2020, pp. 785–797. doi: 10.1007/978-3-030-47436-2_59.

[55] H. Tan and M. Bansal, “LXMert: Learning cross-modality encoder representations from transformers,” in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference,* 2019. doi: 10.18653/v1/d19-1514.

[56] J. Lu, D. Batra, D. Parikh, and S. Lee, “ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” in *Advances in Neural Information Processing Systems,* 2019.

[57] X. Wang, X. Sun, T. Yang, and H. Wang, “Building a Bridge: A Method for Image-Text Sarcasm Detection Without Pretraining on Image-Text Data,” 2020. doi: 10.18653/v1/2020.nlpbt-1.3.

[58] R. Pranesh, “Exploring Multimodal Features and Fusion Strategies for Analyzing Disaster Tweets,” in *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022),* 2022.

[59] P. Krawczuk, S. Nagarkar, and E. Deelman, “CrisisFlow: Multimodal representation learning workflow for crisis computing,” in *Proceedings - IEEE 17th International Conference on eScience, eScience 2021,* 2021. doi: 10.1109/eScience51609.2021.00052.

[60] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017,* 2017. doi: 10.1609/icwsm.v11i1.14955.

[61] D. Du, Y. Ji, L. Li, S. Ren, J. Cheng, and Y. Zheng, "BELSTM: Understanding the Transformer and Bidirectional Long Short-Term Memory for Early Rumor Detection," in *ACM International Conference Proceeding Series,* 2020. doi: 10.1145/3443467.3443836.

[62] E. Min *et al.*, "Divide-and-Conquer: Post-User Interaction Network for Fake News Detection on Social Media," in *Proceedings of the ACM Web Conference 2022,* New York, NY, USA: ACM, Apr. 2022, pp. 1148–1158. doi: 10.1145/3485447.3512163.

[63] Z. Huang, Z. Lv, X. Han, B. Li, M. Lu, and D. Li, "Social Bot-Aware Graph Neural Network for Early Rumor Detection," *Coling 2022,* no. November, 2022.

[64] Q. Li, Q. Zhang, and L. Si, "Rumor Detection by Exploiting User Credibility Information, Attention and Multi-task Learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 1173–1179. doi: 10.18653/v1/P19-1113.

[65] K. Pelrine, J. Danovitch, and R. Rabbany, "The surprising performance of simple baselines for misinformation detection," in *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021*, 2021. doi: 10.1145/3442381.3450111.

[66] X. Chen, H. Wang, L. Ke, Z. Lu, H. Su, and X. Chen, "Identifying Cantonese rumors with discriminative feature integration in online social networks," *Expert Syst Appl*, vol. 215, 2023, doi: 10.1016/j.eswa.2022.119347.

[67] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*, 2021, pp. 5583–5594.

[68] O. Levy *et al.*, “RoBERTa: An optimized method for pretraining self-supervised NLP systems,” *arXiv*, 2020.

[69] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.

[70] P. He, X. Liu, J. Gao, and W. Chen, “DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION,” in *International Conference on Learning Representations*,

[71] N. Houlsby *et al.*, “Parameter-efficient transfer learning for NLP,” in *36th International Conference on Machine Learning, ICML 2019*, 2019.

[72] E. Hu *et al.*, “LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS,” in *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.

[73] B. Lester, R. Al-Rfou, and N. Constant, “The Power of Scale for Parameter-Efficient Prompt Tuning,” in *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2021. doi: 10.18653/v1/2021.emnlp-main.243.

[74] F. Jenhani, M. S. Gouider, and L. Ben Said, “Streaming social media data analysis for events extraction and warehousing using hadoop and storm: Drug abuse case study,” in *Procedia Computer Science*, 2019. doi: 10.1016/j.procs.2019.09.316.

[75] C. Song, C. Yang, H. Chen, C. Tu, Z. Liu, and M. Sun, “CED: Credible Early Detection of Social Media Rumors,” *IEEE Trans Knowl Data Eng*, vol. 33, no. 8, pp. 3035–3047, Aug. 2021, doi: 10.1109/TKDE.2019.2961675.

[76] C. Li *et al.*, “Joint Stance and Rumor Detection in Hierarchical Heterogeneous Graph,” *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 6, 2022, doi: 10.1109/TNNLS.2021.3114027.

[77] T. B. Brown *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, 2020.

[78] J. Ma, W. Gao, and K.-F. Wong, “Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 708–717. doi: 10.18653/v1/P17-1066.

[79] L. Ke, X. Chen, Z. Lu, H. Su, and H. Wang, “A Novel Approach for Cantonese Rumor Detection based on Deep Neural Network,” in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, Oct. 2020, pp. 1610–1615. doi: 10.1109/SMC42975.2020.9283056.

[80] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, “Pre-Training with Whole Word Masking for Chinese BERT,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 29, 2021, doi: 10.1109/TASLP.2021.3124365.

[81] Z. Zhao, P. Resnick, and Q. Mei, “Enquiring minds: Early detection of rumors in social media from enquiry posts,” in *WWW 2015 - Proceedings of the 24th International Conference on World Wide Web*, 2015. doi: 10.1145/2736277.2741637.

[82] L. M. S. Khoo, H. L. Chieu, Z. Qian, and J. Jiang, “Interpretable Rumor Detection in Microblogs by Attending to User Interactions,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8783–8790, Apr. 2020, doi: 10.1609/aaai.v34i05.6405.

[83] P. Zhang, H. Ran, C. Jia, X. Li, and X. Han, “A lightweight propagation path aggregating network with neural topic model for rumor detection,” *Neurocomputing,* vol. 458, pp. 468–477, Oct. 2021, doi: 10.1016/j.neucom.2021.06.062.

[84] X. Liu, Z. Zhao, Y. Zhang, C. Liu, and F. Yang, “Social Network Rumor Detection Method Combining Dual-Attention Mechanism With Graph Convolutional Network,” *IEEE Trans Comput Soc Syst*, pp. 1–12, 2022, doi: 10.1109/TCSS.2022.3184745.

[85] T. Sun, Z. Qian, S. Dong, P. Li, and Q. Zhu, “Rumor Detection on Social Media with Graph Adversarial Contrastive Learning,” in *Proceedings of the ACM Web Conference 2022,* New York, NY, USA: ACM, Apr. 2022, pp. 2789–2797. doi: 10.1145/3485447.3511999.

[86] D. T. Vu and J. J. Jung, “Rumor Detection by Propagation Embedding Based on Graph Convolutional Network,” *International Journal of Computational Intelligence Systems,* vol. 14, no. 1, p. 1053, 2021, doi: 10.2991/ijcis.d.210304.002.

[87] X. Yang, H. Ma, and M. Wang, “Rumor Detection with Bidirectional Graph Attention Networks,” *Security and Communication Networks,* vol. 2022, pp. 1–13, Jan. 2022, doi: 10.1155/2022/4840997.

[88] C. Song, Y. Teng, Y. Zhu, S. Wei, and B. Wu, “Dynamic graph neural network for fake news detection,” *Neurocomputing,* vol. 505, pp. 362–374, Sep. 2022, doi: 10.1016/j.neucom.2022.07.057.

[89] E. Kochkina, M. Liakata, and A. Zubiaga, “All-in-one: Multi-task learning for rumour verification,” in *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings,* 2018.

[90] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, and A. Zubiaga, “SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours,” in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017),* Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 69–76. doi: 10.18653/v1/S17-2006.

[91] P. Wei, N. Xu, and W. Mao, “Modeling Conversation Structure and Temporal Dynamics for Jointly Predicting Rumor Stance and Veracity,” 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)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4786–4797. doi: 10.18653/v1/D19-1485.

[92] J. Yu, J. Jiang, L. M. S. Khoo, H. L. Chieu, and R. Xia, “Coupled Hierarchical Transformer for Stance-Aware Rumor Verification in Social Media Conversations,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 1392–1401. doi: 10.18653/v1/2020.emnlp-main.108.

[93] J. Dougrez-Lewis, M. Liakata, E. Kochkina, and Y. He, “Learning Disentangled Latent Topics for Twitter Rumour Veracity Classification,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 3902–3908. doi: 10.18653/v1/2021.findings-acl.341.

[94] M. Cheng, S. Nazarian, and P. Bogdan, “VRoC: Variational Autoencoder-aided Multi-task Rumor Classifier Based on Text,” in *Proceedings of The*

Web Conference 2020, New York, NY, USA: ACM, Apr. 2020, pp. 2892–2898. doi: 10.1145/3366423.3380054.

[95] Google AI, "Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance," *Google AI Blog,* 2022.

[96] L. Hu, S. Wei, Z. Zhao, and B. Wu, "Deep learning for fake news detection: A comprehensive survey," *AI Open,* vol. 3, 2022, doi: 10.1016/j.aiopen.2022.09.001.

[97] L. Cui, K. Shu, S. Wang, D. Lee, and H. Liu, "dEFEND: A system for explainable fake news detection," in *International Conference on Information and Knowledge Management, Proceedings,* 2019. doi: 10.1145/3357384.3357862.

[98] Z. Yang, J. Ma, H. Chen, H. Lin, Z. Luo, and Y. Chang, "A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection," in *Proceedings of the 29th International Conference on Computational Linguistics,* Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 2608–2621. [Online]. Available: https://aclanthology.org/2022.coling-1.230

[99] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "Generating fact checking explanations," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics,* 2020. doi: 10.18653/v1/2020.acl-main.656.

[100] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971,* 2023.

[101] W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," in *ACL 2017 - 55th Annual Meeting of the Association for*

Computational Linguistics, Proceedings of the Conference, 2017. doi: 10.18653/v1/P17-2067.

[102] A. Paszke *et al.*, “PyTorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems*, 2019.

[103] T. Wolf *et al.*, “Transformers: State-of-the-Art Natural Language Processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[104] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, 2011.

[105] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, “Truth of varying shades: Analyzing language in fake news and political fact-checking,” in *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2017. doi: 10.18653/v1/d17-1317.

[106] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, “Declare: Debunking fake news and false claims using evidence-aware deep learning,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2018. doi: 10.18653/v1/d18-1003.

[107] J. Ma, W. Gao, S. Joty, and K. F. Wong, “Sentence-level evidence embedding for claim verification with hierarchical attention networks,” in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020. doi: 10.18653/v1/p19-1244.