Automatic Incongruent News Detection

(From the Perspective of Body and Headline Centric Representation)

A dissertation submitted in partial fulfillment of the requirements
for the award of the degree of

Doctor of Philosophy

Submitted by

Sujit Kumar

Under the supervision of

Prof. Sanasam Ranbir Singh

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
Guwahati 781039, Assam, India

March 2024

“Our greatest glory is not in never falling, but in rising every time we fall.”

Confucius

Dedicated to my parents and my brothers for their unconditional love and support.

Acknowledgements

I am deeply grateful to Prof. Sanasam Ranbir Singh, my Ph.D. supervisor, for his invaluable guidance and unwavering support throughout my research journey. His mentorship has been instrumental in shaping my academic pursuits, providing me with opportunities to delve into research topics that truly inspire me. Simultaneously, his guidance has been pivotal in transforming abstract ideas into tangible objectives, all while keeping sight of the ultimate goal. I am also grateful to my doctoral committee members Prof. Ashish Anand, Dr. Amit Awekar, and Dr. Mithilesh Kumar Jha for their invaluable feedback and contributions throughout the research process. At this juncture, I extend my heartfelt gratitude to Prof. Sukumar Nandi, Dr. Rashmi Dutta Baruah and Dr. Vijaya Saradhi for their invaluable guidance. I am also sincerely thankful to all the faculty members for their direct and indirect support. Furthermore, I express my appreciation to the Ministry of Human Resource Development (MHRD), Government of India, for their financial assistance throughout my Ph.D. journey. Additionally, I am grateful to various funding agencies and esteemed conferences for awarding travel grants and fellowships, which encouraged my participation across multiple research venues. I am immensely grateful to various funding agencies such as BRNS (project no. 2013/13/8-BRNS/10026), the Department of Biotechnology (project no. BT/COE/34/SP28408/2018), and the Department of Computer Science and Engineering (CSE) at the Indian Institute of Technology Guwahati, for providing essential computing resources utilized throughout this work. Additionally, I would like to extend my acknowledgment to the technical staff and system administrators at IIT Guwahati, as well as the ever-supportive administrative staff in the Department of CSE, Computer & Communication Centre, Academic Affairs, and Student Affairs. I am particularly thankful to my fellow OSINT and CLST lab members at IIT Guwahati. Special mentions to Monika Singh, Akash Anil, Gyanendro Singh, Neelakshi Sharma, Bornali Phukan, Durgesh Kumar, Hemanta Baruah, Anasua Mitra, Jenil Thiyam, Soumyadeep Jana, Shifali Agrahari, Saurabh Kumar, Roshan, and many more. Throughout my PhD journey, I have been fortunate to collaborate with numerous B.Tech students whose unwavering support, affection, and dedication have been invaluable. I express sincere gratitude to Arpit

Gupta, Bhadke Rajas Jagannath, Sahil Garhwal, Mohan Kumar, Shivangi Kumar, Nishtha Sharma, Aditya Sinha, Priyank Soni, Aayush Sachdeva, Bhukya Bharath, Divyam Singal, Pradnesh Prasad Kalkar, Saket Kumar Singh, Ahaan Sameer, Anant Shankhdhar, Bhuvan Aggarwal, Advaita Mallik, Siddharth Hemant Khincha, Arpan Anil Khandare, Naman Anand, Tanveen and others for their contributions. Furthermore, I express my gratitude to the M.Tech students associated with the OSINT Lab at IIT Guwahati, special mentions to Mahima Mallik, Ankit Agrawal, Bawane Akshay Shalikram, Gaurav Kumar, Kunal Warnikar, Ronak Ramesh Chabukswar, Rohan Jaiswal, Aosenba, Mohit Sharma, Sanjiv Kumar, Mohit Ram Sharma, Abhishek Ranjan, Animesh Dey, Debang Joshi, and others, for their unwavering support and camaraderie. I am equally appreciative of the exceptional peers at IIT Guwahati, including Debanjan Roy, Amit Puri, Vanshali Sharma, Karnish Ahmed Tapadar, Himanshu Sharma, Ashita Batra, Abir Banerjee, Pankaj Kumar, Umesh Mishra, Nikhil, Sarthak Saxena, Varenyam Bakshi, Gunjan, Gulve Piyush Ashok, Saurav Yadav and countless others, whose invaluable contributions have enriched my research journey. I extend my gratitude to the rest of my friends at IIT Guwahati, whether they were hostel mates, companions from the Gym, or friends in the Students Gymkhana Council, IIT Guwahati, for infusing my time at the IITG with liveliness and fun. Finally, I express my heartfelt gratitude to IIT Guwahati for offering a beautiful and serene campus equipped with top-notch facilities. Last but certainly not least, I extend my thanks to the medical staff, security personnel, mess staff, and cleaning staff of IIT Guwahati for their invaluable contributions to the well-being and smooth functioning of the campus community.

Declaration

I certify that,

- The work contained in this thesis is original and has been done by myself and under the general supervision of my supervisor(s).
- The work reported herein has not been submitted to any other Institute for any degree or diploma.
- Whenever I have used materials (concepts, ideas, text, expressions, data, graphs, diagrams, theoretical analysis, results, etc.) from other sources, I have given due credit by citing them in the text of the thesis and giving their details in the references. Elaborate sentences used verbatim from published work have been clearly identified and quoted.
- I also affirm that no part of this thesis can be considered plagiarism to the best of my knowledge and understanding and take complete responsibility if any complaint arises.
- I am fully aware that my thesis supervisor(s) are not in a position to check for any possible instance of plagiarism within this submitted work.

Sujit Kumar

March 2024

Abstract

The prevalence of deceptive and incongruent news headlines has demonstrated their significant role in propagating fake news, which worsens the dissemination of both misinformation and disinformation. In the literature, incongruent news article detection has been studies from two aspects- *body-centric* and *headline-centric* encoding. However, earlier headline-centric and body-centric approaches in the literature fail in the following scenarios.(i) The hierarchical encoding in the earlier studies is limited to paragraph level only, and headline-guided attention highlights paragraphs that are contextually similar to headlines. However, considering the underlying incongruent news detection task, highlighting the paragraphs or sentences that are not contextually similar to the headline is essential. (ii) It fails to detect partially incongruent news articles. To address the first limitation of studies in literature, this thesis proposes a `Gated Recursive And Sequential Deep Hierarchical Encoding` **GraSHE** method for detecting incongruent news articles by extending the hierarchy structure of news body from body to word level and incorporating incongruent weights. The proposed model, *(GraSHE)* captures the long-term dependencies and syntactic structure by incorporating sequential information at the paragraph and body level (using BiLSTM), and syntactic structure at the sentence level using child-sum Tree LSTM. Further, unlike headline guided attention models, *(GraSHE)* also incorporates incongruity weight to capture non-dominant textual segments which are not congruent with other part of news body. To address the second limitation of studies in literature, this thesis proposed dual summarization and graph context matching based methods. This thesis proposes dual summarization-based methods *Multi-head Attention Dual Summarization* **MADS** and *dual-summarization based approach,* namely **DuSum**, which divides the news article body into two sets, positive and negative set. Sentences congruent to the headline are placed in the positive set, and sentences incongruent to the headline are placed in the negative set. Then, generate two different summaries of both the positive and negative sets. Next, match the headline with a summary of positive and negative for incongruent news article detection. This thesis proposed graph context matching based methods, *Graph-based Context Matching* **GCM** and *Graph-based Dual Context Matching* (**GDCM**). Both **GCM** and **GDCM** methods first represent headlines

and news bodies as a bigram network to capture contextual relations between words and document structure. Then, for every word in the headline, both methods extract context from the news body Bigram Network. Next, it estimates the similarity between the extracted context and the headline for incongruent news detection. **GCM** extract only positive context (context of headline from paragraphs or sentences where discussion regarding headline key is preset). Whereas, **GDCM** extracts both positive context (context of headline from paragraphs or sentences where discussion regarding headline key is preset) and negative context (context of headline from paragraphs or sentences where discussion regarding headline key is not preset). For all the proposed methods, we conduct extensive experiments on three publicly available benchmark datasets. Our experimental results suggest that the proposed model outperforms existing state-of-the-art models in literature and efficiently detects partially incongruent news.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

With social media becoming ubiquitous today, the misinformation and disinformation circulated over digital platforms is posed as one of the biggest threats to the society, and detection of such information is of paramount important. Misinformation is generally defined as untruthful, inaccurate information in circulation, regardless of intent to mislead. Whereas disinformation (a subclass of misinformation) is distinguished as the information which is circulated with the deliberate intention to mislead people. According to the Statista 2020 report, the proliferation of misinformation (in regard to political, financial, and health care fields only) results in worldwide economy loss of 78 billion US dollars. It affects not only the economy, but also public orders.

According to the study reported at First Draft [1], misinformation and disinformation can be categorized into seven categories based on their purposes and intent, namely satire or parody, misleading content, imposter content, fabricated content, false connection, false context and manipulated content. Though the underlying intent of each type defers, it is hard to clearly distinguish from one type to another from their constructs. This thesis focuses on detecting one particular class of misinformation called *incongruent news*. Incongruent news is not in the above seven classes, but incongruent news is related to fabricated content, false connection, false context and manipulated content. A news article is said to be incongruent if its headline misrepresents its body through fabrication, manipulation, false connections, or incorrect context [2–4]. The thesis explores to handle the problem from three different aspects - (i) *deep hierarchical encoding of documents to capture better contextual representation of the texts*, (ii) *dual-summarization of the documents to capture congruent and incongruent abstractions of the texts*, and (iii) *graph-based headline-centric context extraction from news*

[1] https://firstdraftnews.org/articles/fake-news-complicated/

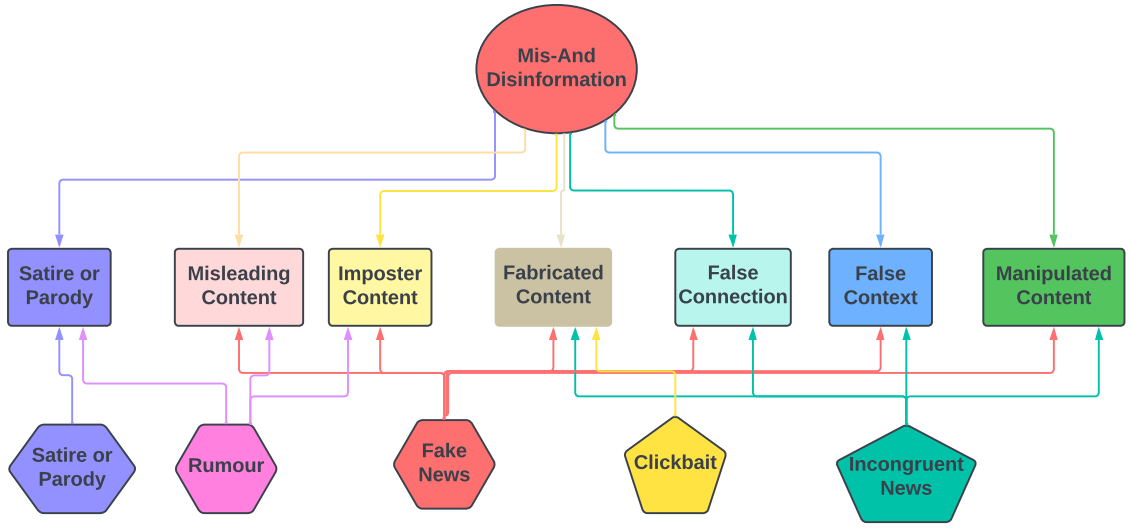


Fig. 1.1 Misinformation Taxonomy

body to verify the consistency between headline and news body proposes various methods, and investigates their effectiveness using various datasets.

1.1 Taxonomy of Misinformation and Disinformation

Following the categorization in First draft, [2] misinformation and disinformation can be categorized into the following seven categories (as shown in first level of Figure 1.1) .

- **Satire or Parody :** This type of content is created by mimicking genuine content with no intention to harm, but the content has potential to fool the reader or audience. Generally, parody content is created for mimic and entertainment purpose.
- **Misleading Content :** Misleading content is created to frame an issue or individual, which are created due to poor journalism, partisanship with the intention to run propaganda or political influence.
- **Imposter Content :** This type of content is created by spoofing legitimate content.Imposter contents are generally created with the intention of parody, provocation, profit and propaganda.
- **Fabricated Content :** Misinformation in this category is created, which is completely false with the intention to deceive and harm. Fabricated content is created for parody, provocation, profit, political influence and propaganda.

[2] https://firstdraftnews.org/articles/fake-news-complicated/

- **False Connection :** False connection in news reporting happens when its headline does not support its body. False connections are created either due to poor journalism or with the motive of profit.
- **False Context :** Legitimate information is presented in the wrong context. False contexts are created with the intention of partisanship, political influence or propaganda.
- **Manipulated Content :** This type of misinformation happens when legitimate information is distorted with the intention to deceive and harm.

In the literature, the above types of misinformation and disinformation are studies relating to different research areas such as *rumor detection, fake news detection, clickbait detection and incongruent news detection* (as shown in second level of Figure 1.1).

1. **Rumor:** It is defined as a piece of information under circulation whose truthfulness is yet to be verified. As shown in Figure 1.1, misinformation in the form of satire or parody, misleading content and imposter content are considered rumors.
2. **Fake News:** In literature, fake news is defined as a story that is internationally created to mislead the readers. According to media scholars, fake news can be defined as *distorted and deceitful content in circulation as news over a communication medium such as print, electronic, and digital communication* [5]. As shown in Figure 1.1, mis-and disinformation in the form of misleading content, imposter content, fabricated content, false connection, false context and manipulated content are considered fake news.
3. **Clickbait:** Clickbait detection focuses on identifying news headlines that are created with several stylistic and linguistic features such as forward-referencing, mentioning of attractive words, public figures and personalities, etc., for the purpose of attracting the attention of the readers to read the article. Since clickbait headlines are fabricated to attract the reader, misinformation in the form of fabricated headlines is considered clickbait.
4. **Incongruent News:** A news article is considered to be incongruent if its headline does not represent its body due to fabricated, manipulated, false connection or wrong context. As indicated in the definition of incongruent news, either a headline misrepresents the news body due to a fabricated or manipulated news body, a false connection between headline and body or a headline is presented in a false context to the news body. Accordingly, misinformation in the form of fabricated, manipulated, false connection or wrong context is considered incongruent news. Though incongruent news article detection and clickbait detection are related with regard to news headlines, incongruent news article detection is comprehensively different from clickbait detection [6][3]. While clickbait detection focuses on identifying news headlines that are created with several stylistic and linguistic features such as forward-referencing, mentioning of attractive words, public figures and personalities, etc. for the purpose of attracting

attentions of the readers to read the article, the incongruent news article detection focuses on identifying news articles whose headlines are incongruent with their respective news bodies [3]. Clickbaits attempt to attract readers to click on the headlines and read the news articles. The incongruent news articles are created for spreading misinformation. Clickbait may be described by only headlines, but incongruent news need to be described by the relation between the headline and the body [6].

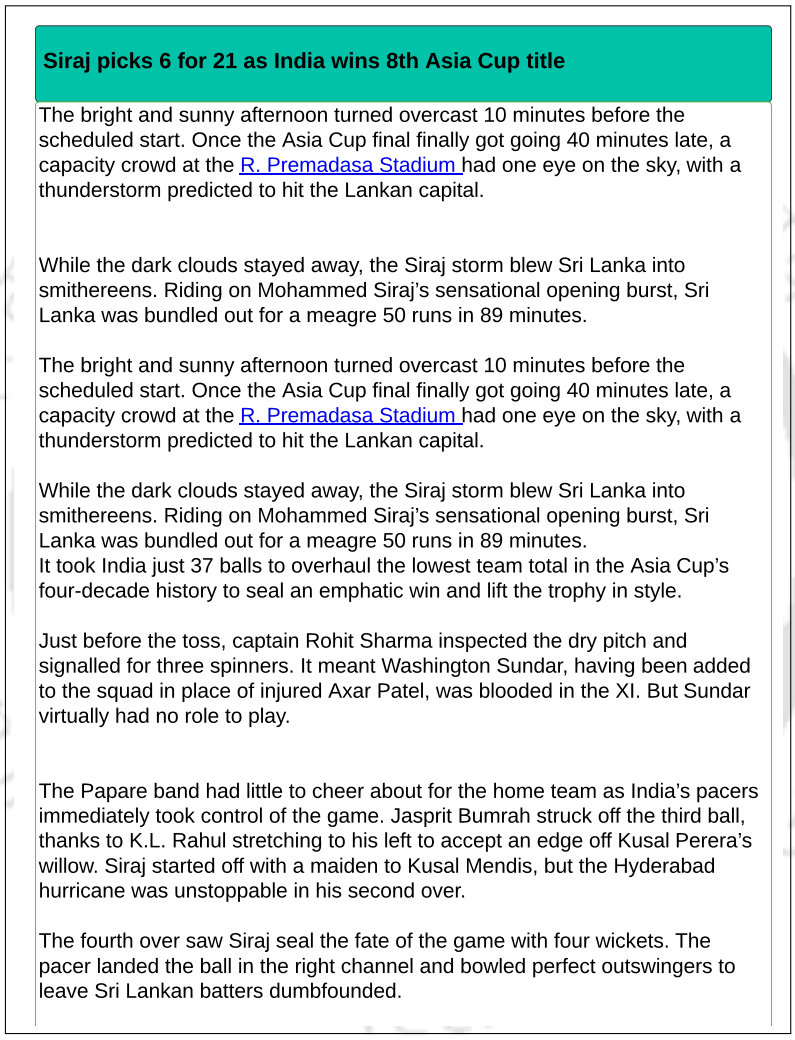


Fig. 1.2 Present example of congruent news. It is evident that the claim made in each paragraph of news body supports the claim made in the news headline. Therefore, it is a fully congruent news article.

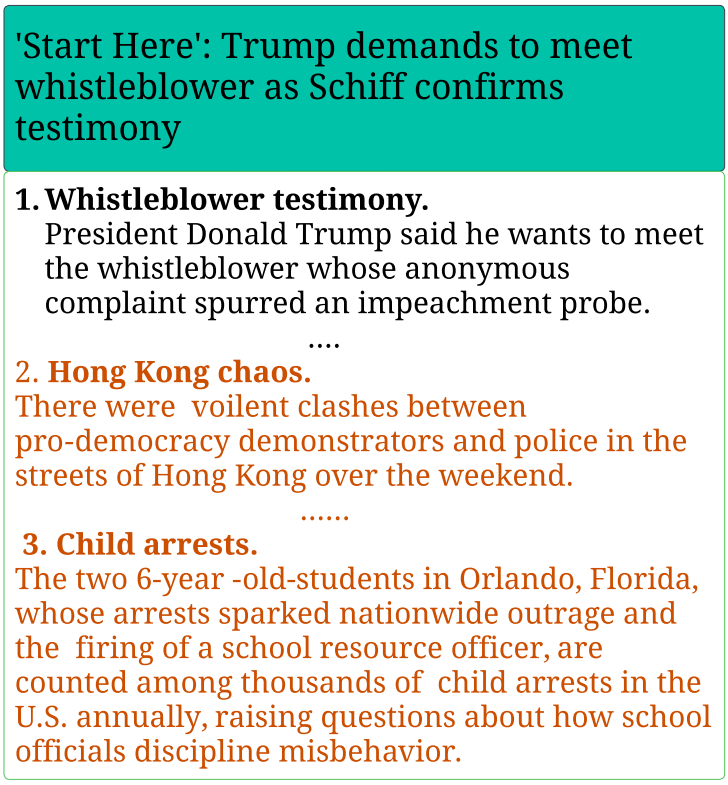


Fig. 1.3 Present example of partially incongruent presented in the study [1]. Where 1. *Whistleblower testimony*, 2. *Hong Kong chaos* and 3. *Child arrests* indicates the news body's first, second and third paragraphs. It is evident that the paragraph 1. *of Whistleblower testimony* is congruent to the headline. But the paragraphs 2. *Hong Kong chaos* and 3. *Child arrests* are not congruent with the headline. Because the headline is in the context of President Donald Trump wanting to meet a Whistleblower, paragraphs 2. and 3. are about *Hong Kong chaos* and *Child arrests*

1.2 Incongruent News Articles

In literature, incongruent news article detection is also known as *misleading headline or deceptive news*. Incongruent news articles can be categorized into two categories: *fully incongruent news and partially incongruent news*. Figure 1.2 presents an example of congruent news where the claim made in the headline supported by the claim made in the news body. A news article is said to be fully incongruent news if news body content does not support the claim made in the headline. Figure 1.4 presents an example of incongruent news, where the headline says, "Why doesn't America Read Anymore?". This indicates that people in America are losing interest in the reading. But the news body does not discuss anything which suggests America is losing interest in reading; instead, it says, "If you are reading, just

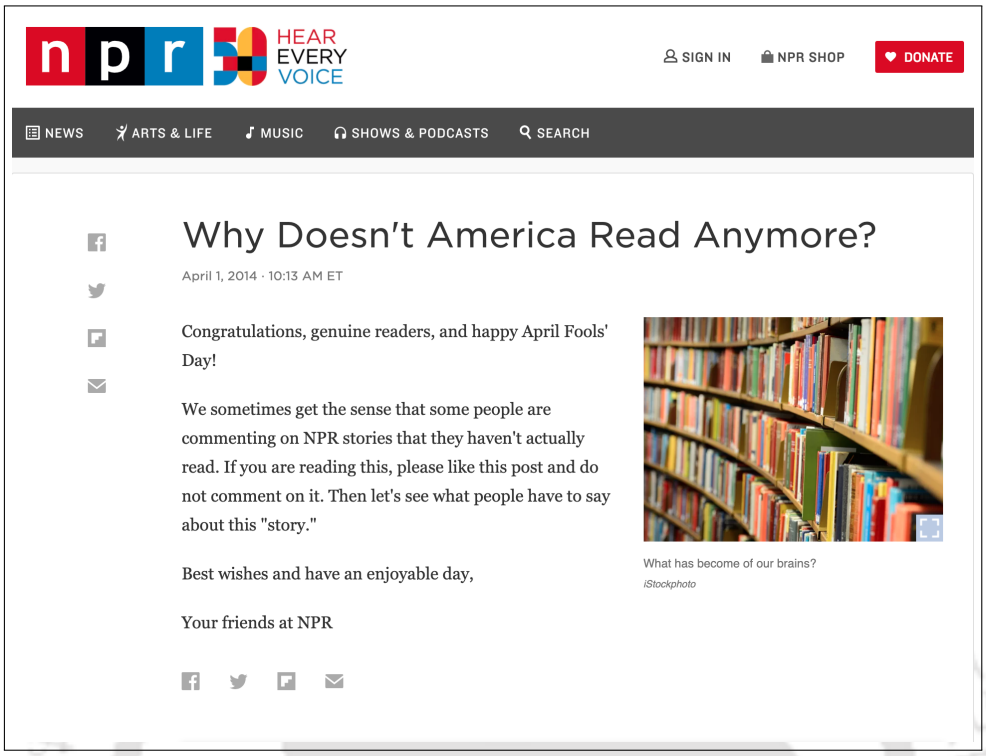


Fig. 1.4 Present example of fully incongruent news published by NPR [3]. Where the headline says, "Why doesn't America Read Anymore?" This indicates that people in America are losing interest in reading. But the news body does not discuss anything which suggests America is losing interest in reading; instead, it says, "If you are reading, just like this post and do not comment". So clearly, the headline's claim is not supported by any relevant discussion in the news body.

like this post and do not comment". So clearly, the headline's claim is not supported by any relevant discussion in the news body. Whereas, partially incongruent news is a special case of incongruent news where one or more paragraphs of the news body are congruent to the headline, and one or more paragraphs of the news body are not congruent to the headline. Figure 1.3 presents an example of partially incongruent presented in the study [1], where "*1. Whistleblower testimony*", "*2. Hong Kong chaos*" and "*3. Child arrests*" indicate the first, second and third paragraphs of the body. It is apparent that the first paragraph is congruent to the headline, but the second and third paragraphs are not congruent to the headline. Because the headline is in the context of President Donald Trump wanting to meet a Whistleblower, whereas the second and third paragraphs are about *Hong Kong chaos* and *Child arrests*.

[3] Example fully incongruent news

1.3 General Characteristics of Incongruent News Articles

Given a pair of news headlines and news bodies, incongruent news article detection aims to determine whether the news headline and body convey the same information to its reader. If a pair of news headlines and body give the same information to its readers, then it is said to be a congruent news article. Otherwise, it is an incongruent news article. Deceptive and incongruent news headlines have proven to be a potent catalyst in disseminating fake news, leading to a dual impact that exacerbates the problem [7]. First, when incorporated into trending news articles, these headlines lure readers into engaging with the content, thereby amplifying the reach of the false information. Second, they perpetuate a cycle of misinformation by distorting facts and manipulating readers' perceptions, undermining the credibility of legitimate news sources. This two-fold harm significantly contributes to the spread of misinformation and disinformation in today's media landscape. In recent times, the usage of deceptive and incongruent news headlines as an effective means to spread misinformation and disinformation over digital platforms is evident [3, 8] [4,5]. As news headlines greatly influence the opinion of the readers [9], it plays a significant role in making a new viral on any social media [10, 11, 4]. As reported in the studies [11, 9, 12], the influence of the news headline is persistent, and approximately 60% of the news circulated on social media is shared without reading the whole article [11]. This indicates that spreading of incongruent headlines leads to the spreading of misinformation and disinformation over digital platforms. A deceitful and incongruent news article can negatively affect readers, such as false beliefs and wrong opinions [6,7] [2, 13, 14].

Studies [7, 15, 16] suggest that most of the time, once individuals have conceived a misleading and deceitful headline, it could be difficult to correct their credence even after reading the full news article. Correcting views gained from misleading and deceitful headlines depends on factors such as their knowledge and understanding of the topic, their initial beliefs, and the perceived strength of the correcting evidence. Furthermore, the study [17] finds that corrections often prove ineffective in diminishing the impact of misinformation. In certain instances, they may even backfire by reinforcing individuals' misconceptions instead. Consequently, detecting deceitful and incongruent news articles [3] [2] [18] [19] [20] is becoming an important research problem to counter the spread of misinformation over digital media. An incongruent news article may be constituted in various forms. As reported in

[4] Examples of misleading headline fake news
[5] Misleading headline fake news over WHO
[6] Impact of the misleading headline in health
[7] Misleading headlines effect on economy news

the studies [3, 21], there are four key characteristics of incongruent news articles as briefly described below.

1. **The headline makes unrelated or opposite claims to its body**: An incongruent news article falls under this category if the news body and headline are from different topics or events, or claims.
2. **Both the headline and body refer to a common topic or event, but the contents are not related** : An incongruent news article falls under this category if both the news headline and news body refer to the same topic or event, but the claim made in the headline is not supported by the content in the news body.
3. **Both headline and body report a genuine event/incident, but the numeric figure or name entities are manipulated**: An incongruent news article falls under this category if both the headline and news body are similar, but name entities or numeric figures are manipulated to mislead readers.
4. **Partially incongruent news**: An incongruent news article falls under this category if the claim made in the headline is supported by one or more paragraphs and the claim made in the headline is not supported by one or more paragraphs of the news body.

1.4 Challenges in Incongruent News Detection

As outlined in Section 1.3, the detection of incongruent news articles stands as a crucial challenge in combating the dissemination of misinformation and disinformation across digital platforms. In general, incongruent news article detection poses the following challenges.

Capturing Contextual Similarity (C1) : A news headline may be congruent even if there is no keyword overlap between the headline and body. In other words, the headline may be expressed in terms of synonyms of keywords used in their respective body. Consequently, models should be able to perform nuanced analysis of the headline and body, consider the contextual meaning of sentences and headline, and resolve ambiguity between words and phrases within the headline and body. This will enable the system to detect incongruity between headline and body based on semantic similarity between headline and body.

Capturing Topic and Semantic Relatedness (C2) : A news article could be incongruent, even if the headline and body are from the same topic. Accordingly, systems must understand the underlying meaning and thematic coherence between the headline and the body. By resolving semantic and topic relatedness, the system can evaluate whether

the article's content aligns with the central theme and intent expressed in the headline. Furthermore, by assessing the semantic and topic relatedness, the system can identify any discrepancies or contradictions between the headline and the body, which can indicate potential instances of misinformation or incongruent news.

Resolve Complex Negation (C3) : A news article could be congruent, even if news articles present conflicting viewpoints, which means that sometimes in a news article, an author may present a conflicting opinion or information where a different perspective regarding the same topic or event is discussed. Consequently, it is essential for a system to identify instances where the content contradicts or deviates from the headline's claim. In other words, the system must resolve complex negations between sentences of body and between headline and body for efficient incongruent news detection. The ability to resolve complex negation helps the systems identify the contradictions and inconsistencies between the headline and the body of an article, leading to more accurate assessments of news content.

Resolve Propositions Between Sentences and Phrases (C4) : Sentences and words within news articles are connected through propositional operators (and or, if then etc.), which indicate a logical relationship between words and sentences of headline and body, respectively. Accordingly, the system must have the ability to resolve propositions between sentences and phrases for efficient incongruent news detections. Understanding propositional content helps to extract the core message or information conveyed by a sentence or between sentences. Hence, the ability to resolve propositions between sentences and words system to assess whether the body of the article aligns with the claims made in the headline or if there are inconsistencies and contradictions between the two.

Overcome Text Length Mismatch Between Headline and Body (C5) : A news headline is usually small and consists of a few words, whereas the news body is generally larger than the headline. This length mismatch between headline and body leads to an unfair comparison between headline and body.

Detect Partially Incongruent News (C6) : A news article could be partially incongruent where one or a few sentences are congruent with the headline while the rest of the sentences are incongruent with the headline. In the case of partial incongruent news, it is important to consider congruity between headline and sentences rather than congruity between headline and body.

Headline Centric Context Extractions (C7) : The main keywords in the headline are discussed in different body segments. So, the system must extract context or discussion regarding keywords from different body segments and aggregate this information to match the discussion or context of keywords in the headline and different body segments. Therefore, it is essential to aggregate contextual information regarding keywords in a headline in different body segments and evaluate to verify the consistency between the headline and body. Incongruence or inconsistency between the aggregated context and the headline may indicate potential misleading or deceptive practices within the news article.

1.4.1 Opportunities and Scope

Earlier studies on incongruent detection can be grouped into two categories, namely *similarity-based* and *summarization-based*. In a similarity-based approach, the idea is to learn the encoding of a news headline and its body, and check their similarities. Initial studies [22, 23] on similarity-based approaches attempt to address the first challenge using bag-of-words-based features. However, as observed in [24], the above studies rely on the lexical overlap between headlines, and need to capture the semantic similarity between headline and body. Realizing the importance of contextual information present in the news articles, the studies in [24, 25] attempt to address the first challenge by utilizing both the contextual information and the bag-of-words-features. Combining bag-of-words-based features with sequential encoding helps the models capture the semantic similarity between the headline and body. However, these models often fail to capture information, such as complex negations and propositional contents, which are important for incongruity detection [24]. Further, the studies in [26–28] exploit the hierarchical structure of news articles (body-paragraph-sentence relation) along with contextual and sequential information present in the news article to address the first, second and third challenges. Our preliminary experiments on similarity-based methods, and also in the study [1], suggest that these methods perform well for small articles of few paragraphs but perform relatively poorly for large articles with a large number of paragraphs. Unlike similarity-based methods, summarization-based studies [29–31] summarize the body of a news article to a synthetic headline to address the fifth challenge. Then, the synthesized headline and the actual headline are matched. The synthetically generated headline from the news article body may not be a faithful or good representation of the news article body [32–34]. As the summarization in studies [29–31] are biased towards the dominant content of the body, such summarization may fail to capture

the embedding noise present in partially incongruent news articles. Consequently, the above summarization-based method fails to detect partially incongruent news articles.

Incongruent news detection studies in literature can also be categorized as body-centric and headline centric. Methods mainly focus on obtaining an effective representation of the body for incongruent news detections is called the body-centric approach. In contrast, methods which extract headline-centric information from the new body are called headline-centric methods. The above-mentioned similarity and summarization-based approaches are body-centric methods that attempt to address incongruent news detection challenges. Table 1.1 presents the mapping between models proposed by studies in literature and the challenges of incongruent news detection. From Table 1.1, it is evident that the similarity-based approach attempts to address challenges *Capturing Contextual Similarity (C1)*, *capturing topic and semantic relatedness (C2)* and *resolve complex negation (C3)* but fails to address challenges *resolve propositions between sentences and phrases (C4)*. This means that similarity-based studies in the literature can easily handle semantic and topic similarity between headline and body but fail to *resolve complex regulation (C3)* between sentences and phrases of the body and fail to understand propositional contents *resolve propositions between sentences and phrases (C4)*. A similar observation is also presented in studies [24, 28]. Accordingly, more sophisticated methods which capture deep semantic relationships between headline and body and the ability to handle complex negation and propositional content between headline and body are needed [24, 28]. Both Similarity and summarization-based methods in the literature fail to detect partially incongruent news efficiently. There is a lack of datasets in resource poor language for incongruent and misleading news detection.

1.5 Research Gaps and Thesis Objectives

Motivated by the challenges of incongruent news detection and the gap in the literature mentioned above, this thesis's research objective is the following.

Deep Hierarchical Encoding (R1) : Proposes a model for detecting incongruent news by considering deep hierarchical encoding, sentence syntactic structure, and incorporating the incongruity weight. Deep hierarchical encodings refer to considering the hierarchical structure of a news document, i.e., *body as a collection of paragraphs, and paragraph as a collection of sentences* and sentence syntactic structure helps to

Table 1.1 Present the mapping between incongruent news detection challenges and literature studies that attempt to address respective challenges. Here, *Capturing Contextual Similarity (**C1**)*, *Capturing Topic and Semantic Relatedness (**C2**)*, *Resolve Complex Negation (**C3**)*, *Resolve Propositions Between Sentences and Phrases (**C4**)*, *Overcome Text Length Mismatch Between Headline and Body (**C5**)*, *Detect Partially Incongruent News (**C6**)* and *Headline Centric Context Extractions (**C7**)* refer to the challenges of incongruent news detection. Similarly, *Deep Hierarchical Encoding (**R1**)*, *Dual Summarization (**R2**)*, and *Graph-based Headline Centric Context Extractions (**R3**)* are the research objectives of this thesis.

Approach		**Models**	**Challanges**						
			C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**
Similairty		FNC [22]	✓						
		Talo Xgboost	✓						
		UCLMR [35]	✓						
		StackLSTM [24]	✓						
		AHDE [28]	✓	✓	✓				
		HDSF [36]	✓	✓	✓				
		Poshan [37]	✓	✓	✓				
		GHDE [1]						✓	
Summarization		FEDS [38, 31]					✓		
		HeadlineStanceChecker [29]					✓		
		MuSeM [30]					✓		
Research	**R1**	GraSHE [39], RoBERT [40]	✓	✓	✓	✓			
	R2	MADS [21], DuSum					✓	✓	
	R3	GCM and GDCM							✓

extend hierarchy from sentence to words and capture long-term dependency among the words within a sentence. As textual documents are order sensitive, extending the hierarchical structure up to the word level helps capture long-term dependencies and syntactic structure between words within a sentence. Similarly, incongruity weight for a paragraph refers to its inability to represent other paragraphs in the news body. Incorporating incongruity weight helps to capture non-dominant textual segments which are not congruent with other parts of the news body. Since all three components of the proposed model, namely deep hierarchical encoding, sentence syntactic structure, and incorporating the incongruity weight, aims to obtain an effective representation of the news body for incongruent news detection, this is a body-centric method.

Dual Summarization (R2) : Propose a model to efficiently detect partially incongruent news using dual summarization. Dual summarization splits the sentences in a news article into two sets, *highly congruent sentences with headline* and *poorly congruent sentences with headline,* based on the relation between headline and sentences. Next, it extracts summaries of the highly and poorly congruent sets separately and performs matching between summaries and the headline. If part of the news article is incongruent with the headline, it is moved to a poorly congruent set, otherwise to a highly congruent one. The intuition is that while generating encoding of the body, one should incorporate both the congruent and incongruent parts of the body to enable capturing minority incongruent content as well. Though this dual summarisation-based approach aims to generate a summary of a new body, but sentences are placed in *highly congruent sentences with headline* and *poorly congruent sentences with headline,* based on the relation between headline and sentence. Consequently, it is a headline centric approach.

Graph-based Headline Centric Context Extractions (R3) : Propose a model to detect incongruent news by matching the headline and headline-centric context extracted from the body. Headline-centric context means the context of each word from the headline present in different segments of the news body. This model solely relies on the matching headline and headline-centric context extracted from the news body; accordingly, this is a headline-centric approach.

1.6 Thesis Overview and Contributions

This thesis aims to address incongruent news article detection challenges to counter the spread of misinformation and disinformation over digital platforms. This thesis has made

three contributions. This thesis first studies the effect of deep hierarchical encoding of news articles by exploiting the syntactic structure of sentences and the hierarchical structure of the new body to capture long-term dependencies between words of sentences for incongruent news article detections. Exploiting the sentence's syntactic structure helps capture long-term dependencies between words, improving the model's deep semantic understanding and resolving complex negation and understanding of propositional content. The second contribution of the thesis proposes dual summarization-based methods to efficiently detect partially incongruent news detection based on headline and sentence relationships and summarizations. The third contribution of the thesis proposes graph-based headline and body context matching methods, which match the headline and context extracted from the news body specific to words in the headline for incongruent news detection. The graph-based headline and body context matching methods helps to aggregate contextual information regarding keywords in the headline from different segment of the body.

1.6.1 Deep Hierarchical Encoding for Detecting Incongruent News Articles

Earlier, researchers have exploited various feature engineering approaches and deep learning models with embedding to capture incongruity between news headlines and their respective bodies. Studies have broadly considered different combinations of bag-of-words-based features, sequential encoding, hierarchical encoding, headline-guided attention-based encoding, etc., of the text in headlines and bodies. This thesis focus on addressing two important limitations observed with hierarchical encoding and headline-guided attention-based encoding methods. Existing hierarchical encoding-based studies limit the hierarchical structure of the body of a news article to paragraph level only, undermining the importance of incorporating long-term dependency from word level to sentence, paragraph and body. Further, existing headline-guided attention-based encoding focuses on contextually similar contents in the body of the headline, undermining the importance of incorporating contextually dissimilar contents. Motivated by the above observations, this thesis proposes a `Gated Recursive And Sequential Deep Hierarchical Encoding` *(GraSHE)* method for detecting incongruent news articles by extending the hierarchy structure of news body from body to word level and incorporating incongruent weights. The proposed model, *(GraSHE)* captures the long-term dependencies and syntactic structure by incorporating sequential information at the paragraph and body level (using BiLSTM) and syntactic structure at the sentence level (child-sum Tree LSTM [41]). Further, unlike headline guided attention models [28][27],

(GraSHE) also incorporates incongruity weight to capture non-dominant textual segments which are not congruent with other parts of news body. From various experimental observations over three publicly available benchmark datasets, it is observed that the proposed method outperforms its bag-of-words, sequential, and hierarchical counterparts.

1.6.2 Dual Summarization for Detecting Incongruent News Articles

Earlier studies [22, 35, 23, 24, 27, 25, 28, 1] on incongruent news article detection mainly focus on estimating the similarity between headlines and bodies. However, the similarity-based method fails in case of the body is too large due to a length mismatch between the headline and body. To overcome such limitations, further studies [29–31] summarize the news article body with summarization methods and then estimates the similarity between the headline and summarized body. While a news article may become incongruent with the presence of a small incongruent text in the body, the majority of the above methods may be biased towards the dominant content (for instance, partially incongruent news articles). As the summarization in these studies are biased towards the dominant content of the body, such summarization may fail to capture the embedding noise present in partially incongruent news articles. Hence, we need an incongruent news article detection-specific summarization technique, which should focus more on the incongruent part of the news article while generating a summary of the news article body.

Motivated by the above issues, this thesis proposes dual summarization-based models, which split the sentences in a news article into two sets; *highly congruent sentences with headline* and *poorly congruent sentences with headline* based on their relationship with the headline. If part of the news article is incongruent with the headline, it is moved to a poorly congruent set, otherwise to a highly congruent one. The intuition is that while generating encoding of the body, one should incorporate both the congruent and incongruent parts of the body to enable capturing minority incongruent content as well. Subsequently, apply a match between the summary of both sets to check consistency and similarity within body content. Similarly, we also apply a match between the headline and summaries of sets to verify the similarity and consistency between the headline and body contents. In this way, our proposed dual summarization models check consistency and similarity within body contents and consistency and similarity between headlines and body contents, which is important for incongruent news articles and fake news article detection. From various experimental observations over three publicly available benchmark datasets, it is evident that the proposed method outperforms its baseline counterparts and favourably handles the (s1), (s2), (s3), (s5)

and (s6) challenges of incongruent news mentioned above. Our proposed models outperform the state-of-the-art summarization-based model in the literature with a significantly high margin over FNC and NELA datasets, respectively.

1.6.3 Graph-Based Context Matching for Incongruent News Detection

Detection of incongruent news has emerged as an essential research problem to counter misinformation and disinformation over digital media. Initial studies on incongruent news detection mainly rely on estimating the similarity between headline and body. However, similarity-based studies fail to handle text length mismatches between headline and body. Summarization-based studies attempted to address the above limitation by summarizing the news body to generate a synthetic headline and estimating the similarity between the synthetic headline and headline. However, summarization-based studies also fail to detect partially incongruent news articles.

Motivated by the above issues, this thesis proposes two context matching models, namely `Graph-based Context Matching` **GCM** and `Graph-based Dual Context Matching` **GDCM**. Initially, **GCM** and **GDCM** construct a bigram network for both headline and body separately, where unigram is a node and edge between nodes if two words co-occur in the headline or body. Next, initialize each node with FastText embedding of the corresponding unigram word associated with the respective node. For each node in headline bigram networks, **GCM**, the model searches for their best matching (maximum similarity) nodes in the body bigram network using cosine similarity between embedding of nodes. Now, considering the neighbour nodes of best matching nodes in the bigram network of the body for each word in the headline, **GCM** aggregates their context from the news body. Subsequently, a context-matching vector is obtained by applying multi-head attention between the headline and the context of each headline's word in the news body. Finally, a feature vector is obtained by combining the encoded representation of headline, body and context-matching features through the attention mechanism and passed through a neural network for incongruent news classification.

Instead of considering only best matching (maximum similarity) nodes in the body bigram of the body for each node in the headline bigram network. **GDCM** construct two sets of matching nodes for each node in the bigram network of headlines: positive and negative sets. Nodes with the least matching (minimum similarity) are placed in the negative set, and the node with the best matching (maximum similarity) is placed in the positive set for each node in the bigram network of headlines. Next, considering the k hop neighbour

nodes of nodes in both positive and negative **GCM** aggregates, the positive and negative context of each node in the headline from the bigram network news body. Subsequently, a positive context-matching feature vector is obtained by applying multi-head attention between the headline and the positive context of each headline's word in the news body. Similarly, a negative context-matching feature vector is obtained by applying multi-head attention between the headline and the negative context of each headline's word in the news body. Here, multi-head attention in the positive and negative sets is different. In the case of multi-head attention, multi-head attention between the headline and the positive context of each headline's word in the news body, high weight is given to context which holds high similarity with the headline. Whereas in the case of multi-head attention between the headline and the negative context of each headline's word in the news body, high weight is given to the context which holds the least similarity with the headline. The prime motivation behind considering positive and negative sets and the above-mentioned multi-head attention for positive and negative sets is that if the context of the word in the headline is close to the context of the least similar node and the maximum similar nodes context of the headline and body are similar and also both headline and body semantically similar. Consequently, it is incongruent news and incongruent otherwise. The high similarity between the context of the words in the headline and the context of words in the positive and negative set also indicated that sentences and paragraphs within the news body are consistent, and also the body is congruent with the headline. Similarly, the case of the partially incongruent context of words in headlines will be similar to the context of words in a positive set, but the context of words in the headline will not be similar (at least similar) to the context of words in a negative set. Finally, a feature vector is obtained by combining the encoded representation of the headline, body, negative context-matching feature and negative context-matching feature through the attention mechanism and passed through a neural network for incongruent news classification. Our experimental results over three benchmark datasets suggest that proposed models **GCM** and **GDCM** outperformed state-of-art models in literature with significantly high margins and efficiently detected partially incongruent news articles. Since both propose models **GCM** and **GDCM**, do not rely on the direct similarity between headline and body, instead **GCM** and **GDCM** rely on the similarity between the context of words in headline and body. Hence, **GCM** and **GDCM** overcome the limitation of similarity-based methods, i.e., text length mismatch between headline and body.

1.7 Thesis Organization

- **Chapter 1 Introduction**: This chapter introduces the problem of incongruent news article detection, the challenges involved, and the motivation of the thesis work. The research objective of this thesis work is formally discussed, followed by an overview of contributions made.
- **Chapter 2 Background Study**: This chapter discusses the different incongruent news detection approaches and prior studies on incongruent news detection.
- **Chapter 3 Datasets**: This chapter discusses the datasets related to incongruent news detection and introduces our proposed datasets for incongruent news detection in The Hindi Language.
- **Chapter 4 Deep Hierarchical Encoding for Detecting Incongruent News** : In this chapter, the first contribution of the thesis work is presented, i.e. the proposed methods `Gated Recursive And Sequential Deep Hierarchical Encoding`, which use the deep hierarchical encoding of a news article by exploiting the hierarchical structure of the new body and syntactic structure of headline and sentences of news body.
- **Chapter 5 Dual Summarization for Detecting Incongruent News** : In this chapter, the second contribution of the thesis work is presented, i.e., the proposed methods `Dual-Summarization based methods`, which generates a dual summary for detecting partial incongruent news.
- **Chapter 6 Graph-Based Context Matching for Incongruent News Detection**: In this chapter, the third contribution of the thesis work is presented, i.e., the proposed methods `it Graph-based Context Matching` **GCM** and `Graph-based Dual Context Matching` **GDCM** which match the context of words in headline and the context of words in the body for incongruent news detection.
- **Chapter 7 Conclusion and Future Work**: This chapter concludes with possible future research directions of this thesis.

Chapter 2

Background Studies

In this chapter, we provide a comprehensive overview of all the essential background information required to comprehend the contributions made in this thesis. We aim to lay the foundation by explaining the fundamental concepts, techniques, and theories relevant to the required topics. These topics can be grouped into four categories: (i) representation of a word, (ii) sentence encoders, (iii) attention mechanism and (iv) network representation and learning. Accordingly, there are four sections in this chapter. Section 2.1 presents the details of word embedding used to represent words in this thesis, Section 2.2 presents details of different encoders used to encode sentences, paragraphs or news body in this thesis, Section 2.3 briefly discuss attention methods used in this thesis and Section 2.4 discusses the details of ngram network used to represent the news body and headline along with message passing-based graph neural network.

2.1 Representation of Words

Distributed representation of words serves as the foundational basis for several natural language processing tasks. Incongruent news detection involves estimating contextual similarity between encoded representations of news headline and body. Word embedding is the basic building block of encoding news headlines and bodies. In this section, we discuss several word embedding methods which have been used to represent the words present in news headlines and bodies in different methods proposed in this thesis. In literature, several word embedding methods have been proposed to obtain a distributed representation of words. The prime objective of every word embedding is to capture the contextual similarity

between words in a numerical format to enable the handling of abstract semantic concepts associated with words. The theoretical basis behind any word representation is that the similarity between two words should be high if they are contextually and semantically similar. These word representation methods may be grouped into two categories: (i) frequency-based methods and (ii) representation learning.

2.1.1 Frequency-based Methods

Initial word representation methods mainly rely on the frequency of words within the corpus (collections of documents) to form word representations. Consequently, frequency-based methods transformed the words into a word representation vector based on the frequency of words within documents.

2.1.1.1 One Hot Encoding

A conventional word representation represents words as one-hot vectors [42]. Given a vocabulary of the n unique word, a one-hot representation for a random i^{th} word from the vocabulary is obtained by creating a vector of length n, where i^{th} the index of the vector is set to one, and all other elements except the i^{th} index are set to zero. Though the one-hot encoding approach is simple, the following are the key limitations with one hot encoding representation : (i) the main limitation with one-hot encoding is that it fails to consider semantic relatedness between words [43], (ii) one hot encoding is a sparse representation (iii) cosine similarity between any two one hot representation will always be zero. Therefore, one hot representation cannot express the similarity between two words.

2.1.1.2 Term Frequency (TF)

Both Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) are popular methods in the literature to obtain weighted word representations. Here, the weight of each word in a document is assigned based on its frequency. Term frequency is the most straightforward approach, which assigns weight to a word based on its frequency within a document. The frequency of words may be high in a large document and low in a small document. So to counter the effect of document size on the term frequency, we normalize the frequency by dividing it by the length of documents. However, term frequency does not consider the importance of rare words beyond individual documents. The term frequency of

word w in a document d can be obtained by the equation below.

$$\mathrm{TF}\Big(\mathrm{w,d}\Big) = \Big(\frac{\mathrm{f_{w,d}}}{\sum_{\acute{w}\in \mathrm{d}} \mathrm{f_{\acute{w},d}}}\Big) \tag{2.1}$$

Where $\mathrm{f_{w,d}}$ indicates the frequency of word w in the document d and $\sum_{\acute{w}\in \mathrm{d}} \mathrm{f_{\acute{w},d}}$ indicates the total number of words in the document d.

2.1.1.3 Term Frequency-Inverse Document Frequency (TF-IDF)

As discussed in subsection 2.1.1.2, term frequency fails to discriminate between common and rare words. So to avoid the impact of common words, information retrieval and natural language processing tasks, TF-IDF multiplies the term frequency of words with their inverse document frequency. Given document collection D and term frequency $\mathrm{TF(w,d)}$ of word w along with its inverse document frequency, $\mathrm{IDFw,D}$ the $\mathrm{TF\text{-}idfw,d,D}$ can be obtained as defined below.

$$\text{TF-idf}\Big(\mathrm{w,d,D}\Big) = \mathrm{TF}\Big(\mathrm{w,d}\Big) \times \mathrm{IDF}\Big(\mathrm{w,D}\Big)\Big) \tag{2.2}$$

Again, inverse document frequency can be obtained by the equation defined below.

$$\mathrm{IDF}\Big(\mathrm{w,D}\Big) = \Big(\log \frac{|\mathrm{D}|}{|d \in D : w \in d|}\Big) \tag{2.3}$$

2.1.2 Representation Learning

As frequency-based word representations are unable to grasp the syntactic and semantic meaning of words effectively, frequency-based-based representations of words also suffer from a curse of dimensional. Such limitations of frequency-based representations methods prompted researchers to explore the acquisition of distributed word representations within a lower-dimensional space. Effective word representations are essential for superior performance of machine learning and deep learning models because the performance of models is heavily influenced by the inputs represented. In literature, several matrix factorization-based methods have been proposed to obtain a representation of words, but matrix factorization-based word representations are not flexible. Deep learning-based models are known for their ability to learn significant features autonomously. Effective representations can be acquired through unsupervised methods. Unsupervised text representation techniques like word embeddings have recently replaced frequency-based representation methods. These word embeddings have evolved into highly effective representation approaches, enhancing

the performance of diverse downstream tasks by leveraging prior knowledge across various machine-learning models. Word embedding constitutes a feature learning approach wherein a word from the vocabulary is linked to a vector in an N-dimensional space. In literature, several distinct word embedding methods have been proposed. However, this thesis discusses well-known word embedding methods such as Word2vec, GloVe, and FastText.

2.1.2.1 Word2Vec

As reported in the studies [44–46],Word2vec is a prevalent and effective method for acquiring efficient word representations from a given corpus. The Word2vec method encompasses two models: Continuous Bag of Words (CBOW) and Skip-gram. Both CBOW and Skip-gram adeptly capture the semantic nuances of words and can be readily adapted to any NLP downstream tasks.

The continuous bag of words (CBOW) method learns to predict the target word given a surrounding context of words in a fixed window size. For example, let us consider a sequence $W_1, W_2, W_3, \ldots, W_n$ of words within a corpus. Then, CBOW model aims to predict a random word W_t given surrounding context W_{t-1}, W_{t-2}, W_{t+1} and W_{t+2} in a window size of two. The main objective of the CBOW model is to maximize the probability of W_t given W_{t-1}, W_{t-2}, W_{t+1}. The CBOW method employs a three-layer neural network to obtain an efficient representation of words. The initial layer represents the context, while the subsequent hidden layer, also known as the projection layer, projects all the surrounding context words into the same vector space and then takes an average of the projected surrounding context vector. Next, the vector obtained by averaging the projected surrounding context is passed through the output layer to predict the target word. Studies [44, 45] present further details of CBOW.

The skip-gram method aims to predict the surrounding context words given the target word. For example, let us consider a sequence $W_1, W_2, W_3, \ldots, W_n$ of words within a corpus. Then, the skip-gram model aims to predict surrounding context W_{t-1}, W_{t-2}, W_{t+1} and W_{t+2} given a target word W_t in a window size of two. The skip-gram model trains a three-layer neural network to obtain an efficient word representation. The input layer of the skip-gram model represents the target word, and the output layer represents the surrounding context words. Studies [44, 45] present further details of CBOW.

2.1.2.2 FastText

Though the word2vec method is suitable for obtaining an efficient representation of words that captures the semantics interpretation of words, word2vec models cannot learn embedding or representations of out-of-vocabulary (OOV). The out-of-vocabulary (OOV) scenario occurs because some words are absent in the corpus used to train the word2vec model. The most common method to handle out-of-vocabulary (OOV) is either the out-of-vocabulary (OOV) is discarded or replaced by the word embedding of *UNK* tag. But the approach to handling out-of-vocabulary (OOV) leads to the loss of sequential and contextual information. Another key limitation of the word2vec model is that word2vec does not consider the morphology of words; instead, it assigns unique word embedding to every word. So, word2vec is unsuitable for languages with many rare words. To overcome the above limitations, study [47] propose skip-gram-based fastText embedding, which represents each word as a bag of character n-grams. A vector represents each character in words, and the vector representation of the word is obtained by summing the vector representation of each character within the word. The implementation details fastText embedding are presented in study [47].

This thesis considers Google's pre-trained word2vec word embedding method to represent words within news headlines and bodies in Chapter 4 and Chapter 5. The thesis also considers pre-trained fastText word embedding as an alternative representation for words within headlines and news bodies in Chapter 3 and Chapter 6.

2.2 Sentence Encoders

A neural network, a mathematical design inspired by the intricacies of biological neurons but not an exact replica of human nerves, has recently undergone rapid advancement, particularly in machine learning. This progress has been particularly evident in applications like image and speech recognition. Moreover, the profound impact of neural networks extends to the natural language processing domain. Incongruent news detection involves estimating contextual similarity between encoded representations of news headlines and the body. A news headline is a sequence of words, a news body is a collection of paragraphs, a paragraph is a collection of sentences, and finally, a sentence is a sequence of words. Accordingly, this thesis uses several sequential encoding and transformer-based models to encode headlines, sentences, paragraphs and news bodies. In this section of the thesis, we discuss different sequential encoding and transformer-based models, which use different proposed methods to encode headlines, sentences, paragraphs and news bodies. Word embedding is the basic building block encoding news headlines and body. In this section, we discuss several word embedding methods which we have used to represent the words of news headlines and bodies in different methods proposed in this thesis. This section briefly overviews several neural networks that find widespread use in natural language processing. In natural language processing, neural networks have been a transformative force, empowering computers to grasp, generate, and manipulate human language. Several neural networks stand out in the field of natural language processing. The majority of models designed to create distributed representations for phrases and sentences utilizing real-valued vectors to encapsulate meaning can be categorized into three primary classes: bag-of-words models, sequence models, tree-structured models and transformer models. The vector representation of the sentences obtained using bag-of-words models does not consider sequential information present in the sentences. In other words, bag-of-words-based models are ordered insensitive.

2.2.1 Sequential Encoding

Sequential models hold a crucial position within *Natural Language Processing* (NLP) owing to the inherently sequential structure of language. These models are tailored to process and interpret text data characterized by an intrinsic order or sequence, encompassing elements like sentences, paragraphs, or complete documents. In the literature, *Recurrent Neural Network* (RNN), *Gated Recurrent Units* (GRUs), *Long Short-Term Memory* (LSTM) are notable sequential models frequently employed in several tasks related to natural language processing.

A recurrent neural network (RNN) is a sequential neural network characterized by a hidden state **h** and an optional output **y**. The RNN is specially designed to process sequential information with variable length, given a sequence of words in sentences, $\mathbf{S} = \left(\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_T \right)$ the hidden state **h** of RNN is defined by the equation below.

$$\mathbf{h}_t = \mathbf{f} \left(\mathbf{U}^{(h)} \mathbf{h}_{t-1} + \mathbf{W}^{(h)} \mathbf{x}_t + \mathbf{b}^{(h)} \right) \tag{2.4}$$

Where $\mathbf{w}_t$ is input representation of the current time stamp, $\mathbf{h}_{t-1}$ is hidden state representation of the previous time stamp, $\mathbf{h}_t$is hidden state representation of the current time stamp and **f** is a non-linear activation function. The hidden state representation vector can be considered a summary of inputs and all intermediate hidden states from time stamp zero to $\mathbf{t}-1$. The weight matrix $\mathbf{w}_t$ is shared across all the time stamps for better generalization. RNN models apply *Backpropagation Through Time* (BPTT) [48–52], which is based on the backpropagation algorithm. However, RNN is simple and effective in little NLP tasks but suffers from the vanishing gradient problem, especially when RNN is applied over a longer input sequence. In the literature, several methods such as close-to-identity matrix [53], long delays [54, 53], Leaky units [55, 56, 53] and echo state networks have been explored to overcome the vanishing gradient problem in RNN.

The *Long Short-Term Memory* (LSTM) architecture proposed by the study overcomes the limitations of RNN by introducing gates that can preserve long and short-term information for longer time stamps. An LSTM (Long Short-Term Memory) unit is a fundamental element within a type of neural network known as recurrent neural networks (RNN). LSTM comprises several vectors in **d** dimensional space $\mathcal{R}^d$, each serving a unique purpose in processing sequential data.

1. **Input Gate** ($\mathbf{i}_j$): At each time step **j**, the input gate $\mathbf{i}_j$ determines how much new information should be incorporated into the memory cell. It takes inputs at the current timestamp $\mathbf{i}_j$ and the previous hidden state $\mathbf{h}_{j-1}$. By producing values between 0 and 1, the input gate $\mathbf{i}_j$ controls the new information introduced to the memory cell.

$$\mathbf{i}_j = \sigma \left(\mathbf{W}^{(i)} \mathbf{x}_j + \mathbf{U}^{(i)} \mathbf{h}_{j-1} + \mathbf{b}^{(i)} \right) \tag{2.5}$$

2. **Forget Gate** ($\mathbf{f}_j$): The forget gate $\mathbf{f}_j$ plays a crucial role in deciding what information from the previous memory cell needs to be retained or discarded. Forget gate considers the last cell memory and the previous hidden state. The forget gate $\mathbf{f}_j$ assigns values between 0 and 1, governing the extent of information to be retained.

$$\mathbf{f}_j = \sigma \left(W^{(f)} \mathbf{x}_j + \mathbf{U}^{(f)} \mathbf{h}_{j-1} + \mathbf{b}^{(f)} \right) \tag{2.6}$$

3. **Output Gate** ($\mathbf{o}_j$): The output gate $\mathbf{o}_j$ determines how much information from the memory cell should be considered for generating the hidden state output. Drawing input from both the current data point and the previous hidden state, the output gate's $\mathbf{o}_j$ values, ranging between 0 and 1, dictate the flow of memory content to the hidden state.

$$\mathbf{o}_j = \sigma\left(\mathbf{W}^{(o)}\mathbf{x}_j + \mathbf{U}^{(o)}\mathbf{h}_{j-1} + \mathbf{b}^{(o)}\right) \tag{2.7}$$

4. **Memory State** ($\mathbf{c}_j$) : Functioning as a central element, the memory cell retains and modifies information over the sequence's course. It adjusts its contents by responding to the input gate, which integrates new data, and the forget gate, which discards outdated information. This interplay empowers the LSTM to make nuanced decisions about what to preserve and what to discard as time progresses.

$$\mathbf{u}_j = tanh\left(\mathbf{W}^{(u)}\mathbf{x}_j + \mathbf{U}^{(u)}\mathbf{h}_{j-1} + b^{(u)}\right) \tag{2.8}$$

$$\mathbf{c}_j = \mathbf{i}_j \odot \mathbf{u}_j + \mathbf{f}_j \odot \mathbf{c}_{j-1} \tag{2.9}$$

5. **Hidden State** ($\mathbf{h}_j$) : The output of the LSTM at each time step is the hidden state, capturing relevant insights accumulated from the input sequence up to the current time stamp. The output gate influences the extent to which the hidden state reflects the memory cell's content.

$$\mathbf{h}_j = \mathbf{o}_j \odot tanh(\mathbf{c}_j) \tag{2.10}$$

Where $\mathbf{x_j}$ denotes the embedding of the associated word $\mathbf{w}_j$, $\mathbf{b}^{(.)}$ denotes the bias, $\mathbf{W}^{(.)}$ and $\mathbf{U}^{(.)}$ denotes the parameter matrices for respective gates. The Gated Recurrent Unit (GRU) [57] is another type of recurrent neural network (RNN) architecture, similar to the Long Short-Term Memory (LSTM). The Gated Recurrent Unit(GRU) details can be found in studies [53, 57].

2.2.2 Tree-LSTM

Though LSTM overcame the vanishing gradient problem of RNN and is also effective for sequential data, it is unsuitable for structured data represented as trees. Because encoding structured data is represented as trees, LSTM needs to incorporate information from multiple children at any time, but LSTM can incorporate information only from the previous hidden state and input at the current time stamp. To overcome such limitation of LSTM, study [41] proposed Tree-LSTM, Tree-LSTM is capable of incorporating information from multiple children units of LSTM and therefore suitable for encoding trees structured data. Given a sentence S and its dependency parse tree, let $ch(j)$ denote the set of children nodes of node j.

The hidden state of a node j is defined by the sum of the initial hidden states of its children nodes, as follows.

$$\acute{\mathbf{h}}_j = \sum_{k \in ch(j)} \mathbf{h}_k \tag{2.11}$$

Using the initial hidden state of node j, i.e., $\mathbf{h}_j$, the corresponding input, output and intermediate gates of node j are estimated as follows.

$$\mathbf{i}_j = \sigma\left(\mathbf{W}^{(i)}\mathbf{x}_j + \mathbf{U}^{(i)}\acute{\mathbf{h}}_j + \mathbf{b}^{(i)}\right) \tag{2.12}$$

$$\mathbf{o}_j = \sigma\left(\mathbf{W}^{(o)}\mathbf{x}_j + \mathbf{U}^{(o)}\acute{\mathbf{h}}_j + \mathbf{b}^{(o)}\right) \tag{2.13}$$

$$\mathbf{u}_j = tanh\left(\mathbf{W}^{(u)}\mathbf{x}_j + \mathbf{U}^{(u)}\acute{\mathbf{h}}_j + b^{(u)}\right) \tag{2.14}$$

Where $\mathbf{x_j}$ denotes the embedding of the associated word w_j, $\mathbf{b}^{(.)}$ denotes the bias, $\mathbf{W}^{(.)}$ and $\mathbf{U}^{(.)}$ denotes the parameter matrices for respective gates. Unlike traditional LSTM, child-sum tree LSTM has multiple forget gates, one for each child node. It allows each child node to incorporate the information selectively. Forget gate for the k^{th} child of the node j is defined as follows.

$$\mathbf{f}_{jk} = \sigma\left(W^{(f)}\mathbf{x}_j + \mathbf{U}^{(f)}\mathbf{h}_k + \mathbf{b}^{(f)}\right) \tag{2.15}$$

The final cell state and hidden state of node j are defined as follows.

$$\mathbf{c}_j = \mathbf{i}_j \odot \mathbf{u}_j + \sum_{\mathbf{k} \in \mathbf{ch}(j)} \mathbf{f}_{jk} \odot \mathbf{c}_k \tag{2.16}$$

$$\mathbf{h}_j = \mathbf{o}_j \odot tanh(\mathbf{c}_j) \tag{2.17}$$

The hidden state of the root node defines the encoding of the sentence.

This thesis used LSTM to encode news headlines, sentences, paragraphs and news bodies. Similarly, it also considered tree-LSTM to encode headlines and sentences of news bodies.

2.2.3 Transformer

The Transformer model [58] is a well-known deep learning architecture with extensive use across domains like natural language processing, computer vision, and speech processing. The first Transformer model was introduced as a sequence-to-sequence model [59] for machine translation, and subsequent research has revealed that Transformer-based pre-trained models [60] can achieve state-of-the-art performance across various tasks. Consequently, the Transformer architecture has become the preferred choice in NLP, particularly for pre-trained

models. Apart from applications of the Transformer in natural language processing, the Transformer has also been embraced in [61–63] audio processing [64–66] and even in other fields such as chemistry [67] and life sciences [68].

As reported in studies, Transformer based pre-trained language model is effective in several natural language processing tasks [69, 70, 60, 71], such as natural language inference [72, 73] and paraphrasing [74]. Motivated by the above observations in the literature regarding the performance of Transformer based pre-trained language models, this thesis uses two transformers-based pre-trained *Bidirectional Encoder Representations from Transformers* (BERT) and Sentence-BERT (S-BERT) to encode the sentences of the news articles. The main reason behind considering the BERT and S-BERT to encode the sentences of news articles is that transformers-based pre-trained BERT and Sentence-BERT S-BERT help to capture long-term dependencies between words of sentences or phrases. The model and implementation details of BERT and S-BERT can be found in study [75] and study [76], respectively. This dissertation considers BERT to encode headlines, sentences of news body and news body. Similarly, this study also considers S-BERT to encode headlines and sentences of the news body.

Within the domain of literature studies [24, 25, 40, 77, 1, 27, 26], researchers explore sequential encoding-based models utilizing techniques such as Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), and Long Short-Term Memory (LSTM) for encoding both headlines and various segments of news bodies. However, as evidenced in studies by [78, 41, 79], recursive encoding of sentences using Tree-LSTM has been shown to capture long-term dependencies among words and the syntactic structure of sentences. These studies have also noted enhanced performance across various applications, including determining semantic relatedness of sentence pairs, sentiment classification, and natural language interface tasks. Inspired by these findings regarding Tree-LSTM, we incorporate the child-sum Tree LSTM approach proposed by [41] in Chapter 4 of this thesis to encode both headlines and sentences within news articles. Transformer-based encoding models such as BERT, sentence transformer, and sentence BERT have gained popularity due to their ability to capture long-term dependencies between words of sentences while encoding the sentences. Motivated by such observations, we consider BERT [80] and sentence transformer [76] to encode sentences of headline and news body in Chapter 3 and Chapter 5.

2.3 Attention Mechanism

The concept of emulating human attention into deep learning models was first introduced in computer vision [81, 82]. Introducing attention was an effort to simplify the computational demands of image processing, enhancing performance by introducing a model that would selectively focus on specific image regions rather than the entire picture. However, the true origins of the attention mechanisms we are familiar with today are often traced back to the field of natural language processing [83]. In this context, the study [83] integrated attention into a machine translation model to tackle specific challenges associated with recurrent neural network structures. Following the study [83], which underscored the benefits of attention, the techniques related to attention underwent refinement [84]. Subsequently, these techniques gained rapid popularity across a diverse range of tasks. These tasks encompass text classification [85, 86], image captioning [87, 88], sentiment analysis [89, 90], as well as speech recognition [91, 92]. Attention [58] methods basic building block of recently transformer based pre-trained language models. This thesis used scaled [58], multi-head attention [58] and additive attention [83] in different contributions made in this thesis with the following objective. (i) Select the sentence of a headline which is contextually similar to the headline. (ii) Highlight the outlier sentences within the news body.

2.3.1 Scaled Attention

Given a set of queries, keys and values in the form of a query matrix $\mathbf{P}^q$, keys matrix $\mathbf{P}^k$ and value matrix, $\mathbf{P}^v$ the scaled dot product attention is defined by equation 2.18, 2.19 and 2.20 below.

$$\mathbf{M} = \left(\frac{\mathbf{P}^q \cdot (\mathbf{P}^k)^\top}{\sqrt{\mathbf{z}}} \right) \tag{2.18}$$

Where $\mathbf{z}$ is the dimension of queries and keys. $\mathbf{M}$ is a similarity matrix where $\mathbf{M}[\mathbf{i}, \mathbf{j}]$ represents the similarity between $\mathbf{i}^{th}$ query and $\mathbf{j}^{th}$ keys. Subsequently, the Softmax is applied over the similarity matrix $\mathbf{M}$ to obtain a probability distribution matrix $\mathbf{A}$ using equation as defined below 2.19.

$$\mathbf{A}_{i,j} = \left(\frac{\exp(\mathbf{M_{i,j}})}{\sum_{k,l} \exp(\mathbf{M_{k,l}})} \right) \tag{2.19}$$

Next, the final representation **p** is obtained by multiplying the probability distribution matrix **A** and value matrix **V** using the equation 2.20 as defined below.

$$\mathbf{p} = \left(\mathbf{A} \cdot \mathbf{P}^{v}\right) \tag{2.20}$$

2.3.2 Multi-head Attention

Given a set of queries, keys and values in the form of a queries' matrix **Q**, keys matrix **K** and value matrix, **V** first we estimate similarity between queries and keys by equation 2.21 as defined below.

$$\mathbf{P}_c^q, \mathbf{P}_c^k, \mathbf{P}_c^v = \mathbf{Q} \cdot \mathbf{W}_c^q, \mathbf{K} \cdot \mathbf{W}_c^k, \mathbf{V} \cdot \mathbf{W}_c^v \tag{2.21}$$

Where $\mathbf{W}_c^q$, $\mathbf{W}_c^k$ and $\mathbf{W}_c^v$ are learnable parameter matrices of query, key and value projections respectively, for c^{th} attention head of multi-head self attention and $\cdot$ is the dot product between matrix. Subsequently, attention weigh $\mathbf{A}_c$ is defined as follows:

$$\mathbf{M} = \left(\frac{\mathbf{P}_c^q \left(\mathbf{P}_c^k\right)^{\top}}{\sqrt{\mathbf{z}}}\right) \tag{2.22}$$

$$\mathbf{A}_{c,i,j} = \left(\frac{\exp(\mathbf{M_{ij}})}{\sum_{k,l} \exp(\mathbf{M_{k,l}})}\right) \tag{2.23}$$

Here M is matching matrix and $\mathbf{A}_c$ is attention weight matrix of c^{th} attention head. $\mathbf{A_c}[\mathbf{i},\mathbf{j}]$ entry represents the similarity probability between i^{th} query and j^{th} key. **z** is the dimension of $\mathbf{P}_c^q$. Next, we weighted summation is applied over value and attention weight matrix.

$$\mathbf{u}_{c,i} = \left(\sum_{j=1, i \neq j}^{k} \mathbf{A}_{c,ij} \mathbf{P}_{c,i}^{v}\right) \tag{2.24}$$

Now we concatenate the representation obtained by different attention heads and pass it to a dense layer to obtain the final sentence representation **U**.

$$\mathbf{U} = \left(\mathbf{U}_1 \oplus \mathbf{U}_2 \oplus .. \mathbf{U}_c \oplus . \oplus \mathbf{U}_l\right) \mathbf{W}_u \tag{2.25}$$

Where $\mathbf{W}_u$ is the trainable parameter matrix and $\mathbf{U}_c$ is c^{th} attention head. **U** is the representation matrix obtained by concatenating the representation of c^{th} attention head.

In Chapter 5, this thesis employs multi-head attention to encode sentences and derive a dual summary of the news body by exploring the interrelationships between sentences. Likewise, Chapter 6 adopts a two-step approach. Firstly, it represents the headline and news body as *n*-grams networks, followed by the construction of subgraphs for each node in the headline *n*-grams network. These subgraphs are generated by extracting the k-hop neighborhood of the best matching node in the *n*-grams network of the news body. Subsequently, multi-head attention is applied between the headline *n*-grams network and these subgraphs to capture the contextual relationships between the headline and various segments (subgraph extracted from *n*-grams network of news body) of the news body.

2.4 Graph Neural Network

Deep learning has transformed various machine learning tasks, ranging from image recognition and speech understanding to natural language processing. While these tasks often involve data represented in Euclidean spaces, there are growing applications where data exists in non-Euclidean forms, taking the shape of graphs with intricate relationships and dependencies among entities. This shift has posed significant challenges to conventional machine-learning techniques. Recently, numerous studies have emerged that extend deep learning methods to accommodate graph data. These graph-structured data present complexities that have yet to be encountered in traditional data forms. Researchers have introduced innovative approaches that bridge deep learning with graph data to address this. Graph Neural Networks (GNNs) stand out among these approaches. GNNs operate by adapting the principles of convolutions and aggregations to graph-like structures. They enable nodes within a graph to acquire representations by considering their neighbouring nodes' characteristics, proving effective in various graph-related tasks. Central to GNNs is the concept of message passing, whereby nodes iterative exchange information with their neighbours. This iterative process empowers nodes to combine insights from nearby entities and update their own features accordingly. GNNs have been instrumental in tasks such as graph classification and regression, where entire graphs are assigned labels or continuous values. They also excel in node-centric jobs like node classification and link prediction, which are crucial in fraud detection and network analysis domains. Graph Attention Networks (GATs) are a specialized type of GNN that employ attention mechanisms to intelligently prioritize neighbors during information aggregation. This adaptive information propagation enhances the flexibility and contextual awareness of GNNs. Beyond their applications, GNNs and their counterparts, like graph autoencoders and generative models, have encountered challenges unique to graph data. These challenges include handling the irregular nature of graphs, efficiently managing large and sparse graph structures, and maintaining computational efficiency. The realm of applications for graph-based deep learning spans diverse fields, encompassing social networks, bioinformatics, recommendation systems, and more. As these techniques continue to evolve, they enable the resolution of intricate problems that involve interconnected data points, reshaping the landscape of machine learning and AI. This thesis proposed graph context matching based methods which represent news headlines and body in the form of a ngram network and then match the context of headline and context of headline-centric context to the different contexts of news body. Our proposed graph context matching-based methods use graph neural network methods to update the information on each node in headline and body ngram network-based context of neighbour node and local neighbourhood structure.

This section of the thesis presents a graph neural network used in different proposed methods to obtain the encoding of news headlines and body bigram networks.

Graph Neural Network (GNNs) have emerged as a powerful framework for performing various machine learning tasks on graph-structured data. Graphs are versatile data structures that represent relationships between entities. These relationships are typically captured through nodes (vertices) and edges (connections between nodes), making graphs an apt representation for a wide range of real-world scenarios, including social networks, citation networks, biological systems, and recommendation systems. Traditional deep learning models, designed primarily for grid-like data such as images or sequences, struggle to effectively capture the inherent structure and connectivity present in graphs. To address the above limitations of traditional deep learning models, several graph neural networks have been proposed in the literature. In this thesis, we briefly discuss two graph neural networks, namely Graph Convolutional Neural Networks (GCNs) and Graph Attention Neural Network (GAT). Both GCNs and GAT are based on the principle of message passing methods. Message Passing Neural Networks (MPNNs) stand as a compelling intersection of graph theory and deep learning, offering a versatile framework for learning representations from graph-structured data. This section of the thesis explores the core concepts, design principles, applications, and advancements of MPNNs, shedding light on their pivotal role in capturing complex relationships and enabling predictive modeling within various domains. In graph-based data, entities are represented as nodes, and relationships as edges. The crux of MPNNs lies in their ability to propagate information between connected nodes, simulating a process akin to messages being exchanged among neighbors. These messages capture local context and are aggregated to update node representations. This iterative process allows MPNNs to distill both node-specific attributes and relational dependencies, encapsulating them into refined embeddings. The message passing neural networks (MPNNs) encompass several fundamental components:

1. **Message Function:**This function computes messages based on the sender and receiver nodes' attributes and the edge connecting them. It encodes the sender's information and imparts it to the receiver, encapsulating contextual knowledge.

2. **Aggregation Function:** The aggregation function merges received messages to create a refined representation for each node. This function considers the neighborhood's influence, facilitating richer information fusion.

3. **Update Function:** The update function combines the aggregated message with the node's current representation to generate an updated embedding. This step allows nodes to integrate the propagated information into their own context.

2.4.1 Graph Convolutional Neural Networks (GCNs)

The core functionality of GCNs lies in the graph convolutional layer, which generalizes the concept of convolution from grid-like data to graphs. The convolution operation in GCNs involves aggregating information from neighboring nodes and updating the representation of the central node. Mathematically, this operation can be expressed as a weighted sum of the feature vectors of neighboring nodes. The Graph Convolution operation incorporates the essence of the message-passing paradigm explained earlier. The two primary operations involved in Graph Convolution Neural Networks, namely aggregate and update. The primary motivation behind GCN is to update information on the embeddings of a node in a graph based on the information on the embeddings of neighbors nodes. The aggregate operations collect and aggregate the information from neighbors nodes and then update operation updates the embeddings or information of nodes based on neighbors nodes' information aggregated by the aggregate function. The aggregate and update operations are represented by equation 2.27. Initially, we assign a node feature $\mathbf{x}_v$ is assigned as initial node embedding $\mathbf{h}_v$ to each node's v in the graph based on the domain and underlying characteristics of the graph (as defined in Equation 2.26). However, the initial node feature is unavailable. In that case, we assign a random vector to each node and learn the local neighbour information and structure by updating the node embeddings based on the neighbour information. Next, information or embedding of each node is updated by considering information or embedding of neighbour nodes as defined in Equation 2.27.

$$\mathbf{h}_v^0 = \mathbf{x}_v \quad \forall v \in \mathcal{V} \tag{2.26}$$

Where $\mathbf{x}_v$ is the node feature of node v and $\mathbf{h}_v$ is the initial node embedding of node v.

$$\mathbf{h}_v^{(k)} = \mathbf{f}\left(\mathbf{W}_n^{(k)} \cdot \frac{\sum_{u \in \mathcal{N}(v)} \mathbf{h}_u^{(k-1)}}{|\mathcal{N}(v)|} + \mathbf{W}_m^k \cdot \mathbf{h}_v^{(k-1)} \right) \forall v \in \mathcal{V} \tag{2.27}$$

Where $\mathbf{W}_n^{(k)}$ and $\mathbf{W}_m^k$ are a learnable parameter matrix, $\mathbf{f}$ is the update function and $|\mathcal{N}(v)|$ is the number of neighbour for node v.

2.4.2 Graph Attention Network (GAT)

Though GCNs is an effective method for learning embeddings of nodes in a graph by local neighbour structure and context of nodes in a neighborhood, one key limitation of GCNs

is that it gives equal weight or importance to information from all the nodes in neighbour while updating the knowledge of node in Equation 2.27. However, in some domains, it is crucial that information from some neighbour nodes should be given high importance, and information in some neighbour nodes should be given less importance.

$$\mathbf{h}_v^0 = \mathbf{x}_v \quad \forall v \in \mathcal{V} \tag{2.28}$$

$$\mathbf{h}_v^{(k)} = \mathbf{f}\left(\mathbf{W}_n^{(k)}\left[\cdot \sum_{u \in \mathcal{N}(v)} \alpha_{vu}^{(k-1)} \mathbf{h}_u^{(k-1)} + \alpha_{vv}^{(k-1)} \cdot \mathbf{h}_v^{(k-1)}\right]\right) \forall v \in \mathcal{V} \tag{2.29}$$

$$\alpha_{vu}^k = \left(\frac{\mathbf{A}^{(k)}(\mathbf{h}_v^{(k)}, \mathbf{h}_u^{(k)})}{\sum_{w \in \mathcal{N}(v)} \mathbf{A}^{(k)}(\mathbf{h}_v^{(k)}, \mathbf{h}_u^{(k)})}\right) \quad \forall v \in \mathcal{V} \tag{2.30}$$

For example, consider a graph of words in a document where words in the document are a node and word embedding of the word is assigned as the initial embedding of the node. Next, while updating the embedding of nodes based on the information of the neighbour, it makes sense to give more importance to neighbour nodes which are highly contextually similar and give less importance to neighbour nodes which are less contextually similar. So, instead of giving equal weights to information from all the neighbour nodes, GAT assigns a unique weight α_{vu}^k to each node u in the neighbour of node V based on the similarity between u and v $u \in \mathcal{N}(v)$ (as defined in Equation 2.30).

In Chapter 6, the initial step involves representing both the headline and news body as n-grams networks. Each node in these networks is initialized with fastText embeddings of the words associated with the respective nodes [47]. Subsequently, a graph attention network [93] is applied over the n-grams network of the headline and the n-grams network of the news body. This process updates the node embeddings based on the context of neighboring nodes and the local neighborhood structure.

Summary

This chapter presents a comprehensive overview of all the essential background information required to comprehend the contributions made in this thesis. We discuss (i) word embedding methods used in this thesis to represent words,(ii) different sequential encoding and transformer-based models used to encode headlines, sentences, paragraphs and news bodies, and (iii) Several attention mechanisms are used to highlight sentences and paragraphs of the news body with respect to the headline.

Chapter 3

Datasets

With the increasing concerns of disinformation shared over digital platforms, detecting fake news articles in resource-poor languages is becoming an important research problem. While several datasets are under circulation in the public domain for the English language for fake news detection research, such datasets are not readily available for resource-poor languages. In this thesis, we curate and propose four types of large-scale hybrid (real samples and fake synthetic samples) *Hindi* datasets, suitable for fake news detection research in news articles from different content and linguistic aspects for public access. Though few small-scale Hindi datasets for fake news detection are reported in the literature, they are neither readily available nor linguistically annotated. Appropriate annotation is important for developing a linguistically complex model and explainability study. The quality and reliability of the proposed datasets are further evaluated using different state-of-the-art methods over real fake news samples.

3.1 Introduction

Detecting fake news articles has evolved as an important research problem to handle the early detection of misinformation on digital platforms [3] [2] [94], including publications in vernacular or resource-poor languages. A news article is said to be fake if it presents intentionally created false information to mislead its readers. Fake news articles are created with several motivations[1], such as political polarization, propaganda, diverting the issue and spreading hate against an individual, organization or community. While several datasets

[1] Types of Fake News Under circulations

are under circulation in the public domain for the English language for fake news detection research, such datasets are not readily available for resource-poor languages, Indian languages in particular. In this thesis, we curate and propose four types of large-scale hybrid (real samples and synthetic fake samples) datasets for *Hindi language*, suitable for fake news detection research from different content and linguistic aspects for public distribution. Based on the 2011 census, a total of 52.8 Crore Indians speak Hindi as their first language, and about 13.9 Crore as second language (it is about 55% of the total Indian population). Annual Statements, Press in India, 2020-21 indicates that there are 4,349 dailies with a total circulation of 10,36,19,621 copies in India. As per recent report of National Crime Records Bureau[2], in India there is a 214% hike in crimes related to fake news[3].

Curating real fake news articles in circulation on a large scale is a challenging and expensive task. Further, the popular datasets like NELA [28], FNC [22] in English languages for fake news detection research are also created synthetically. Considering that there are almost no publications of Hindi dataset in the public domain for fake news detection research (either synthetic or real), our dataset is a first of its kind. Further, it not only adapts the strategies proposed in NELA and FNC, it also incorporates two more linguistic strategies (POS and NE replacement) to allow the researchers to explore not only the content, but also the linguistic aspects. The proposed datasets are further evaluated with state-of-the-art methods using explicitly extracted features, and also with neural embedded features.

3.2 Datasets in Literature

3.2.1 Dataset for English Language

There are three publicly available datasets, namely ISOT (Information Security and Object Technology) fake news dataset [95, 96], FNC (Fake News Challenge) dataset [22], and NELA-17 (News Landscape) dataset [97, 28]. Table 3.1 presents the characteristics of these datasets. Study [96, 95] curate the ISOT dataset by compiling news articles published by reliable and unreliable sources. News articles published by reliable sources are labelled as *True* class samples, and news articles published by unreliable sources are labelled as *Fake* news. The FNC dataset has four classes: agree, disagree, discuss, and unrelated. The samples in agree, disagree and discuss classes are merged and named as *True* class, whereas the samples in unrelated classes are considered as *Fake* class. *Fake* class samples of the FNC

[2] National Crime Records Bureau, India report
[3] Fake news challenge in India

Table 3.1 Characteristics of Experimental Datasets

	Dataset	True	Fake	Total	#Head	#Body	#Para	#Sen
ISOT	Train	17083	18232	35315	9.438	244.325	3.799	16.955
	Test	1726	1815	5313	9.377	236.379	3.729	16.606
	Dev	2607	2706	3541	9.388	241.136	3.733	16.607
FNC	Train	12057	32917	44974	8.478	217.216	11	19.465
	Test	7064	18349	25413	8.503	213.757	10.523	18.744
	Dev	1370	3628	4998	8.465	216.347	10.808	19.215
NELA-17	Train	35710	35710	71420	10.558	551.923	13.494	26.649
	Test	3151	3151	6302	10.529	566.921	13.851	27.526
	Dev	3151	3151	6302	10.547	541.188	13.49	26.256

dataset are generated by mismatching the headline and body of authentic news articles from different topics [98]. So in the case of an *Fake* class sample of the FNC dataset, though the headline and body are different, the content within the news article body is coherent. Hence, an *Fake* class sample of the FNC dataset is considered a fully incongruent news article. Similarly, a *True* class sample of the FNC dataset is considered a fully congruent news article. The NELA-17 dataset has been widely used for misinformation detection tasks, including incongruent news detection studies [28, 30, 37]. We also curate the NELA dataset following the procedure reported in [28] [4] over the news corpus provided at, [97] [5]. The news articles published by authenticated sources are labelled as *congruent* class, and *incongruent* news articles are generated by inserting noisy contents in congruent news articles. The noisy contents are paragraphs randomly selected from some other news article. Considering the procedure of incongruent sample generating in the NELA dataset, it is evident that incongruent samples in the NELA dataset are partially incongruent. Because every paragraph except the noise (randomly selected paragraph) which is inserted will be congruent with the headline, so, an incongruent class sample of the NELA dataset represents the true nature of a partially incongruent news article.

3.2.2 Datasets for Hindi Language

Initial studies, [99, 100], compiled hostile and fake news datasets in social media for the Hindi language. Further studies [101, 102] attempt to compile datasets for fake news article

[4] NELA-17 dataset generation code
[5] NELA-17 Dataset corpus

detection in the Hindi language by collecting fake news articles from fact-check websites. But these datasets are publicly not available for research and development purposes. These datasets have a small sample size and lack topic and event diversity. Hence, they are insufficient to train robust state-of-the-art models for fake news article detection. Compiling large-scale human labelled datasets is a challenging task [103–105]. Considering such challenges and limitations, this study curated and proposed four types of large-scale datasets for fake news article detection in the Hindi language. Our proposed dataset curation methods could be used to curate a large-scale dataset for fake news article detection in any vernacular or resource-poor language.

3.3 Proposed Datasets

Headline	नेटो की आपूर्त‍ि को लेकर पाक-अमरीका में समझौता
True Body	11 सतिंबर 2001 के बाद अफ़ग़ानस्‍तिान में नेटो सेना के लिए सामान की आपूर्त‍ि पर इस प्रकार का पहला लखिति समझौता हुआ है.इस्‍लामाबाद स्‍थति अमरीका के वरष्‍ठि राजनयकि रचिर्ड होगलैंड और पाकस्‍तिान के रक्षा मंत्रालय के अतरिक्‍ति सचवि एडमरिल फारुख़ ने इस समझौते पर दस्‍तख़त कए.अफ़ग़ानस्‍तिान में मौजूद नेटो सेना के लिए सामान की आपूर्त‍ि को लेकर इससे पहले कोई लखिति समझौता नहीं था.समझौते पर हस्‍ताक्षर के बाद अमरीकी राजनयकि रचिर्ड होगलैंड ने पत्रकारों से बातचीत करते हुए कहा क‍ि लखिति समझौते के बाद पाकस्‍तिान को 'कोअलशिन स्‍पोर्ट फंड' के मद में तुरंत एक अरब डॉलर की वत्‍तिीय सहायता दी जाएगी.गौरतलब है क‍ि दोनों देशों के बीच तनाव के चलते पाकस्‍तिान को क़रीब दो सालों से इस मद में वत्‍तिीय सहायता नहीं दी गई है.इस समझौते को फ़लिहाल सार्वजनकि नहीं कयिा गया है और लेकनि दोनों पक्षों ने बताया है क‍ि इसको जल्‍द ही सार्वजनकि कयिा जाएगा.पछिले साल 26 नंबर को क़बायली इलाक़े मोहमंद एजेंसी में नेटो हैलकॉप्‍टरों ने पाकस्‍तिानी सेना की चौकी पर हमला कयिा था, जसिमें 24 पाकस्‍तिानी सैनकि मारे गए थे और 15 अन्‍य घायल हो गए थे.पाकस्‍तिानी सरकार ने उस हमले का कड़ा वरिोध कयिा था और अफ़ग़ानस्‍तिान में नेटो सेना के लिए सामान आपूर्त‍ि पर रोक लगा थी.भारी अंतरराष्‍ट्रीय दबाव और अमरीका और पाकस्‍तिान के बीच हुई बातचीत के बाद पाकस्‍तिान ने क़रीब सात महीनों बाद उसपर लगी रोक हटा दी थी.
Named Entity Replacement (NE-R)	डपिार्टमेंट ऑफ़ स्‍कूल एजुकेशन एंड लटिरेसी के बाद तीन इमली सुरंगवायर में डायरेक्‍टरेट जनरल ऑफ सेंट्रल एक्‍साइज इंटेलजिंस के लिए सामान की आपूर्त‍ि पर इस प्रकार का पहला लखिति समझौता हुआ है.भीमावरम स्‍थति थोक के वरष्‍ठि राजनयकि राचनु और दाड़लाघाट के रक्षा मंत्रालय के अतरिक्‍ति सचवि रामवलिाश ने इस समझौते पर दस्‍तख़त कए.तीन इमली सुरंगवायर में मौजूद डायरेक्‍टरेट जनरल ऑफ सेंट्रल एक्‍साइज इंटेलजिंस के लिए सामान की आपूर्त‍ि को लेकर इससे पहले कोई लखिति समझौता नहीं था.समझौते पर हस्‍ताक्षर के बाद कंट्रोलर जनरल ऑफ डफिन्‍स एकाउंट्स (सगड़ा) राचनु ने पत्रकारों से बातचीत करते हुए कहा क‍ि लखिति समझौते के बाद दाड़लाघाट को 'कोअलशिन स्‍पोर्ट फंड' के मद में तुरंत एक अरब डॉलर की वत्‍तिीय सहायता दी जाएगी.गौरतलब है क‍ि दोनों देशों के बीच तनाव के चलते दाड़लाघाट को क़रीब दो सालों से इस मद में वत्‍तिीय सहायता नहीं दी गई है.इस समझौते को फ़लिहाल सार्वजनकि नहीं कयिा गया है और लेकनि दोनों पक्षों ने बताया है क‍ि इसको जल्‍द ही सार्वजनकि कयिा जाएगा.पछिले साल 26 नंबर को जमानवाव इलाक़े मोहड़लेक्‍ट्रॉनक्‍स कारपोरेशन ऑफ़ इंडयिा (केलि) में नेटो हैलकॉप्‍टरों ने दाड़लाघाटी सेना की चौकी पर हमला कयिा था, जसिमें 24 दाड़लाघाटी सैनकि मारे गए थे और 15 अन्‍य घायल हो गए थे.दाड़लाघाटी सरकार ने उस हमले का कड़ा वरिोध कयिा था और तीन इमली सुरंगवायर में डायरेक्‍टरेट जनरल ऑफ सेंट्रल एक्‍साइज इंटेलजिंस के लिए सामान आपूर्त‍ि पर रोक लगा थी.भारी नेशनल कमीशन फॉर बैकवर्ड क्‍लासेज दबाव और थोक और दाड़लाघाट के बीच हुई बातचीत के बाद दाड़लाघाट ने क़रीब सात महीनों बाद उसपर लगी रोक हटा दी थी.
Part of Speech Replacement (POS-R)	11 सतिंबर, 2001 के बाद अफगानस्‍तिान में नाटो सेनाओं को माल की आपूर्त‍ि के बारे में मध्‍यवर्त‍ी इस प्‍रकार की वचनबद्ध समझौता की गई है। इस समझौते को इस्‍लामाबाद में वरष्‍ठि अमेरकी राजनयकि रचिर्ड होगलैंड और पाकस्‍तिान रक्षा मंत्रालय के उपाध्‍यक्ष एडमरिल फौरूख ने हस्‍ताक्षर नही कए। इससे पहले अफगानस्‍तिान में नाटो सेनाओं को आपूर्त‍ि के बारे में कोई बातचीत में कोई सहमत‍ि नहीं थी। समझौते पर हस्‍ताक्षर करने के बाद अमेरकी राज्‍य सचवि रचिर्ड होगलैंड ने रपिोर्टरों से कहा क‍ि कथति समझौते के बाद पाकस्‍तिान को एक बलियिन डॉलर का गैर वत्‍तिीय सहायता को सहयोग नधि में शामलि कयिा जाएगा। यह उल्‍लेखनीय है क‍ि दोनों देशों के बीच तनाव के कारण पाकस्‍तिान को लगभग दो वर्ष तक इस मामले मे गैर वत्‍तिीय सहायता नहीं दी गई है। यह समझौता अभी तक नजिी नहीं हुआ है और दोनों पक्षों ने प्रतवािद कयिा है क‍ि यह शीघ्र ही नजिी नहीं होगा। पहली वर्ष 26 को एक नेतो हेलीकॉप्‍टर ने मोमान एजेंसी के काबायी जलि में पाकस्‍तिानी सेना का बचाव कयिा, जसिमें 24 पाकस्‍तिानी सैनकि मारे गए और 15 अन्‍य घायल हुए। पाकस्‍तिान सरकार ने इस हमले और अफगानस्‍तिान में नाटो सेनाओं को अनावश्‍यक आपूर्त‍ि का जोरदार वरिोध नहीं कयिा। अमेरकिा और पाकस्‍तिान के बीच हल्‍के राष्‍ट्रीय दबाव और बातचीत के बाद पाकस्‍तिान ने लगभग सात महीने बाद उस पर प्रतबिंध को कम कर दयिा। वास्‍तव में, यह ऐसा नहीं है मानो हम इसे छोड़ने में असमर्थ होने के लिए अपने स्‍थान पर रहे हैं।
Part of Speech and Named Entity Replacement (POS-NE-R)	11 सतिंबर, 2001 के बाद कवति में नाटो सेनाओं को माल की आपूर्त‍ि के बारे में मध्‍यवर्त‍ी इस प्‍रकार की वचनबद्ध समझौता की गई है। इस समझौते को कतसौर में वरष्‍ठि गरिगढ़ा राजनयकि अर्थी और इंडयिन इंस्‍टट्यिूट ऑफ़ एन्‍त्रेप्रेयूर्शपि रक्षा मंत्रालय के उपाध्‍यक्ष रूपलाल ने हस्‍ताक्षर नहीं कए। इससे पहले कवति में नाटो सेनाओं को आपूर्त‍ि के बारे में कोई बातचीत में कोई सहमत‍ि नहीं थी। समझौते पर हस्‍ताक्षर करने के बाद गरिगढ़ा राज्‍य सचवि अर्थी ने रपिोर्टरों से कहा क‍ि कथति समझौते के बाद इंडयिन इंस्‍टट्यिूट ऑफ़ एन्‍त्रेप्रेयूर्शपि को एक बलियिन डॉलर का गैर वत्‍तिीय सहायता को सहयोग नधि में शामलि कयिा जाएगा। यह उल्‍लेखनीय है क‍ि दोनों देशों के बीच तनाव के कारण इंडयिन इंस्‍टट्यिूट ऑफ़ एन्‍त्रेप्रेयूर्शपि को लगभग दो वर्ष तक इस मामले में गैर वत्‍तिीय सहायता नहीं दी गई है। यह समझौता अभी तक नजिी नहीं हुआ है और दोनों पक्षों ने प्रतवािद कयिा है क‍ि यह शीघ्र ही नजिी नहीं होगा। पहली वर्ष 26 को एक नेतो हेलीकॉप्‍टर ने मोमान एजेंसी के काबायी जलि में इंडयिन इंस्‍टट्यिूट ऑफ़ एन्‍त्रेप्रेयूर्शपि सेना का बचाव कयिा, जसिमें 24 इंडयिन इंस्‍टट्यिूट ऑफ़ एन्‍त्रेप्रेयूर्शपि सैनकि मारे गए और 15 अन्‍य घायल हुए। इंडयिन इंस्‍टट्यिूट ऑफ़ एन्‍त्रेप्रेयूर्शपि सरकार ने इस हमले और कवति में नाटो सेनाओं को अनावश्‍यक आपूर्त‍ि का जोरदार वरिोध नहीं कयिा। अमेरकिा और इंडयिन इंस्‍टट्यिूट ऑफ़ एन्‍त्रेप्रेयूर्शपि के बीच हल्‍के राष्‍ट्रीय दबाव और बातचीत के बाद इंडयिन इंस्‍टट्यिूट ऑफ़ एन्‍त्रेप्रेयूर्शपि ने लगभग सात महीने बाद उस पर प्रतबिंध को कम कर दयिा। वास्‍तव में, यह ऐसा नहीं है मानो हम इसे छोड़ने में असमर्थ होने के लिए अपने स्‍थान पर रहे हैं।

Fig. 3.1 An example of a fake news article body generated by NE-R, POS-R and POS-NE-R approaches. The headline and True Body constitute a real news article. Words and sentences replaced by respective dataset curation methods are marked in red. The English translation of this example is presented in Figure 3.2.

The underlying hypothesis behind the proposed fake news datasets are as follows.

- The content of the body of a fake news article should be coherent with regard to the underlying topic/theme.

Headline	Pak-US agreement on supplies to Nato
True Body	This is the first written agreement on the supply of goods to NATO forces in Afghanistan since September 11, 2001. Richard Hoagland, a senior US diplomat based in Islamabad, and the agreement was signed by Admiral Farooq, Additional Secretary, Ministry of Defense of Pakistan. Earlier, there was no written agreement regarding the supply of goods to the Nato forces present in Afghanistan. There was no agreement. American diplomat Richard Hoagland, while talking to reporters after the agreement was signed, said that after the written agreement, Pakistan will be given the 'coalition support fund'. Financial assistance of one billion dollars will be given immediately. It is worth noting that due to the tension between the two countries, Pakistan has not been given any financial assistance for the last two years. Till now, it has not been made public, but both sides have said that it will be made public soon. Last year on the 26th number of Nato helicopters in tribal area Mohmand Agency, there was an attack on a Pakistani army post, in which 24 Pakistani soldiers were killed, and 5 soldiers were injured. The Pakistani government strongly condemned the attack and, in Afghanistan, there was a ban on the supply of goods to the NATO army. After heavy international pressure and talks between the US and Pakistan, Pakistan lifted the ban on it after about seven months.
Named Entry Replacement (NE-R)	This is the first written agreement on the supply of goods to the Directorate General of Central Excise Intelligence forces in Teen Imli Square since September 11, 2001. Rachnu, a senior US diplomat based in Bhimavaram, and the agreement was signed by Ramvalash, Additional Secretary, Ministry of Defense of Darlaghat. Earlier, there was no written agreement regarding the supply of goods to the Central Excise Intelligence forces present in Teen Imli Square. There was no agreement. American diplomat Rachnu, while talking to reporters after the agreement was signed, said that after the written agreement, Darlaghat will be given the 'coalition support fund'. Financial assistance of one billion dollars will be given immediately. It is worth noting that due to the tension between the two countries, Darlaghat has not been given any financial assistance for the last two years. Till now, it has not been made public, but both sides have said that it will be made public soon. Last year on the 26th number of Central Excise Intelligence helicopters in tribal area Jamanwav Agency, there was an attack on a Darlaghati army post, in which 24 Darlaghati soldiers were killed, and 5 soldiers were injured. The Darlaghati government strongly condemned the attack and, in Teen Imli Square, there was a ban on the supply of goods to the Central Excise Intelligence army. After heavy international pressure and talks between the Bhari National and Darlaghat, Darlaghat lifted the ban on it after about seven months.
Part of Speech Replacement (POS-R)	This is the first spoken agreement on the supply of goods to NATO forces in Afghanistan since September 11, 2001. Richard Hoagland, a senior US diplomat based in Islamabad, and the agreement was not signed by Admiral Farooq, Additional Secretary, Ministry of Defense of Pakistan. Earlier, there was no written agreement regarding the supply of goods to the Nato forces present in Afghanistan. There was no agreement. American diplomat Richard Hoagland, while talking to reporters after the agreement was signed, said that after the written agreement, Pakistan will be given the 'coalition support fund'. Financial assistance of one billion dollars will be given later. It is worth noting that due to the tension between the two countries, Pakistan has not been given any financial assistance for the last two years. Till now, it has not been made personal, but both sides have said that it will be made personal soon. First year on the 26th number of Nato helicopters in tribal area Mohmand Agency, there was an attack on a Pakistani army post, in which 24 Pakistani soldiers were killed, and 5 soldiers were injured. The Pakistani government did not condemn the attack and, in Afghanistan, there was a ban on the supply of goods to the NATO army. After heavy international pressure and talks between the US and Pakistan, Pakistan did not lift the ban on it after about seven months.
Part of Speech and Named Entity Replacement (POS-NE-R)	This is the first spoken agreement on the supply of goods to the NATO forces in Kavita since September 11, 2001. Arthi, a senior US diplomat based in Garigdhara, and the agreement was not signed by Rooplal, Additional Secretary, Ministry of Defense of Indian Institute of Entrepreneurship. Earlier, there was no written agreement regarding the supply of goods to the NATO forces present in Kavita. There was no agreement. American diplomat Rachnu, while talking to reporters after the agreement was signed, said that after the written agreement, Indian Institute of Entrepreneurship will be given the 'coalition support fund'. Financial assistance of one billion dollars will be given later. It is worth noting that due to the tension between the two countries, the Indian Institute of Entrepreneurship has not been given any financial assistance for the last two years. Till now, it has not been made personal, but both sides have said that it will be made personal soon. First year on the 26th number of NATO helicopters in tribal area Jamanwav Agency, there was an attack on an Indian Institute of Entrepreneurship army post, in which 24 Indian Institute of Entrepreneurship soldiers were killed, and 5 soldiers were injured. The Indian Institute of Entrepreneurship government strongly did not condemn the attack and, in Kavita, there was a ban on the supply of goods to the NATO army. After heavy international pressure and talks between the America and Indian Institute of Entrepreneurship, Indian Institute of Entrepreneurship did not lift the ban after about seven months.

Fig. 3.2 The English translation of the example is presented in Figure 3.1. An example of a fake news article body generated by NE-R, POS-R and POS-NE-R approaches. The headline and True Body constitute a real news article. Words and sentences replaced by respective dataset curation methods are marked in red.

- There should be noise embedded into the body, to make the reporting in the articles fake.

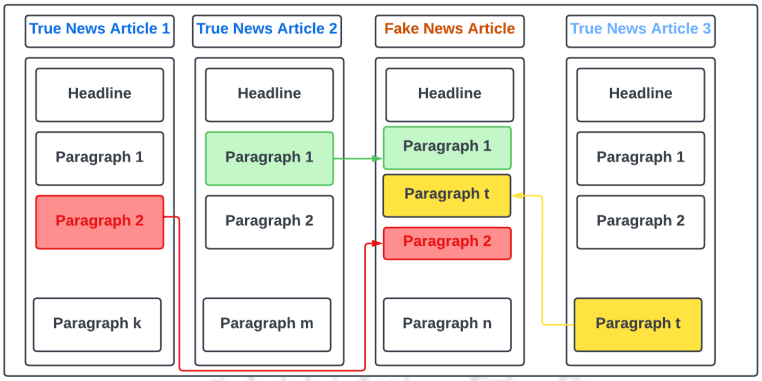


Fig. 3.3 Schematic diagram of split and merge method for fake news article body generation.

- The embedded noise should be linguistically valid.

We propose four types of large-scale hybrid datasets; real news articles for true class, and synthetic articles for fake class. First, we crawled newspapers published by authentic media houses between 2002 and 2021 in the Hindi language and considered them as real samples. To generate a fake news article sample, we propose the following four different strategies, namely (i) split and merge, (ii) named-entity replacement, (iii) part-of-speech replacement, and (iv) part-of-speech and named-entity replacement. The proposed datasets draw inspiration from the patterns observed in real fake news circulating in the Hindi language. The Split and Merge (SM) dataset is motivated by the fake sample generation method employed in the NELA dataset for English fake news article detection. The Named Entity Replacement (NE-R) dataset is constructed based on the observation that fake news often involve modifications to named entity information before being circulated on digital platforms. Similarly, the Part of Speech Replacement (POS-R) approach is informed by real-world instances of fake news articles where certain parts of the news body are altered to contradict the actual claim of the news, thereby resembling true news but with a divergent viewpoint. These tactics are often employed to propagate propaganda and malicious campaigns, aiming to confuse and mislead readers.

3.3.1 Split and Merge Approach (SM)

This approach is motivated by the NELA-17 dataset [28]. It splits a true news article into paragraphs, and then the fake news articles are generated by inserting a paragraph selected randomly from a random news article into a true news article. As reported in [28], NELA-17 may have false-negative samples if the randomly selected article is similar to the article

in which the selected paragraph is inserted. To reduce this effect, we propose to choose more paragraphs from more articles. In this study, we consider randomly selected three articles, and select one random paragraph from each of these articles. Figure 3.3 illustrates a schematic diagram of the proposed *split and merge* approach.

3.3.2 Named-Entity Replacement (NE-R)

If we consider a news reporting of an event, replacing the associated entities with another entity may make the reporting fake. For example, claim made in news article[6] *Corona affects Hindus, Muslims have dua, don't need vaccine: Kolkata's Maulana Barkati*, if name Maulana Barkati is replaced by another entity World Health Organization, then it becomes fake news of serious concern. Another similar incident was reported by news article [7] where the location of Hyderabad was changed to Assam to create and spread fake news. Motivated by this, the proposed named-entity based fake news articles are generated by impersonating individual, personality, celebrity or organization. Given a news article consisting of a pair of headline $\mathcal{H}$ and body $\mathcal{B}$, let $\mathcal{E}$=$\{\mathcal{E}_1 \cup \mathcal{E}_2 \cup \cdots \cup \mathcal{E}_k\}$ be the set of named entities present in $\mathcal{H}$ and $\mathcal{B}$. The $XLM-Roberta-base$[8] [106] is used to parse the document and extract the named entities. Once $\mathcal{E}$ is obtained, the named entities in $\mathcal{E}$ are replaced with randomly selected matching named entities from Indian name dataset[9] person with person, location with location[10], and organization with organization[11] [12] [13]. Once all named entities are replaced, the resultant articles are labelled as fake. Figure 3.4 illustrates a schematic diagram of the proposed approach. Instead of replacing all the entities, replacing few of them may also be sufficient to make a news fake. Considering the diverse nature of the articles curated in our dataset, we replaced all the entities to increase the likelihood of being fake after replacement.

3.3.3 Part of Speech Replacement (POS-R)

Given a news article, if we replace some of the adjectives or verbs by their antonyms, the resultant article may represent fake news. For example, the first paragraph of the report

[6] Named entity example
[7] Named entities replacement real exmaple
[8] XLM-Roberta-base model setting from Hugging face for NER
[9] Indian names dataset
[10] List of cities in India
[11] List of government organizations in India
[12] List of companies names in India
[13] International organizations name list

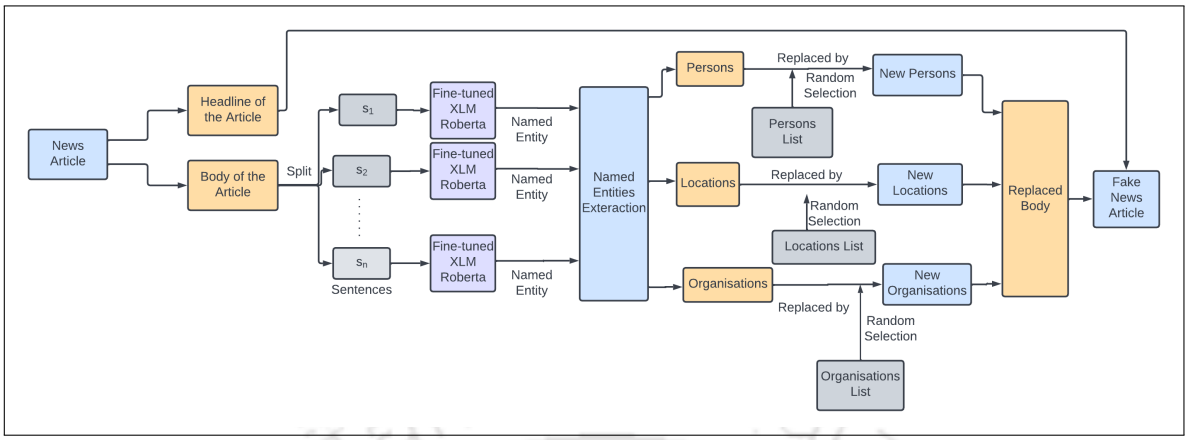


Fig. 3.4 Named Entity Replacement Approach

claims [14], *The two mRNA vaccines, Pfizer and Moderna, authorized by the U.S. Food and Drug Administration (FDA) and recommended by the Centers for Disease Control and Prevention (CDC), are very safe and very good at preventing serious or fatal cases of COVID-19.* Here, *safe* and *good* are adjectives. If these adjectives are replaced with their antonym *harmful* and *bad* respectively, this becomes fake news. This approach generates fake news articles by altering the adjectives and verbs present in the articles with their antonyms. For this task, we need language tools such as parser, POS tagger, Wordnet etc. for Hindi language. Though some of these tools are reported in literature, considering the challenges of procuring them for offline processing, we have relied on a more convenient approach, i.e., translate the documents in English, perform the required prepossessing in English, and then translate back to Hindi. The schematic diagram of the process is presented in Figure 3.5. We use the mBART-50[15] many to many multilingual machine translation model [107] for translating the document back and forth between Hindi and English. The translated sentences in English are passed to Stanza [108] to extract part of speech information (adjective and verbs). The antonyms of the extracted adjectives and verbs are obtained using NLTK WordNet Interface[16]. We consider the closest antonym as provided by the tool as the candidate for replacement. Though we do not need to replace all the verbs and adjectives, like in NER, but we replace every verb and adjectives with their antonyms to increase the likelihood of being fake. The resultant document is translated back to Hindi using mBART-50.

[14] POS fake news example source
[15] XLM-Roberta-base setting Hugging face
[16] NLTK Wordnet

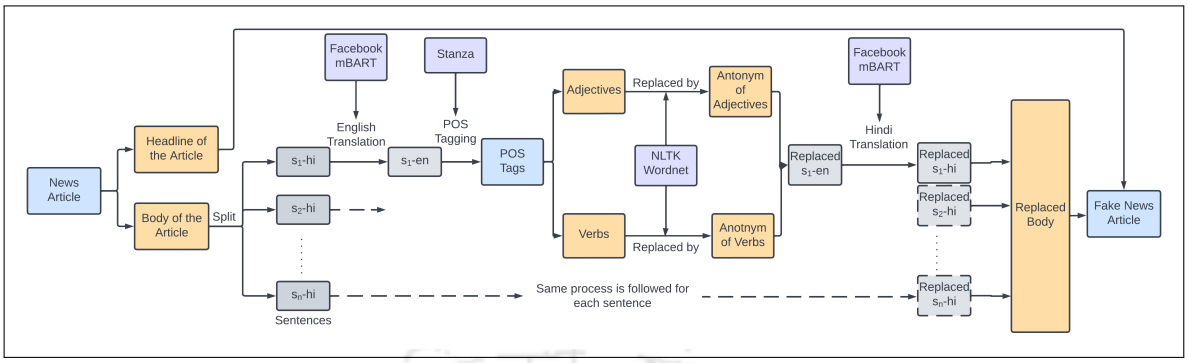


Fig. 3.5 Part of Speech Replacement Approach

3.3.4 Combine POS and NER Replacement (POS-NE-R)

This hybrid approach combines named-entity and part-of-speech replacement to generate the fake news. Given a news article with its headline and body, the adjectives and verbs are replaced with their antonyms, following the same procedure used in POS replacement approach. Then the persons, organizations and locations named entities are replaced following the same procedure used in NE replacement approach.

3.3.5 Human Crafted Fake News and Real Fake News

The primary motivation behind curating a human-crafted fake news dataset and real fake news dataset is to validate and assess the quality and reliability of datasets generated in sections 3.3.1, 3.3.2, 3.3.3 and 3.3.4. The human-crafted fake news dataset and real fake news dataset are used only as a test set.

Human-Crafted Fake News: This dataset is built using news articles manually collected from dailies between March 2022 to May 2022. Each news article is manually verified by eight human annotators and labelled them as true sample. Fake news samples are further generated by manually editing true news articles. We manually edited the following information on news article. (i) change named entities mentioned in body, to similar class of named entity of body. (ii) modify the main claim of body to make contextually similar fake news.

Real Fake News Dataset: To understand and observe the patterns of real-world fake news under circulation, we curate a real fake news dataset by manually collecting the Hindi language real fake news under circulation from different digital platforms such as Facebook, Twitter, Reddit, Koo etc. News flagged as fake news by different fact-checking media houses (Opindia, Boom live, the Fauxty and News 18) were also used as sources. For true

news articles, news articles from reputed media houses were collected and verified by eight annotators.

3.4 Statistics of Datasets

We consider two publicly available news sources for curating the true news articles, namely BBC news corpus [17] and Navbharat Times (NAV) [18]. While the BBC distribution has only, 4335 articles, we crawl about 0.26 million articles published during 2002 to 2021 from Navbharat Times (NAV) [19]. The collected news articles from Navbharat Times are from different topics, including regional and international politics, business, sports, current affairs, entertainment and reporting. The documents are further cleaned by removing unwanted punctuations, HTML tags, images and links. Fake version of these news articles are then generated by applying the methods discussed in section 3.3. Table 3.2 presents the characteristics of the datasets.

3.5 Experimental Setups and Discussions

To evaluate the quality of proposed datasets, we consider the following baseline models and features.

Features: Considering the significance of several bag-of-words-based features reported in studies [109, 110, 35, 24, 111] for the English language fake news article detection, we use count Overlap (CO) and Term Frequency–Inverse Document Frequency (TF-IDF) to evaluate the effectiveness of our proposed datasets. The implementation details of these features are as follows:

1. **Count Overlap (CO):** Count overlap features count the number of common tokens in the headline and body of a news article. It measures the number of common tokens between the headline and body and helps to find similarities between them in terms of the number of common tokens. We first extract all unigrams, bigrams, and trigrams in the headline and body. Subsequently, we count the number of common unigrams, bigrams, and trigrams.

[17] BBC News Corpus
[18] Navbharat Times
[19] Navbharat Times old archive

Table 3.2 Characteristics of Experimental Datasets. (i) **#Head** denote average number of words in headline. (ii) **#Body** denote average number of words in body. (iii) **#Para** denote average number of paragraphs in body. (iv) **#Sen** denote average number of sentence in body

Data Set Curation Method	Corpus	Dataset	Total	True	Fake	#Head	#Body	#Para	#Sen
Split and Merge (SM)	NAV	Train	261511	130570	130941	7.077	204.637	4.827	15.595
		Test	72642	36550	36092	7.078	203.997	4.973	15.633
		Dev	29057	14485	14572	7.084	204.462	4.642	15.618
	BBC	Train	6242	3108	3134	7.427	383.64	5.756	21.851
		Test	1734	882	852	7.455	376.84	5.369	21.145
		Dev	694	2706	3541	9.388	241.136	5.713	22.407
Named Entity Replacment (NE-R)	NAV	Train	261511	130742	130333	7.218	436.291	4.812	16.036
		Test	72522	36368	36154	7.213	1030.11	4.656	16.151
		Dev	29009	14495	14514	7.234	368.184	4.356	16.064
	BBC	Train	6026	3121	2905	7.434	994.5	6.983	27.635
		Test	1674	876	798	7.464	1326.007	7.239	28.073
		Dev	694	345	349	7.567	374.051	7.496	29.437
Part of Speech Replacment (POS-R)	NAV	Train	226476	130785	95691	7.56	245.107	4.895	19.324
		Test	62909	36261	26648	7.412	244.747	4.921	19.381
		Dev	25166	14551	10615	7.250	244.209	4.891	19.159
	BBC	Train	6241	3127	3114	7.520	1670.80	5.140	16.612
		Test	1734	872	862	7.507	556.96	3.371	16.715
		Dev	649	349	345	7.6	512	5.126	17.175
Part of Speech and Named Entity Replacement (POS-NE-R)	NAV	Train	254792	130692	124100	7.255	212.128	4.998	19.324
		Test	70777	36387	34390	7.253	212.474	5.12	19.381
		Dev	28312	14521	13791	7.243	212.176	4.912	19.159
	BBC	Train	6241	3128	3113	7.55	222.37	5.081	16.838
		Test	1734	862	872	7.403	273.13	5.134	17.163
		Dev	694	349	345	7.599	719.18	4.952	17.42
Human crafted fake news	–	–	706	263	443	19.859	414.835	5.8	22.406
Real fake news	–	–	3984	1992	1992	9.656	291.86	3.021	10.066

2. **Term Frequency–Inverse Document Frequency (TF-IDF)** As reported in study [35] the TF-IDF features are effective for fake news classification. Subsequently, we compute the cosine similarity between the TF-IDF vector of the headline and the TF-IDF vector of the body. Finally, these three features, namely TF-IDF of headline, body and cosine similarity are fed to classifiers for fake news classification.

Classifiers: This study builds several classifiers over CO and TF-IDF to validate the response of different classifiers over proposed datasets. Decision Tree (DT) [112], Support Vector Machine (SVM) [113], over CO and TF-IDF features separately and then combined. Considering the superiority of ensemble learning for fake news classification task[20][114–117], this study utilizes ensemble learning methods, including Adaboost(Ada) [118, 119], Bagging (Bag) [120, 121] and XGBoost [122] over CO and TF-IDF features separately and

[20] Xgboost model for fake news.

then combined. This study used the default setting of all Decision Tree, Support Vector Machine and Adaboost, Bagging, and XGboost available at scikit-learn[21].

BiLSTM : The main objective of the *BiLSTM* model is to estimate the similarity and entailment between the headline and body. Initially, the headline and body are fed into bidirectional LSTM [123] to obtain an encoded representation of the headline and body. Subsequently, we estimate the angle and difference between the encoded representation of the headline and body. Finally, the encoded representations are concatenated with the difference and angle features and fed to a two-layer fully connected neural network for fake news classification.

Multihead Cross Attention over BERT Encoding (MCA): The main objective of this baseline model is to give high importance to sentences similar to the headline in the news article body. To achieve this objective, we apply multi-head cross attention between the encoding of the headline and the sentences in the news article body. Initially, the news article body is split into several sentences. Subsequently, we feed the headline and sentences into IndicBERT [22] [124] to obtain encoded representations of headline and sentences. Next, news article body representation is obtained by applying multi-head cross attention between the headline and body. Multi-head [58] cross-attention assigns a unique weight to each sentence based on its similarity with the headline. We restrict the maximum number of attention heads to two. Eventually, feature estimation between headline and body and MLP for classification is exactly similar to the `BiLSTM` model discussed above.

Recurrence over BERT (RoBERT): Motivated by study [125] we use `RoBERT` model. `RoBERT`, first, splits the body into sentences. Next, we encode the headline and sentences of the body using IndicBERT [124]. We apply bidirectional LSTM over sentence-encoded representation to obtain the encoded representation of the body. The main motivation behind `BiLSTM` over the encoding of sentences is that every sentence in the body is contextually related to the previous and next sentences. Bidirectional LSTM captures context from both directions. Eventually, feature estimation between headline and body and MLP for classification is exactly similar to the `BiLSTM` model discussed above.

We consider pre-trained FastText[23] [126] embeddings for *BiLSTM* model. F-measure (F) and Accuracy (Acc) have been used as evaluation metrics. We consider the default setting of the pre-trained IndicBERT [24] hosted on Hugging Face. Table 3.4 presents the details of the hyperparameter used to produce results presented in Table 3.5 and 3.3. Though we have

[21] scikit-learn
[22] IndicBERT
[23] FastText embedding
[24] IndicBERT

Table 3.3 Comparing the performance of models over human-crafted fake news dataset. (i) $F1$: denotes count overlapping features, (ii) $F2$ denotes TF-IDF features, (iii) $\oplus$: denotes concatenations operation between features.

Models		BBC								NAV							
		SM		NE-R		POS-R		POS-NE-R		SM		NE-R		POS-R		POS-NE-R	
		Acc	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**
Feature	$DT(F1)$	0.359	0.286	0.379	0.310	0.373	0.273	0.373	0.273	0.381	0.286	0.382	0.307	0.386	0.296	0.386	0.294
	$DT(F2)$	0.441	0.420	0.470	0.460	0.429	0.410	0.416	0.371	0.454	0.434	0.454	0.437	0.460	0.445	0.468	0.568
	$DT(F1 \oplus F2)$	0.420	0.386	0.467	0.449	0.398	0.332	0.400	0.325	0.447	0.443	0.488	0.487	0.423	0.362	0.426	0.41
	$SVM(F1)$	0.381	0.286	0.386	0.494	0.381	0.286	0.381	0.286	0.386	0.296	0.381	0.286	0.381	0.286	0.386	0.296
	$SVM(F2)$	0.402	0.328	0.420	0.357	0.406	0.334	0.392	0.308	0.420	0.357	0.413	0.346	0.392	0.308	0.419	0.355
	$SVM(F1 \oplus F2)$	0.382	0.288	0.400	0.321	0.382	0.288	0.381	0.286	0.390	0.303	0.577	0.494	0.381	0.286	0.395	0.31
	$Ada(F1)$	0.359	0.286	0.379	0.310	0.373	0.273	0.373	0.273	0.381	0.286	0.381	0.286	0.386	0.296	0.386	0.296
	$Ada(F2)$	0.385	0.304	0.413	0.349	0.386	0.296	0.386	0.296	0.437	0.385	0.427	0.368	0.419	0.355	0.443	0.401
	$Ada(F1 \oplus F2)$	0.371	0.280	0.390	0.327	0.379	0.283	0.381	0.286	0.379	0.283	0.463	0.431	0.355	0.298	0.390	0.3
	$Bag(F1)$	0.373	0.273	0.379	0.310	0.373	0.273	0.373	0.273	0.381	0.286	0.382	0.307	0.386	0.296	0.386	0.294
	$Bag(F2)$	0.424	0.395	0.450	0.432	0.410	0.387	0.407	0.357	0.460	0.440	0.450	0.437	0.460	0.443	0.473	0.454
	$Bag(F1 \oplus F2)$	0.390	0.315	0.444	0.41	0.388	0.304	0.395	0.313	0.449	0.425	0.491	0.488	0.409	0.338	0.412	0.363
	$XGBoost(F1)$	0.373	0.273	0.386	0.296	0.373	0.273	0.373	0.273	0.381	0.286	0.381	0.286	0.386	0.296	0.386	0.296
	$XGBoost(F2)$	0.388	0.298	0.410	0.341	0.396	0.314	0.386	0.296	0.422	0.359	0.409	0.339	0.415	0.348	0.432	0.377
	$XGBoost(F1 \oplus F2)$	0.373	0.291	0.407	0.337	0.381	0.286	0.382	0.577	0.382	0.288	0.475	0.450	0.388	0.298	0.386	0.325
LSTM	*BiLSTM*	0.381	0.347	0.412	0.361	**0.583**	**0.571**	**0.627**	0.352	**0.577**	**0.563**	0.419	0.405	0.386	0.302	0.386	0.306
BERT	*MCA*	0.372	0.271	0.43	0.422	0.432	0.529	0.432	0.429	**0.562**	**0.5402**	**0.627**	0.385	0.393	0.324	0.432	0.4295
	RoBERT	0.372	0.298	0.478	0.403	**0.502**	**0.5**	**0.572**	0.4	0.446	0.443	0.432	**0.429**	0.424	0.399	0.409	0.369

experimented with several hyperparameter values, we present the values that give the best result.

Table 3.4 Present details of hyperparameter values

Corpus	Model	Parameter	Value
BBC	BiLSTM	# word in headline	10
		# word in body	539
	RoBERT and MCA	# sentence in body	10
NAV	BiLSTM	# word in headline	10
		# word in body	350
	RoBERT and MCA	#sentence in body	10
Word embedding Dimension			300
Learning Rate			0.01
#MLP layer			2
LSTM hidden state dimension			100
# Attention head in MCA Model			2
# Epoch			40

Table 3.5 Comparing the performance of models over test datasets. (i) $F1$: denotes count overlapping features, (ii) $F2$ denotes TF-IDF features, (iii) $\oplus$: denotes concatenations operation between features.

Models		BBC								NAV							
		SM		NE-R		POS-R		POS-NE-R		SM		NE-R		POS-R		POS-NE-R	
		Acc	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**
Feature	$DT(F1)$	0.825	0.823	0.661	0.661	0.676	0.673	0.726	0.726	0.744	0.744	0.615	0.614	0.675	0.670	0.707	0.707
	$DT(F2)$	0.678	0.678	0.605	0.603	0.593	0.593	0.668	0.667	0.672	0.667	0.631	0.631	0.605	0.595	0.647	0.647
	$DT(F1 \oplus F2)$	0.740	0.739	0.597	0.596	0.601	0.601	0.668	0.668	0.680	0.680	0.624	0.623	0.618	0.613	0.646	0.646s
	$SVM(F1)$	0.813	0.813	0.661	0.661	0.678	0.675	0.702	0.696	0.734	0.732	0.615	0.614	0.672	0.663	0.704	0.703
	$SVM(F2)$	0.754	0.753	0.687	0.687	0.671	0.670	0.747	0.745	0.723	0.723	0.690	0.689	0.648	0.627	0.706	0.706
	$SVM(F1 \oplus F2)$	0.807	0.806	0.690	0.689	0.691	0.690	0.754	0.753	0.744	0.743	0.691	0.686	0.672	0.663	0.714	0.716
	$Ada(F1)$	0.825	0.823	0.661	0.661	0.676	0.673	0.726	0.726	0.744	0.743	0.615	0.614	0.675	0.673	0.704	0.703
	$Ada(F2)$	0.767	0.763	0.687	0.686	0.665	0.661	0.743	0.743	0.723	0.722	0.688	0.686	0.652	0.645	0.703	0.70
	$Ada(F1 \oplus F2)$	0.825	0.823	0.686	0.684	0.688	0.687	0.752	0.747	0.747	0.746	0.694	0.693	0.677	0.673	0.715	0.715
	$Bag(F1)$	0.825	0.823	0.661	0.661	0.676	0.673	0.726	0.726	0.744	0.743	0.615	0.614	0.675	0.673	0.704	0.703
	$Bag(F2)$	0.685	0.684	0.607	0.610	0.603	0.602	0.673	0.673	0.674	0.674	0.635	0.634	0.604	0.604	0.649	0.649
	$Bag(F1 \oplus F2)$	0.772	0.771	0.605	0.601	0.621	0.620	0.683	0.682	0.699	0.699	0.635	0.634	0.625	0.619	0.661	0.661
	$XGBoost(F1)$	0.826	0.823	0.661	0.661	0.676	0.673	0.726	0.726	0.744	0.743	0.615	0.614	0.675	0.673	0.704	0.703
	$XGBoost(F2)$	0.768	0.763	0.684	0.682	0.668	0.667	0.743	0.743	0.723	0.723	0.689	0.689	0.653	0.645	0.7	0.706
	$XGBoost(F1 \oplus F2)$	0.825	0.822	0.690	0.689	0.692	0.692	0.750	0.750	0.747	0.746	0.694	0.693	0.677	0.674	0.712	0.716
LSTM	*BiLSTM*	**0.837**	**0.8361**	**0.969**	0.969	0.95	0.95	0.974	0.973	**0.926**	**0.925**	**0.919**	**0.919**	**0.999**	**0.999**	**0.98**	**0.975**
BERT	*MCA*	0.497	0.332	0.519	0.388	0.88	0.878	0.869	0.868	0.525	0.512	0.497	0.322	0.718	0.659	0.555	0.528
	RoBERT	0.76	0.757	0.905	0.904	**0.989**	**0.988**	**0.987**	**0.986**	0.791	0.789	0.858	0.856	0.938	0.936	0.897	0.897

3.5.1 Results and Discussion

Table 3.5 presents the performance of models over test datasets. Similarly, Table 3.3 presents the response of models over manually crafted fake news datasets. As presented in Table 3.5 and 3.3, the baseline models are grouped into three categories, namely *Feature*, *LSTM* and *BERT*. From Table 3.5, it is evident that the performance of *Feature*-based models is better over *SM* dataset, compared to the performance over *NE-R, POS-R and POS-NE-R*. With respect to the dataset curation process and characteristics of *SM*, although noisy content is inserted, but the existing content of the body is not modified. Hence, the lexical overlap between the headline and body doesn't change. But in the case of *NE-R, POS-R and POS-NE-R*, essential words of the body are modified for generating fake news articles; hence lexical overlap is reduced. From Table 3.5, by comparing the performance of the *Feature*-based model, the following observations can be made: (i) *Feature*-based models are unsuitable for fake news article detection where datasets lack lexical overlap. (ii) Our proposed dataset curation methods *NE-R, POS-R and POS-NE-R* effectively modify the news article body to generate fake news articles. Further, comparing the performance of *LSTM, BERT* based models, it is evident that the performance of *BiLSTM* model over *NE-R, POS-R and POS-NE-R* datasets are significantly high compared to the performance of the *BiLSTM* model over *SM* dataset. The possible reason behind such performance in the case of *SM* dataset is due to the existing content of body not being modified, even in case of fake news sample. Hence, it affects the discriminative capability of models. But in the case of

Table 3.7 Comparing the performance of models over real fake news datasets. (i) **Acc** : indicates the accuracy, (ii) **T** and **F** indicates F-measure score for *True* news and *Fake* news class respectively.

		SM			NE-R			POS-R			POS-NE-R		
	Model	**Acc**	**T**	**F**	**Acc**	**T**	**F**	**Acc**	**T**	**F**	**Acc**	**T**	**F**
BBC	BiLSTM	**0.7**	**0.7**	**0.723**	0.583	0.677	0.413	0.476	0.212	**0.643**	0.472	0.318	**0.638**
	MCA	0.5	0.66	0.23	0.5	0.66	0.291	0.533	**0.656**	0.274	**0.6**	**0.7**	0.282
	RoBERT	0.563	0.652	0.414	**0.704**	**0.737**	**0.663**	**0.561**	**0.651**	0.408	0.57	0.558	0.582
NAV	BiLSTM	0.548	0.59	0.496	**0.653**	**0.714**	**0.6**	0.5	0.666	0.1	0.635	0.731	0.5
	MCA	0.461	0.216	**0.6**	0.468	0.216	0.6	0.533	0.656	0.274	0.582	0.453	**0.654**
	RoBERT	**0.563**	**0.61**	**0.53**	0.618	0.612	0.624	0.512	0.667	0.08	0.637	0.66	0.61

NE-R, POS-R and POS-NE-R datasets, important words of body are modified during the fake sample generation process. This improves the discriminative capability of models for the classification of fake news. From such observations, it can be concluded that our dataset curation methods effectively generate fake news datasets, which are adequate for models to discriminate between *True* and *Fake* news. To further validate this claim, we compare the response of *MCA* and *RoBERT* across datasets. From Table 3.5 we observe that *RoBERT* performs better over *NE-R, POS-R and POS-NE-R* datasets compared to the performance over *SM* datasets. Subsequently, on comparing the performance of models from Table 3.5 it is evident that *BiLSTM* outperforms all other baseline models except for *POS-R* and *POS-NE-R* dataset over BBC corpus. The performance of *RoBERT* and *MCA* could be further improved with fine-tuning of BERT and usage of the optimal number of attention heads for *MCA*.

Human Crafted Fake News Dataset: As discussed in section 3.3.5, human crafted fake news dataset was used only to validate the response of models trained over proposed datasets. From Table 3.3, it is evident that the performance of *BiLSTM* and *RoBERT* over *POS-R* and *POS-NE-R* is superior to performance over *SM and NE-R* datasets for BBC corpus. From Table 3.3, it is also apparent that *BiLSTM* and *MCA* performance over *SM* and *NE-R* datasets is superior to performance over *POS-R and POS-NE-R* for NAV corpus. From the above observation and observing the response of models over human-crafted fake news, establish that our proposed dataset is effective for fake news detection task model training.

3.6 Quality and Reliability of Proposed Datasets

To assess the quality and reliability of the proposed datasets, we perform the following evaluations over proposed datasets: *(i)* human annotation and evaluations, *(ii)* validation of

models trained on proposed datasets for real fake news classification, (iii) error analysis of models response over real fake news datasets.

3.6.1 Human Annotation and Evaluations

Following the procedure reported in studies, [28, 127], we perform human Turing test on the proposed datasets to assess and analyze their quality and reliability. We randomly selected 500 fake news samples generated by each dataset curation methods: SM, NE-R, POS-R and POS-NE-R. These randomly selected 2000 samples with only headline and body pair were given to 8 annotators without any information regarding the veracity of these news samples. To ensure high-quality annotations, we selected annotators with a literary background and people who are avid readers of the news. They were asked to verify the authenticity of news by the following measures : (i) check the topic and information consistency within the news article and incongruity between headline and body, (ii) verify the central claim, facts, numeric values, and dates associated, (iii) verify the information associated with events and named entities by exploring Wikipedia, and news reporting by authentic media houses such as ANI, PTI, Navbharat times, India Today newsgroup etc. Given 2000 headline and body pair, annotators were tasked to assign a label *True* or *Fake* based on their assessment. Annotators assigned *Fake* label to 83.4% and *True* label to 16.6% headline body pair out of 2000. The 83.4% accuracy of human annotators suggests that fake news generated by our proposed dataset generation methods is non-trivial for even humans to determine its veracity. To measure the quality of annotations, we estimate inter-annotator agreement between eight annotators using Fleiss' kappa (k). The inter-annotator agreement score using Fleiss' kappa (k) is 0.862.

3.6.2 Models Response on Real Fake News Dataset

This subsection is dedicated to evaluating the effectiveness of models trained on the proposed datasets in detecting fake news disseminated across digital platforms. The primary objective of this assessment is to ascertain the fidelity of the proposed datasets in representing the essential characteristics of fake news articles circulated online. Specifically, we endeavor to determine whether the proposed datasets accurately capture the distinct traits exhibited by fake news articles within the digital sphere. Our analysis focuses on discerning if models trained on the proposed datasets exhibit a notable proficiency in identifying genuine instances of fake news. High accuracy achieved by these models would suggest that the datasets

Table 3.8 Present the statistical comparison between NAV and BBC corpus in terms of the number of POS and NE words. Here, STD indicates Standard Deviation.

	POS-R		**NE-R**	
Corpus	NAV	BBC	NAV	BBC
Mean	24.42	50.757	12	31.827
STD	17	32.827	9	20.257
Minimum	1	1	1	1
Maximum	163	297	7	19
25%	15	29	10	27
50%	21	42	17	39
75%	36	64	197	177

successfully encapsulate the defining features of authentic fake news articles. We select three best-performing models (from Table 3.5) trained over the proposed dataset by considering real fake news datasets as a test set. Table 3.7 presents the responses of models over real fake news datasets. From Table 3.7, the performance of *RoBERT* and *BiLSTM* over BBC corpus indicate that models trained over our proposed datasets are adequate for real fake news detection. Similar conclusions can be drawn by observing the performance of *BiLSTM* and *RoBERT* over *SM*, *NER*, and $POS-NE-R$ dataset generation methods for the NAV corpus. From such observations, it is apparent that models trained over our proposed datasets are effective for real fake news classifications. This establishes that our fake news dataset, generated by the proposed dataset generation methods, is reliable for real-world fake news article detection. In our work, we have presented only the baseline models. However, the performance may be improved after addressing language and dataset-specific concerns, which we have planned to explore in our future work. However, the response of *MCA* model over real fake news dataset is poor over few dataset generation methods. This indicates that the optimal number of attention heads needs to be examined. From Table 3.7, it can be observed that the performance of models trained on $POS-R$ NAV corpus is poor in predicting real fake news. But for BBC, the corpus performance of models trained on $POS-R$ is good. To understand the difference in the performance of models over $POS-R$ for NAV and BBC corpus, we count the number of POS words in each news article replaced with their antonyms during dataset curation for both BBC and NAV corpus. Table 3.8 presents the statistics of the POS word count of news articles in the NAV and BBC corpus. From Table 3.8, it is apparent that the number of POS words in news articles of the NAV corpus is less compared to the number of POS words in a news article of the BBC corpus. On average, BBC corpus has

50.757 POS words, whereas NAV corpus has 24.40, which is more than half of BBC corpus. So during dataset curation using $POS-R$, more words are replaced in a news article of BBC corpus compared to the number of words replaced in news articles of NAV corpus. The more POS words are replaced with their antonyms, the high degree of fake news articles generated. The fewer POS words replaced with antonyms during dataset generations, the less degree of fake news articles generated. This is the possible reason for the decrease in performance of models over $POS-R$ data built over NAV corpus compared to $POS-R$ dataset over BBC Corpus in Table 3.7. From such observations, it is also evident that the quality of fake news samples generated by $POS-R$ data curation methods depends on the number of POS words in the news article corpus. So, to curate a high-quality fake news dataset using $POS-R$, the news article corpus with a significant number of POS words must be selected.

3.6.3 Error Analysis

To investigate the kind of real fake news samples, correctly classified and misclassified by models, we conducted an error analysis of models' predictions over real fake news datasets. Our error analysis shows the models trained over proposed datasets can detect different kinds of fake news, such as satires, false connections, misleading content etc. Our findings from error analysis confirm our hypothesis discussed in the methodology section. To further validate the effectiveness of models trained over proposed datasets in real fake news detection, we study the prediction of models over real fake news dataset by manually identifying the kind of real fake news correctly predicted by models and the kind of real fake news that is misclassified. Our manual investigations on the prediction of models suggest that models are able to detect different kind of fake news. An example presented in Figure 3.7 is satirical fake news, which is correctly predicted by models trained over proposed datasets. Figure 3.6 and 3.8 present examples of real fake news from which our SM and $NE-R$ based approaches are inspired. These are also correctly predicted by $BiLSTM$ model trained over proposed datasets. Figure 3.9 presents a misclassified example of real fake news, which is a claim made by a filmmaker. To detect such kind of fake news, external knowledge is required to verify the authenticity of a claim. Such fake news is difficult to detect based on linguistics and the content of the news documents.

Headline	इंस्टा पे पिक अपलोड करने के लिए बांधी बॉयफ्रेंड को राखी, बॉयफ्रेंड ने की आत्मदाह की कोशिश **Translation :** Girl tied rakhi to boyfriend for uploading pic on insta, boyfriend attempted self-immolation.
Body	**अभी हाल में ही 2 लड़कियां सेल्फी लेने के चक्कर में बाढ़ में फंस गई, अंत में पुलिस प्रशासन के रेस्क्यू ऑपरेशन के बाद वो सुरक्षित निकाली गईं।** एक लड़की ने जिसके इंस्टाग्राम और ट्विटर में काफी फॉलोवर्स है अपने बॉयफ्रेंड के राखी बांध दी जिससे क्षुब्ध होकर बॉयफ्रेंड ने आत्मदाह की कोशिश की, घटना की पड़ताल करने पर पता चला कि लड़की घर की अकेली संतान थी और सोशल मीडिया में चल रहे ट्रेंड " शो योर राखी पिक्स " पे फ़ोटो अपलोड करने के चक्कर में बॉयफ्रेंड को घर बुलवाया बॉयफ्रेंड ने तुरंत अपने व्हाट्सऐप पे "आज तेरा भाई करके आया" स्टेटस लगाया और लड़की के घर पहुंच गया लेकिन उसके मंसूबों पर तब पानी फिर गया जब लड़की ने उसकी कलाई पे राखी बांध दी। **Translation:** **Recently, 2 girls got trapped in the flood while taking a selfie, later they were evacuated safely after the rescue operation of the police administration.** A girl who has a lot of followers on Instagram and Twitter tied rakhi on her boyfriend, due to which the boyfriend attempted self-immolation. After investigating the incident, it was found that the girl was the only child of the house and invited boyfriend home to follow the trend going on in social media "Show Your Rakhi Pics". The boyfriend immediately put the status "Aaj Tera Bhai Karke Aaya" on his WhatsApp and reached the girl's house but his plans were shattered when the girl tied a rakhi on his wrist.

Model and dataset description

Model	BiLSTM
Dataset generation approach	Split and Merge(SM)
Corpus	BBC

Fig. 3.6 Example of real fake news correctly predicted by BiLSTM model trained over dataset generated by *SM* method over BBC corpus. This is also an example of real fake news where content is unrelated to the headline and other paragraphs of the body. So, our split and merge approach *SM* is inspired by such observations over such real fake news. In this figure, content unrelated to the headline and other body paragraphs is marked in red.

Headline	BCCI ने बुलाई थी चयनकर्ताओं की मीटिंग, अनुष्का शर्मा शूटिंग में व्यस्त होने के कारण रद्द हुई मीटिंग **Translation:** BCCI had called the meeting of the selectors, the meeting was canceled due to Anushka Sharma being busy with the shoot.
Body	भारतीय क्रिकेट अपने बेहतरीन वक़्त से तो एक बड़े बदलाव के दौर से गुज़र रहा है, अभी अभी बंगाल टाइगर दादा सौरव गांगुली ने bcci अध्यक्ष पद संभाला है जिस से क्रिकेट जगत में काफी खुशी है। आते ही दादा ने चयनकर्ताओं की एक आपातकालीन मीटिंग बुलाई थी ताकि आगे होने वाले चयन में मुद्दे और लक्ष्य स्पष्ट हो सके मगर एक ऐसी वजह से मीटिंग को कैंसिल करना पड़ा जिसे सुन खुद दादा स्तब्ध रह गए।दरअसल दादा भी आपकी तरह नहीं जानते थे कि भारतीय टीम के लिए खिलाड़ियों का जो चयन होता है वो दरअसल चयनकर्ता नहीं श्रीमती कोहली यानी बॉलीवुड अभिनेत्री अनुष्का शर्मा करती हैं बाकी चयन करता तो उनके लिए चाय नाश्ते के इंतेज़ाम करते हैं। अब जब अनुष्का ही नहीं थी तो बाकी चयनकर्ताओं ने भी हाथ खड़े कर लिये और दादा से कहा हम कोई निर्णय नहीं ले पाएंगे क्योंकि हमने कभी लिए ही नहीं हमसे न हो पायेगा। इसके बाद दादा ने कोहली से बात कर अनुष्का को बुलाने की कोशिश की मगर अनुष्का अपनी ऐड की शूटिंग में व्यस्त होने के कारण भारतीय टीम के चयनकर्ताओं की मीटिंग को रद्द करना पड़ा।अनुष्का के चयन प्रक्रिया में एक छत्र राज करने वाले दावे को साबित किया भूतपूर्व महान खिलाड़ी फारुख इंजीनियर ने। फारुख कहते हैं चयनकर्ता सिर्फ अनुष्का के लिए चायपत्ती का चयन करते हैं खिलाड़ी का चयन उनकी श्रीमती के प्रति सेवा पर निर्भर रहता है। इस दावे में काफी दम इसलिए भी लगता है क्योंकि अचानक विश्वकप जैसी प्रतियोगिता में विजय शंकर जैसे खिलाड़ी जिसको आप अपनी गली की टीम में न शामिल करें उनको अम्बाती रायडू जैसे बल्लेबाज़ के बदले टीम में भेजा गया। माना जा रहा है विजय शंकर एक साथ 20 शॉपिंग बैग उठा सकते हैं और इसका उन्हें फायदा मिला।और फिर कुंबले जैसे महान कोच की जगह रवि शास्त्री का अपॉइंटमेंट हो या टीम में टिकटोक बॉय के एल राहुल का चयन यह सब अनुष्का जी की मेहरबानियों का नतीजा लग रहा है। **Translation** : Indian cricket is going through a big change from its best time, right now, Bengal Tiger Dada Sourav Ganguly has taken over the post of BCCI President, due to which there is a lot of happiness in the cricket world. Dada had called an emergency meeting of the selectors as soon as he arrived so that the issues and goals could be clarified in the upcoming selection, but had to cancel the meeting due to a reason, listening to which Dada himself was shocked. Actually dada also did not know like you that the selection of players for the Indian team is actually not done by the selectors, but by Mrs. Kohli, i.e. Bollywood actress Anushka Sharma. The rest of the selectors make arrangements for tea and snacks for them. Now that Anushka was not there, the rest of the selectors also gave up and said to Dada that they will not be able to take any decision because we have never taken before. After this, Dada talked to Kohli and tried to call Anushka, but Anushka was busy in her ad shooting, so the Indian team selectors meeting had to be canceled. Former great player Farooq Engineer proved Anushka's claim that there was an umbrella rule in the selection process. Farooq says that the selectors are there only for choosing tea leaves for Anushka. The player's selection depends on their service to his wife. There is a lot of merit in this claim because suddenly in a tournament like the World Cup, a player like Vijay Shankar, whom you should not include in your gully team, was added to the team instead of a batsman like Ambati Rayudu. It is believed that Vijay Shankar can lift 20 shopping bags at once, and he got benefit from it. And then the appointment of Ravi Shastri instead of a great coach like Kumble or the selection of the TikTok boy KL Rahul in the team, it all seems to be the result of Anushka ji's kindness.

Model and dataset description

Model	BiLSTM
Dataset generation approach	Split and Merge(SM)
Corpus	BBC

Fig. 3.7 Example of real fake news correctly predicted by BiLSTM model trained over dataset generated by *SM* method over BBC corpus. This is also an example of real fake news, where mocking of a news statement made by a former cricketer is circulated as fake satire news. Satire is a dangerous kind of fake news, which has the potential to bypass fact-checkers. Correctly predicting satire kind of fake news indicates that the models trained over our proposed dataset can detect different types of fake news.

Headline	हिंद लिख साइन बोर्ड कालिख पोत तस्वीर किसान आंदोलन नह
Body	सिख समुदाय के लोग हिंद साइन बोर्ड कालिख पोत दिख तस्वीर सेट वीडिय वायरल रह सोशल मीडिय यूज़र्स वर्तमान चल रह किसान आंदोलन जोड़ शेयर रह यूट्यूब मौजूद वायरल वीडिय तस्वीर हिंद इम्पोज़िशन के ख़िलाफ़ प्रदर्शन हिस्स ना वर्तमान चल रह किसान आंदोलन सम्बंधित हैंहिंद इम्पोज़िशन ग़ैर हिंद भाष थोप इसक ख़िलाफ़ प्रदर्शन 2017 दक्षिण भारत शुर कर्नाटक तमिलनाड जगह लोग हिंद साइन बोर्ड कालिख पोत दिय प्रदर्शनकार कह उनक भाष हिंद नह हिंद साइन बोर्ड सबस उनक भाष के क्य लिख गय रिपोर्ट पढ़ें जनसंख्य नियंत्रण क़ानून के दाव के पीएम मोद एडिटेड तस्वीर वायरलफ़ेसबुक कृष्णकांत सिंह नाम के यूज़र तस्वीर सेट शेयर लिख रिलायंस जिय के टॉवर तोड़ के अगल काम हिंदी_नही_चल क्य किसान समाधान नह व्यवधान चाह हैंय शांत नह संघर्ष चाह हैय विकास नह विनाश चाह स्वतंत्र नह स्वछंद चाह सड़क नह स्पीड ब्रे चाह हैं। #फर्जी_किसान_आन्दोलनपोस्ट देख आर्काइव वर्ज़न देखेंट्विटर श्रीश त्रिपाठ नाम के यूज़र वीडिय तस्वीर कोलाज शेयर कैप्शन लिख असल चेहर साम आ रह टॉवर तोड़ के पंजाब हिंद नह चल किसान आन्दोलन बह हिन्द हिन्द विरोध असल मकसद किसान आंदोलन हकीकत खालिस्ता आंदोलन किसान के भेष आतंक उनक समर्थक उनक एजेंडाअराजक फैल देश तोड़ लोग पंजाब के दुश्मन इन्ह पूर भारत रह रह पंजाब के बार चिं ऐस कभ नह दोगल बामपंथ असर खालिस्तान किय कभीपोस्ट आर्काइव वर्ज़न देख पोस्ट देखेंअसल चेहर साम आ रह है।टॉवर तोड़ के पंजाब हिंद नह चलेगी।किसान आन्दोलन बह हिन्द हिन्द विरोध असल मकसद है।य हैकिसान आंदोलन हकीकतय खालिस्ता आंदोलन किसान के भेष आतंकीउनक समर्थक है।उनक एजेंडाअराजक फैल तस्वीर के शेयर किय पोस्ट देख के दाव के फ़ेसबुक बड़ पैम पोस्ट शेयर किय हैंफ़ैक्ट चेक रेलव स्टेशन बन मस्जिद तस्वीर कह हैफ़ैक्ट चेकबूम तस्वीर रिवर्स इमेज सर्च खोज इन्ह तस्वीर के साल 2017 मीडिय रिपोर्ट्स साम आईं जिनम तस्वीर के सेट इस्तेमाल किय गय मीडिय रिपोर्ट्स पढ़ेंहम पाय सोशल मीडिय वायरल तस्वीर 2017 पंजाब हिंद के ख़िलाफ़ प्रदर्शन प्रदर्शन के पंजाब सरकार पंजाब भाष डेस्टिनेशन सबस लिख शुर किय था25 अक्टूबर 2017 हिंदुस्तान टाइम्स प्रकाशित रिपोर्ट के अनुसार पंजाब संगठन द्वार हिंद अंग्रेज के नीच साइनबोर्ड नंबर तीन पंजाब लिख आरोप लग विरोध प्रदर्शन किय गय था।25 अक्टूबर 2017 इंडिय टीव प्रकाशित रिपोर्ट कह गय कट्टरपंथ सिख समूह बठिंडा-फ़रीदकोट राष्ट्रीय राजमार्ग के किनार साइनबोर्ड हिंद अंग्रेज शब्द काल के बड़ पैम अभियान चल मांग रह साइनबोर्ड पंजाब सभ भाष वरीय मिले असदुद्दीन ओवैस बध दे अमित शाह तस्वीर एडिटेड हैवायरल वीडियोवायरल क्लिप के स्क्रीनशॉट रिवर्स इमेज सर्च 14 सितंबर 2020 साथियम न्यूज़ चैनल द्वार अपलोड किय गय यूट्यूब वीडिय मिलातमिल वीडिय के शीर्षक हिंद विरोध के ख़िलाफ़ विरोध-प्रदर्शन दक्षिण भारत नह बल्क उत्तर दिख है

Model and dataset description

Model	BiLSTM
Dataset generation approach	Name Entity Replacement (NE-R)
Corpus	BBC

Fig. 3.8 Example of real fake news correctly predicted by BiLSTM model trained over dataset generated by $NE-R$ method over BBC corpus. This is also an example of real fake news; different entities are not correlated in body, and the entity between headline and body has no relations. So, our Named Entity Replacements $NE-R$ is inspired by such observations over such real fake news. In this figure, named entities are marked in red.

Headline	मधुर भंडारकर ने नसीरुद्दीन के 'डर' को नकारा **Translation** : Madhur Bhandarkar denies Naseeruddin's 'fear'
Body	कुछ दिन पहले ही अभिनेता नसीरुद्दीन शाह ने एक विवादित बयान देकर 'असहिष्णुता' की चर्चा को एकबार फिर चिंगारी दे दी है। उन्होंने कहा कि वो एक ऐसी परिस्थिति के बारे में चिंतित हैं जहाँ उनके बच्चों को उग्र भीड़ घेरकर उनसे पूछ रही है कि उनका धर्म हिन्दू है या मुस्लिम? उन्होंने यह भी कहा कि भारतीय समाज में ज़हर फ़ैल चुका है। मशहूर अभिनेता के बयान ने बहुत सारी प्रतिक्रियाओं को निमंत्रण दे दिया है। बॉलीवुड से भी बहुत से लोगों ने इस पर प्रतिक्रिया दी है। कुछ मीडिया रिपोर्ट्स की मानें तो डायरेक्टर मधुर भंडारकर और आशुतोष राणा भी नसीरुद्दीन शाह के 'समर्थन में खड़े हो गए हैं'। इंडियन एक्सप्रेस के एक लेख की हेडलाइन थी, 'कानून और व्यवस्था पर शाह के बयान के बाद आशुतोष राणा, मधुर भंडारकर ने दिखाई एकजुटता'। जबकि 'Quint' ने लिखा कि आशुतोष राणा, मधुर भंडारकर और अन्य लोगों ने किया नसीरुद्दीन शाह का बचाव'। द न्यू इंडियन एक्सप्रेस ने अपनी रिपोर्ट में लिखा कि मधुर भंडारकर ने नसीरुद्दीन शाह का बचाव किया। बिज़नेस स्टैंडर्ड ने लिखा है कि 'आशुतोष राणा, मधुर भंडारकर ने नसीरुद्दीन शाह के बयान का बचाव किया'। **Translation** : A few days ago, actor Naseeruddin Shah has once again sparked the discussion of 'intolerance' by making a controversial statement. He said he was worried about a situation where his children are surrounded by a mob and being asked if their religion is Hindu or Muslim? He also said that poison has spread in Indian society. The famous actor's statement has invited a lot of reactions. A lot of people from Bollywood have also reacted to this. According to some media reports, directors Madhur Bhandarkar and Ashutosh Rana have also "stood in support" of Naseeruddin Shah. An Article in The Indian Express had the headline, "After Shah's statement on law and order, Ashutosh Rana, Madhur Bhandarkar show solidarity". While 'Quint' wrote that Ashutosh Rana, Madhur Bhandarkar and others defend Naseeruddin Shah. The New Indian Express wrote in its report that Madhur Bhandarkar defended Naseeruddin Shah. Business Standard has written that 'Ashutosh Rana, Madhur Bhandarkar defend Naseeruddin Shah's statement'.

Model and dataset description

Model	BiLSTM
Dataset generation approach	Split and Merge(SM)
Corpus	BBC

Fig. 3.9 Example of real fake news misclassified by BiLSTM model trained over dataset generated by *SM* method over BBC corpus. This is an example of real fake news, which is a statement and claim made by an eminent Bollywood filmmaker. To detect such kind of fake, external knowledge is required to verify the authenticity of a claim. Such fake news is difficult to detect based on linguistics and the content of the documents.

Summary

This chapter of the thesis curates and proposes four types of large-scale Hindi datasets for fake news article detection, namely, split and merge *SM*, named-entity replacement *NE-R*, part of speech replacement *POS-R*, part-of-speech and named-entity replacement *POS-NE-R*. The quality and reliability of proposed datasets are evaluated using several baseline models over the test, human-crafted fake news datasets and real fake news datasets. We also applied human annotation and evaluations, and error analysis on models' predictions over real fake news datasets to further assess and validate the proposed datasets' quality and reliability. Our proposed dataset curations methods can be easily extended to curate large-scale datasets in any vernacular language for several tasks, such as fake news article detection, document similarity and entailment classification. Further, we identify the following four future research directions: (i) evaluate the performance of models trained over hybrid datasets by combining false samples generated by all four proposed dataset curation methods. (ii) Explore methods to generate contextual fake news samples. (iii) Curate large-scale fake news article detection datasets for other regional languages in India. (iv) Explore a state-of-the-art model for fake news article detection over proposed datasets.

Chapter 4

Deep Hierarchical Encoding for Detecting Incongruent News

With the increase in misinformation across digital platforms, incongruent news detection is becoming an important research problem. Earlier, researchers have exploited various feature engineering approaches and deep learning models with embedding to capture incongruity between news headlines and their respective bodies. Studies have broadly considered different combinations of bag-of-words-based features [22] [35] [23], sequential encoding [23], hierarchical encoding [1] [28], headline-guided attention-based encoding [1] [28], etc., of the text in headlines and bodies. This chapter addresses two significant limitations observed with hierarchical encoding and headline-guided attention-based encoding methods. Existing hierarchical encoding-based studies limit the hierarchical structure of the body of a news article to paragraph level only, undermining the importance of incorporating long-term dependency from word level to sentence, paragraph and body. Further, existing headline-guided attention-based encoding focuses on contextually similar contents in the body of the headline, undermining the importance of incorporating contextually dissimilar contents. Motivated by the above observations, this chapter of the thesis proposes a Gated Recursive And Sequential Deep Hierarchical Encoding (GraSHE) method for detecting incongruent news articles by extending the hierarchical structure of the news body from the body to the word level and incorporating incongruity weight. From various experimental setups over three publicly available benchmark datasets, the experimental results indicate that the proposed model outperforms baseline models with bag-of-word-based features, and sequential, hierarchical, and headline-guided attention-based encoding methods. We conducted several ablation studies to further validate the proposed model's performance. The following key observations can be

made from the ablation study: *(i)* models with hierarchical encoding outperform models with non-hierarchical encoding. *(ii)* recursive encoding of sentences boosts the performance of models as compared to sequential encoding of sentences within paragraphs. *(iii* incongruent news article detection is domain-dependent. Incorporating explicit features further boosts the performance of our proposed model and also decreases the domain dependency of models.

4.1 Introduction

Detecting incongruity between news headlines and their corresponding bodies has emerged as a crucial research challenge in recent times, aimed at early detection of misinformation in electronic media [2][3]. An article is deemed incongruent if its headline fails to accurately represent its body due to fabrication, manipulation, false connections[1], or miscontextualization[2]. Notably, individuals tend to primarily read news headlines rather than the full article [11], and the impressions formed by these headlines significantly contribute to the virality of news stories on social media platforms [128]. Consequently, the detection of incongruent news headlines plays a pivotal role in combating misinformation in electronic media.

Though the initial studies on incongruity detection in a news article can be credited to `Fake News Challenge`, *(FNC-1)* [22], the importance of the problem can be traced back to the year 2007 [129]. In recent times, researchers have exploited various methods for detecting incongruent news articles such as simple *n*-gram features-based models [35] [23], summarization-based models [30] [29], and hierarchical encoding-based models [1] [28]. While incongruent news articles with distinctive features between their headlines and bodies are easy to identify, detecting a systematically created incongruent news article is a non-trivial task. From the FNC-1 challenge [22], it is evident that incorporating features extracted from diverse perspectives helps in detecting incongruent news articles better. A classic example is *XGBoost*, the winner of the FNC-1 challenge, which considers various types of features such as *n*-grams, latent features, sentiment, etc. However, as observed in [24], the models with bag-of-words based features often fail to capture information like complex negations, deep semantic relationships, and propositional contents which are important for incongruity detection. To capture deeper contextual and sequential semantic relationship between headline and news body, studies in [23][27][25] combine embeddings obtained from sequential models (like LSTM) and the explicit features like *n*-grams. While the above studies define a news

[1]Instances where the caption of an image does not align with the image itself or the headline does not support the content.

[2]Legitimate information presented in an incorrect context.

article as a sequence of texts, studies in [28] [1] have defined a news article as a hierarchical structure, i.e., *body as a collection of paragraphs, and paragraph as a collection of sentences*. They also apply headline-guided attention to select paragraphs which are contextually similar to the headline. Though, the above hierarchical encoding and headline-guided, attention-based method provide promising results as compared to their sequential and bag-of-word counterparts (also observed in this study). However, the hierarchical structure in the above studies is limited to paragraph level only, and headline-guided attention highlights paragraphs that are contextually similar to headlines. As textual documents are order sensitive, extending the hierarchical structure up to the word level helps capture not only long-term dependencies but also syntactic structure between words within a sentence [78]. Further, highlighting contextually dissimilar paragraphs or sentences in news paragraphs will also help to detect partially incongruent news articles. Here, a partially incongruent news article refers to a news article with few segments of incongruent content while the majority of the content is congruent. Motivated by the above observations, this Chapter proposes a `Gated Recursive And Sequential Deep Hierarchical Encoding` *(GraSHE)* method for detecting incongruent news articles by extending the hierarchy structure of news body from body to word level and incorporating incongruent weights. The proposed model, *(GraSHE)* captures the long-term dependencies and syntactic structure by incorporating sequential information at the paragraph and body level (using BiLSTM), and syntactic structure at the sentence level (child-sum Tree LSTM [41]). Further, unlike headline guided attention models [28][27], *(GraSHE)* also incorporates incongruity weight to capture non-dominant textual segments which are not congruent with other part of news body. From various experimental observations over three publicly available benchmark datasets, it is observed that the proposed method outperforms its bag-of-words, sequential, and hierarchical counterparts.

4.2 Research Objective

In this work, the research objective is to capture the long-term dependencies and syntactic structure by incorporating sequential information at the paragraph and body level (using BiLSTM) and syntactic structure at the sentence level (child-sum Tree LSTM [41]). Further, unlike headline guided attention models [28][27], *(GraSHE)* also incorporates incongruity weight to capture non-dominant textual segments which are not congruent with other part of news body.

4.3 Contributions

The key highlights of contributions are summarized as follows:

1. Propose a Gated Recursive And Sequential Deep Hierarchical Encoding *(GraSHE)* model, which can effectively identify incongruent news articles of different characteristics.
2. Perform various ablation studies to understand the importance of considering the deeper hierarchy to capture long-term dependencies and syntactic structure.
3. Ablation study to understand the effect of incorporating select gate.
4. Conduct an empirical study to investigate the domain dependency of incongruent news article detection tasks.

4.4 Literature Review

In the literature, studies [130][131][132][133][134][135][136] have briefly reviewed and analyzed works related to misinformation and disinformation detection. In this study, we only retrospect works related to incongruent news article detection. The study [3] suggests that incongruent news differs from other types of misinformation, and the spread of incongruent news leads to the spread of fake news articles over digital platforms. Though incongruent news article detection and clickbait detection are related with regard to news headlines, incongruent news article detection is comprehensively different from clickbait detection [6][3]. While clickbait detection focuses on identifying news headlines that are created with several stylistic and linguistic features such as forward-referencing, mentioning of attractive words, public figures and personalities, etc. for the purpose of attracting attention of the readers to read the article, the incongruent news article detection focuses on identifying news articles whose headlines are incongruent with their respective news bodies [3]. Clickbaits attempt to attract readers to click on the headlines and read the news articles. The incongruent news articles are created to spread misinformation. Clickbait may be described by only headlines, but incongruent news needs to be described by the relation between the headline and the body [6]. In literature, studies on incongruent news article detection can be visualized from two different aspects; feature engineering and classification approaches. Though earlier studies on incongruent news article detection mostly considered bag-of-words-based features such as n-grams, tf-idf and topic modeling features [22][23][137][35][138], recent studies

mostly explore neural based encoding models [27][28][139][77][26]. Considering the significance of incorporating contextual and sequential information present in headlines and bodies, the studies [24][25][40] consider both the sequential encoding and the bag-of-words-based features. The incongruent news detection methods can be broadly divided into summarization and similarity-based approaches. The methods reported in the studies [77][140][1][141] encode the headline and body using sequential neural models, and then apply similarity matching methods between the encoding of the headline and the body. Further, in the studies [27][26][28], authors have encoded news body by exploiting the structural properties of news body. The study [27] uses an inverted pyramid writing style as reported in study [142] to give more attention to the first few lines of news body, then apply similarity matching between the encoding of headlines and bodies.

Study [26] learns hierarchical discourse structure between sentences of the news article to encode the relationship between sentences of the news article while encoding the news article. A recent similarity-based study [28] exploits the hierarchical structure of the body, which splits the body into paragraphs and encodes each paragraph separately, then applies headline-guided attention to select paragraphs that are highly contextually similar to the headline. However, the hierarchical structure in the above studies is limited to paragraph level only. For incongruent news articles detection, extending the hierarchical structure up to word level, and consideration of contextually dissimilar paragraphs or sentences in addition to contextually similar paragraphs sentences may also be important.

Further, recent summarization-based studies for incongruent news article detection [30, 29, 143] summarize body of the news article and subsequently determine similarity between headlines and summary of the news bodies. As the summarization in these studies are biased towards the dominant content of the body, such summarization may fail to capture the embedding noise present in partially incongruent news articles. To overcome such limitations in summarizations-based study [21] proposed a Multi-head Attention Dual Summarization-based model which splits the news article sentences into two sets based on their similarity with the headline. And generate a summary of each set separately for incongruent news article detections. In this thesis, our proposed model *GraSHE* focuses on a hierarchical structure. However, unlike existing hierarchical methods, the proposed methods extend the hierarchical structure of the news body from body to word level, and also assigns incongruity weight to each paragraph based on its inability to represent other paragraphs of the body.

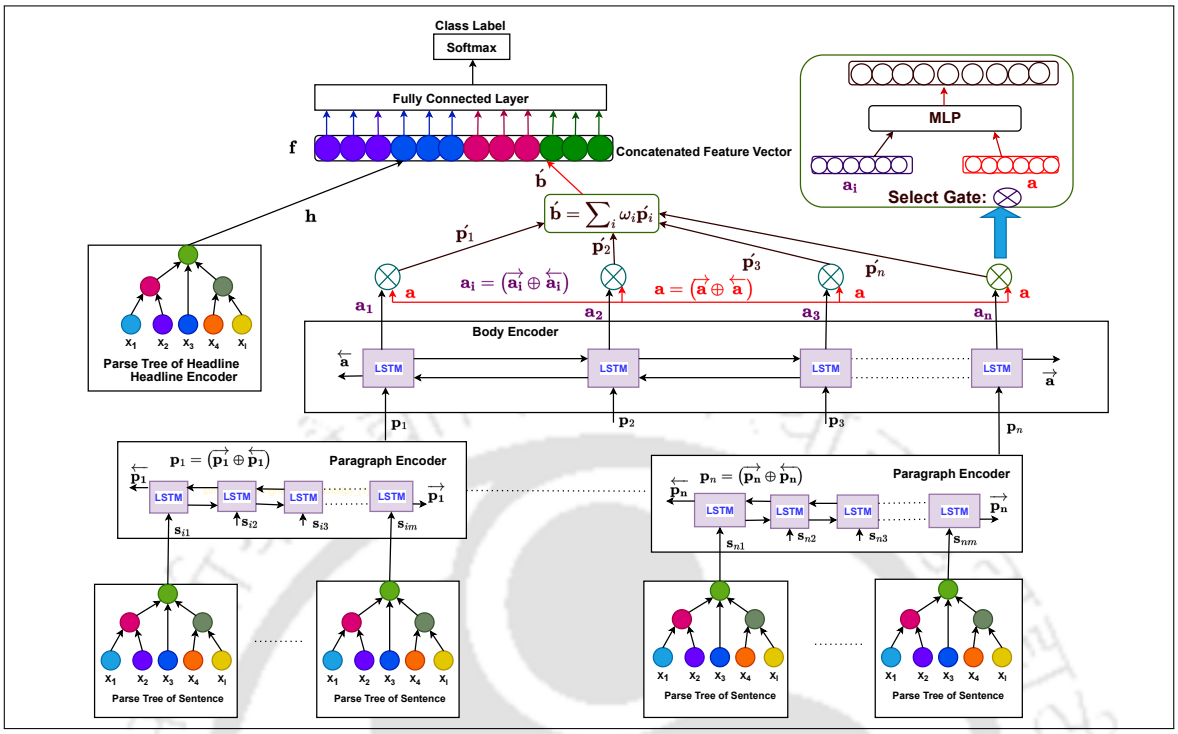


Fig. 4.1 Schematic diagram of the proposed **GraSHE** model. Unidirectional LSTM can also be used as a paragraph and body encoder. Encoded represented of the body is represented by $\overrightarrow{\mathbf{a}}$ and similarly $\overrightarrow{\mathbf{p}}_i$ is considered as representation of i^{th} paragraph. Incongruity weight ω_i is estimated using equation 4.15, 4.16 and 4.17 .

4.5 Proposed Framework: Gated Recursive And Sequential Deep Hierarchical Encoding

As mentioned above, the objective of the Chapter is to study the effect of two important aspects of encoding news articles while detecting incongruity; (i) the effect of deeper hierarchical encoding by extending the hierarchical structure of news body from the body to the word level (ii) incorporating the incongruity weight of different paragraphs defining the inability to represent the encoding of the entire document.

Figure 4.1 shows the schematic diagram of the proposed `Gated Recursive And Sequential Deep Hierarchical Encoding` *(GraSHE)* model. It defines the hierarchical structure as follows. A body $\mathcal{B}$ of a news article is a sequence of n paragraphs, $\mathcal{B}$=$\{\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_n\}$. A paragraph P_i is a sequence of m sentences, $\mathcal{P}_i$=$\{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_m\}$. A sentence $\mathcal{S}_j$ is defined recursively by a dependency parse tree $\mathcal{S}_j$ consisting of l words, $\mathcal{S}_j$=$\{w_1, w_2, \cdots, w_l\}$. The body is encoded using a gated sequential model over paragraphs. The paragraphs are encoded

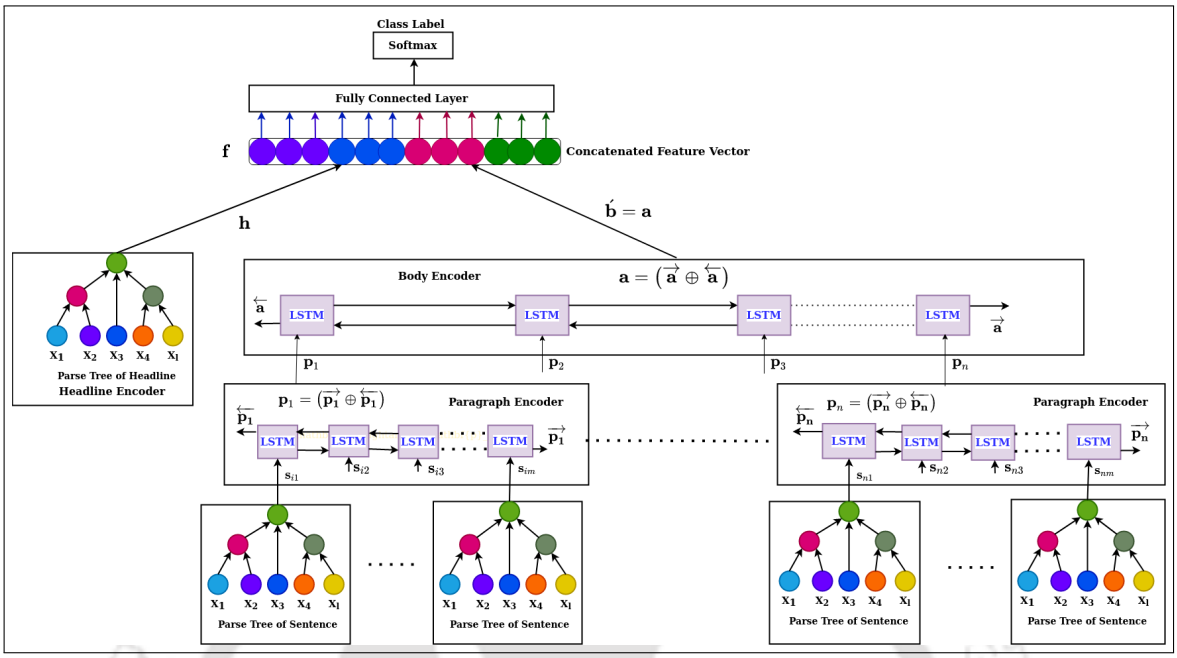


Fig. 4.2 Schematic diagram of the proposed **RaSHE** model.(i) **RaSHE**$^{(\mathbf{Ui})}$ is similar to **RaSHE** except bidirectional LSTM is replaced by unidirectional LSTM for paragraphs and body encoding. (ii) **HoBERT** is similar to **RaSHE**. Only BERT is used as headline and sentence encoder instead of child-sum Tree LSTM. (iii) **HeLSTM** is also similar, as **RaSHE** unidirectional LSTM has used to encode sentences and headlines, and unidirectional LSTM is also used as a paragraph and body encoder.

using a sequential model over the sentences. The sentences are encoded using a tree-based encoding model over the dependency parse tree of the sentence. Similarly, the headline is also encoded using a tree-based model.

Instead of recursive, earlier studies [28][1] of incongruity detection have considered only sequential structure. Earlier studies [78] [41][79] have reported that recursive encoding of sentences, such as dependency parse trees, helps in capturing long-term dependencies between words and the syntactic structure of the sentence. The studies in [78] [41][79] have also observed better performance for various applications such as semantic relatedness of sentence pairs, sentiment classification and natural language interface. Motivated by these observations, we proposed to use a dependency parse tree to represent sentences. In the literature, several neural network-based encoders such as *mTreeLSTM* [79], *child-sum Tree LSTM* [41], *tree-transformer* [144] are proposed to encode trees. Because of its ability to capture long-term dependency, child-sum Tree LSTM [41] has been considered in our proposed model. Ideally, we should be able to use any suitable state-of-the encoder for the same purpose. Given a sentence S and its dependency parse tree, let $ch(j)$ denote the set of

children nodes of node j. The hidden state of a node j is defined by the sum of the initial hidden states of its children nodes, as follows.

$$\acute{\mathbf{h}}_j = \sum_{k \in ch(j)} \mathbf{h}_k \tag{4.1}$$

Using the initial hidden state of node j i.e., $\mathbf{h}_j$, the corresponding input, output and intermediate gates of node j are estimated as follows.

$$\mathbf{i}_j = \sigma\left(\mathbf{W}^{(i)}\mathbf{x}_j + \mathbf{U}^{(i)}\acute{\mathbf{h}}_j + \mathbf{b}^{(i)}\right) \tag{4.2}$$

$$\mathbf{o}_j = \sigma\left(\mathbf{W}^{(o)}\mathbf{x}_j + \mathbf{U}^{(o)}\acute{\mathbf{h}}_j + \mathbf{b}^{(o)}\right) \tag{4.3}$$

$$\mathbf{u}_j = tanh\left(\mathbf{W}^{(u)}\mathbf{x}_j + \mathbf{U}^{(u)}\acute{\mathbf{h}}_j + b^{(u)}\right) \tag{4.4}$$

Where $\mathbf{x_j}$ denotes the embedding of the associated word w_j, $\mathbf{b}^{(\cdot)}$ denotes the bias, $\mathbf{W}^{(\cdot)}$ and $\mathbf{U}^{(\cdot)}$ denote the parameter matrices for respective gates. Unlike traditional LSTM, child-sum tree LSTM has multiple forget gates, one for each child node. It allows each child node to incorporate the information selectively. Forget gate for the k^{th} child of the node j is defined as follows.

$$\mathbf{f}_{jk} = \sigma\left(W^{(f)}\mathbf{x}_j + \mathbf{U}^{(f)}\mathbf{h}_k + \mathbf{b}^{(f)}\right) \tag{4.5}$$

The final cell state and hidden state of node j are defined as follows.

$$\mathbf{c}_j = \mathbf{i}_j \odot \mathbf{u}_j + \sum_{\mathbf{k} \in \mathbf{ch}(j)} \mathbf{f}_{jk} \odot \mathbf{c}_k \tag{4.6}$$

$$\mathbf{h}_j = \mathbf{o}_j \odot tanh(\mathbf{c}_j) \tag{4.7}$$

The hidden state of the root node defines the encoding of the sentence. A sequence of sentences defines a paragraph. Once encoding of the sentences is obtained, the encoding of a paragraph can be estimated using a sequential model such as Recurrent Neural Network (RNN), Gated Recurrent Units (GRUs), Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT) etc. We have considered unidirectional LSTM and bidirectional LSTM (BiLSTM) in this study.The prime motivation behind applying sequential encoding (LSTM or BiLSTM) over the encoding of sentences to obtain the paragraph's encoding is that. A paragraph of the news article body is a sequence of sentences, and every sentence in the paragraph is sequentially and contextually related to the previous and next sentences in the news body. If $\mathbf{s}_{ij}$ denotes the encoding of a sentence $\mathcal{S}_j$ in a paragraph $\mathcal{P}_i$, the encoding $\mathbf{p}_i$ of the paragraph $\mathcal{P}_i$ is obtained either with LSTM or

BiLSTM as follows.

$$\mathbf{p}_i = LSTM\left(\mathbf{s}_{i1}, \mathbf{s}_{i2}, ..., \mathbf{s}_{im}\right) \tag{4.8}$$

$$\mathbf{p}_i = BiLSTM\left(\mathbf{s}_{i1}, \mathbf{s}_{i2}, ..., \mathbf{s}_{im}\right) \tag{4.9}$$

The structure of the news body is considering a sequence of paragraphs. Every paragraph is sequentially and contextually related to the previous and next paragraph. Hence, we apply sequential encoding (LSTM or BiLSTM) over the encoding of paragraphs to obtain the encoding of the body. Encoding of a body $\mathcal{B}$ of a news article can be learned from the sequence of underlying paragraph encoding using either LSTM or BiLSTM, as defined below.

$$\mathbf{a} = LSTM\left(\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n\right) \tag{4.10}$$

$$\mathbf{a} = BiLSTM\left(\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n\right) \tag{4.11}$$

where $\mathbf{a}$ represents encoding of the body $\mathcal{B}$. As the encoding of the body $\mathbf{a}$ is likely to bias either by the first and the last paragraph depending on the direction of the sequence (an inherent effect of LSTM models), we regularize the encodings of the intermediate LSTM units with reference to a global encodings of the first and last LSTM units) through a gate as shown in Figure 4.1. Like in [145], we employ a multi-layer perceptron based select gate as defined in equation 4.12 and 4.13.

$$\mathbf{c}_i = \sigma\left(\mathbf{W}^p \mathbf{p}_i + \mathbf{U}^p \mathbf{a} + \mathbf{b}^p\right) \tag{4.12}$$

$$\acute{\mathbf{p}}_i = \sigma\left(\mathbf{p}_i \odot \mathbf{c}_i\right) \tag{4.13}$$

where W^p, U^p and b^p are the select gate parameters and $\odot$ denotes element wise multiplication. The main motivation behind using a select gate is to capture the importance of intermediate paragraphs while generating an encoding of the body, which is defined by a weighted summation of the gated representation of the paragraphs, as defined below.

$$\acute{\mathbf{b}} = \sum_i \omega_i \acute{\mathbf{p}}_i \tag{4.14}$$

where ω_i is the incongruity weight of the paragraph, $\mathcal{P}_i$ representing its inability to represent the encoding of the entire document (other paragraphs). If $\mathbf{M}$ represents an attention matching matrix of a paragraph with another paragraph, the matrix $\mathbf{M}$ is defined as below.

$$M_{ij} = \frac{exp(\acute{\mathbf{p}}_i^T \cdot \acute{\mathbf{p}}_j)}{\sum_{k \neq i} exp(\acute{\mathbf{p}}_i^T \cdot \acute{\mathbf{p}}_k)}. \tag{4.15}$$

where M_{ij} denotes the probability of paragraph $\mathcal{P}_i$ representing paragraph $\mathcal{P}_j$. The strength of the paragraph $\mathcal{P}_i$ (i.e., ω_i) is then defined by the Softmax of the entropy of the probability distributions of the paragraph $\mathcal{P}_i$ representing all other paragraphs, as follows.

$$\omega_i = \frac{exp(H_i)}{\sum_{k \neq i} exp(H_k)} \tag{4.16}$$

where $H(i)$ is the entropy of paragraph $\mathcal{P}_i$.

$$H(i) = -\sum_{j \neq i} \Big(1 - M_{ij}\Big) \log \Big(1 - M_{ij}\Big) \tag{4.17}$$

The higher the entropy, the higher is the strength of representation. Since the underlying task is incongruent news article detection, hence paragraph which cannot represent other paragraphs of news article should be given higher importance. So, we consider $(1 - M_{ij})$ for entropy estimation in equation 4.17. If the $Pr(\mathcal{P}_i \rightarrow \mathcal{P}_j)$ for all j, $j \neq i$ is uniformly distributed, it indicates that $\mathcal{P}_i$ represents all other paragraphs equally likely, i.e., $\mathcal{P}_i$ can represent the body. If we want to assign equal weight to all the paragraphs, then we can simply set all the weights to, $1/n$ i.e., $\omega_i = 1/n, \forall i = 1..n$ (it is equivalent to average). We denote this model as $GraSHE^{(=)}$ in the experimental results and discussion section. Once we obtain the encoding of the body $\acute{\mathbf{b}}$ and headline $\mathbf{h}$ (obtained using child-sum Tree LSTM), we further estimate the following two vectors **sim** and **diff** capture the similarity and the difference between the body and headline.

$$\mathbf{sim} = \acute{\mathbf{b}} \odot \mathbf{h} \tag{4.18}$$

$$\mathbf{diff} = \acute{\mathbf{b}} - \mathbf{h} \tag{4.19}$$

Now, we define the final feature for the classification as follows.

$$\mathbf{f} = \bar{\mathbf{b}} \oplus \mathbf{h} \oplus \mathbf{sim} \oplus \mathbf{diff} \tag{4.20}$$

where $\oplus$ denotes concatenation of vectors. The feature vector $\mathbf{f}$ is then passed through a dense layer with a *Softmax* output layer. We apply cross entropy loss to learn the parameters.

4.6 Evaluation Methodology

In this section, we describe the experiment settings for accessing the performance of the candidate methods. We also provide reproducibility information for the shown results and analyses.

4.6.1 Dataset Characteristics

This study uses three publicly available datasets, namely ISOT fake news dataset [96] [95], FNC dataset [22], and NELA-17 dataset [97] [28]. The FNC dataset has four classes, namely *agree*, *disagree*, *discuss*, and *unrelated*. The samples in agreed, disagree, discuss classes are merged and named a *True* class, whereas the samples in unrelated class are considered *fake* class. For NELA dataset, we curate the samples following the procedure reported in [28][3] over the news corpus provided at [97] [4]. The news articles published by authenticated sources are labelled as *true* class, and the fake samples are generated by randomly inserting paragraphs from another news article into true class news articles. Section 3.2 presents further details of datasets, and Table 3.1 presents the characteristics of these datasets.

4.6.2 Experimental setups

This study uses Google's worde2vec [44] pretrained word embeddings. F-measure (F), Accuracy (Acc) and class-wise F-measure have been used as evaluation metrics. We consider the default setting of the pretrained BERT as defined in[5]. We have used the Standford CoreNLP dependency parser[6] to convert a sentence into a sentence tree based on the syntactic structure. Subsequently, we pass the sentence tree (dependency parse tree of sentence) to *child-sum Tree LSTM* [41] to encode the sentence tree and obtain an encoded representation of the sentence. Table 6.1 presents the details of hyperparameters that have been used to produce the results presented in this thesis. Though we have experimented with different values of hyperparameters, Table 6.1 presents the value which gives the best accuracy over the development split of the dataset. Table 4.2 presents the hyperparameters value related to the hierarchical structure of news articles. The hyperparameter value presented in the table is obtained through statistical analysis. We first counted the number of paragraphs

[3] NELA-17 dataset generation code
[4] NELA-17 dataset corpus
[5] Pretrained BERT By Huggingface
[6] Standford CoreNLP dependency parser

Table 4.1 Details of hyperparameters used to produce results.

Hyperparameters	Values
Epoch	40
Batch Size	50
Word Embedding Dimension	200
Learning Rate	0.01
Loss Function	Cross Entropy
Cell State Dimension of LSTM	100
Hidden State Dimension of LSTM	100
Number of Layer in MLP	2
Hidden State Dimension of child-sum Tree LSTM	100
Cell State Dimension child-sum Tree LSTM	100

Table 4.2 Details of hyperparameters related to the hierarchical structure of news articles.

Datatset	Hyperparameter	Value
FNC	Number of sentence within a paragraph Number of paragraphs within a news body,	5 18
ISOT	Number of sentence within a paragraph Number of paragraphs within a news body	5 5
NELA	Number of sentence within a paragraph Number of paragraphs within a news body	22 5
Number word in a sentence for LSTM in **HeLSTM** model		12

within a news article for all sample datasets, then verified the minimum, maximum and average number count of paragraphs within a dataset. We selected hyperparameter values satisfied by a minimum of 75% samples in a dataset. For example, at-least 75% of news articles have a minimum of 18 paragraphs in the FNC dataset. Similarly, we select other hyperparameter values presented in Table 4.2. We use padding (a random vector) if the number of sentences or paragraphs is less than the number of sentences and paragraphs defined in Table 4.2. Similarly, in case the number of sentences or paragraphs is more than the number of sentences and paragraphs defined in Table 4.2 we discarded the sentences and paragraphs. Our code repository is publicly available[7] to reproduce the proposed models' results.

[7] Deep-Hierarchical-Encoding Code Repository

4.6.3 Baseline Models

To compare the performance of the proposed methods with the state-of-art models from the literature, we have considered the following baseline systems.

- ***(FNC-1)*** [22]: It is the system provided by the $FNC-1$ organizer. It uses gradient gradient-boosting classifier with manually identified features. We consider the default hyperparameters as defined in [8].
- `XGBoost`: Talo XGBoost model [9] trains an ensemble classifier XGBoost over several feature vectors such as TF-IDF, overlapping n-grams, count of refuting words, SVD and sentiment features etc. We consider the default hyperparameters as defined in https://github.com/Cisco-Talos/fnc-1.
- `UCLMR` [35]: It is a multi-layer perceptron-based classifier built over n-gram features in headlines and bodies, and the similarity between them. We consider the default hyperparameters as defined in [10].
- `StackLSTM` [24] [11]:It combines various topic modelling features (LSI-topic, NMF-topic, NMF-cos, LDA-cos) with embedding obtained with StackLSTM over the top two hundred words in the news article.
- `Attentive Hierarchical Dual Encoder` ***(AHDE)*** [28][12]: It is the first hierarchical model reported in literature for detecting incongruity.
- `Graph-based Hierarchical Dual Encoder` ***(GHDE)*** [1][13]: It is the most recent study in incongruity detection. As it needs paragraph-level annotations, it has been tested only with the NELA dataset, where the inserted paragraphs are annotated as fake.
- `Fake News Detection Using Summarization` ***(GSFD)*** [143]: It uses a summarization-based method for fake news detection by exploiting contextual graph relation between sentences in a document.

In addition to the above state-of-the-art baseline models from the literature, we further build the following baseline to study the effectiveness of our proposed model.

- `LSTM`: This is a non-hierarchical sequential model applied over concatenated headline and body using Long short-term memory (LSTM) [123]. The hidden state of the last input is passed to a fully connected layer classifier.

[8] FNC-1
[9] XGBoost
[10] UCLM
[11] StackLSTM
[12] ADHE
[13] GHDE

- `Multi-Layer Perceptrons (MLP)`: Motivated by XGBoost, a multilayer perceptron classifier is over the *n*-gram overlapping features between headline and body, TF-IDF, SVD, and sentiment features used in XGBoost.
- `Hierarchical encoding with LSTM (HeLSTM)`: This model is similar to, $RaSHE^{(Ui)}$ except that unidirectional LSTMs are applied to encode the sentences, instead of Tree-LSTMs. The hidden state of the last word in a sentence is considered as the encoded representation of the sentence. All other setting of *HeLSTM* is similar to $RaSHE^{(Ui)}$.
- `BERT` [75]: Considering the encouraging observations of BERT (for various NLP tasks) in recent studies, we also build a classifier using BERT embedding generated over the text in the headline and body. First, we concatenate the headline and body. If the number of tokens in the concatenated headline and body exceeds 512 tokens, we divide the concatenated headline and body into several segments, where the number of tokens in each segment is less or equal to 512 tokens. Encoding obtained for each segment using pre-trained BERT are concatenated and passed to a fully connected layer, followed by Softmax for classification.
- `Recurrence over BERT (RoBERT)`: To compare the performance of our proposed model and its variants with hierarchical transformer-based model, this study considers the hierarchical transformer-based model (RoBRT) proposed in study [125]. We consider each sentence and headline of a news article as a segment, obtain an encoded representation of each segment using BERT, and then apply LSTM over encoded representations of segments. The hidden state vector of the last segment is passed to a fully connected layer and Softmax for classification. Further, the details of *RoBERT* can be studied in [125].
- `Hierarchy over BERT (HoBERT)`: *HoBERT* model is similar to, *RaSHE* except that BERT is applied to obtain the encoded representation of a sentence instead of Tree-LSTM. Figure 4.2 presents a block diagram of *HoBERT* and *RaSHE*. The primary motivation behind this model is to capture a deep hierarchical structure of news articles over the encoding of sentences using BERT.

4.7 Performance Analysis

Table 4.3 compares the performance of different methods over three datasets. As shown in the table, the methods are grouped into two; *baseline systems* and *proposed systems*. We first study the responses of the baseline systems, which are further sub-grouped into *Features* and *Encoding* based models. As noted in section 4.6.1, the three datasets, namely ISOT, FNC, and NELA-17 differ widely in their characteristics. The incongruent news articles in NELA and FNC datasets are synthetically created by embedding noise (randomly selected

Table 4.3 Comparison of the performances of different models over three benchmark datasets. Here, (Acc) and (F) indicate accuracy and F-measure, respectively. Similarly, (Cong.) and (Incong.) indicate F-measure of congruent and incongruent class, respectively.

			NELA-17				ISOT				FNC			
		Models	**Acc**	**F**	**Cong.**	**Incong.**	**Acc**	**F**	**Cong.**	**Incong.**	**Acc**	**F**	**Cong.**	**Incong.**
Baseline Systems	**Features**	*FNC* − 1	0.586	0.586	0.564	0.608	0.844	0.844	0.847	0.842	0.586	0.496	0.282	0.709
		XGBoost	**0.699**	**0.699**	0.694	0.704	0.989	0.989	0.989	0.989	**0.977**	**0.971**	0.958	0.984
		UCLMR[35]	0.589	0.588	0.608	0.569	**0.997**	**0.997**	0.997	0.997	0.964	0.955	0.934	0.975
		MLP	0.629	0.629	0.574	0.628	0.985	0.985	0.985	0.985	0.926	0.909	0.939	0.909
		StackLSTM[24]	0.597	0.591	0.541	0.641	0.992	0.992	0.992	0.992	0.971	0.963	0.946	0.982
	Encoding	*LSTM*	0.555	0.555	0.563	0.547	0.99	0.99	0.99	0.99	0.616	0.504	0.269	0.740
		HeLSTM	0.602	0.602	0.607	0.598	0.997	0.997	0.997	0.997	0.689	0.597	0.402	0.7901
		AHDE[28]	**0.606**	**0.606**	0.614	0.598	0.913	0.913	0.909	0.909	0.666	0.487	0.158	0.797
		GHDE [1]	0.55	0.331	0.331	0.332	-	-	-	-	-	-	-	-
		GSFD [143]	0.533	0.532	0.550	0.5153	**0.998**	**0.998**	0.998	0.998	0.615	0.493	0.244	0.742
		BERT	0.572	0.563	0.624	0.503	0.894	0.894	0.894	0.891	**0.722**	0.524	0.21	0.838
		RoBERT	0.615	0.613	0.54	0.642	0.996	0.996	0.996	0.996	0.664	0.583	0.4	0.767
		HoBERT	**0.635**	**0.634**	0.65	0.618	0.991	0.991	0.991	0.991	0.686	**0.632**	0.491	0.773
Proposed Systems	**Encoding**	$GraSHE^{(Ui,=)}$	0.664	0.663	0.629	0.632	**0.999**	**0.999**	0.999	0.999	0.715	0.629	0.366	0.814
		$GraSHE^{(=)}$	**0.70**	**0.699**	0.71	0.688	**0.999**	**0.999**	0.999	0.999	0.723	0.528	0.226	0.831
		$GraSHE^{(Ui)}$	0.63	0.63	0.66	0.668	0.998	0.998	0.998	0.998	0.718	0.505	0.028	0.835
		GraSHE	0.676	0.675	0.679	0.672	0.998	0.998	0.998	0.998	0.715	0.514	0.203	0.826
		$RaSHE^{(Ui)}$	0.652	0.652	0.642	0.661	**0.998**	**0.998**	0.998	0.998	0.711	0.624	0.442	0.805
		RaSHE	0.677	0.677	0.678	0.999	**0.999**	**0.999**	0.999	0.999	**0.743**	**0.668**	0.51	0.826
	Enc + Fea	$HeELSTM^{(F)}$	0.626	0.626	0.619	0.633	0.995	0.995	0.995	0.995	0.969	0.961	0.944	0.978
		$GraSHE^{(=,Ui,F)}$	**0.656**	**0.656**	0.656	0.656	**0.999**	**0.999**	0.999	0.999	**0.971**	**0.964**	0.948	0.98
		$RaSHE^{(Ui,F)}$	0.652	0.651	0.648	0.655	0.999	0.999	0.999	0.999	**0.964**	**0.955**	0.935	0.975

paragraph from other document), and by taking headline and body from different news articles, respectively. Thus, NELA has embedded noise, whereas FNC has coherent content in the bodies. Furthermore, the incongruent news articles in ISOT dataset are fake news articles published and reported as fake by third-party fact-checker. Because of the differences in dataset characteristics, some classifiers respond differently over different datasets.

Feature vs Encoding: Except for FNC dataset, all the classifiers (in both features and encoding) provide results with small marginal differences. The wider margin is observed in FNC, and it is because of its test samples drawing from different domain, which is verified section 4.7.3. It indicates that with carefully chosen features, one can get a comparable, even superior, performance as compared to state-of-the-art encoding methods. For instance, XGBoost outperforms others (both feature and encoding-based encoding) over NELA and FNC datasets. Therefore, *incorporating explicit features may still be important for boosting the performance of encoding-based models.*

Hierarchical vs Non-hierarchical: Among the encoding-based baseline methods reported in Table 4.3, except for the LSTM and BERT, all other models use hierarchical encoding over body. It is evident from the table that, in the majority of the cases, the hierarchical encoding-based models provide superior performances as compared to their non-hierarchical counterparts, for all the datasets. It can also be noted that many of the hierarchical models outperform many of the models with explicit features. Therefore, *hierarchical encoding is important for generating embedding of a large document.*

Dataset Characteristics: Response of a model greatly depends on the nature of the samples, and underlying distribution. Therefore, considering the three datasets with differing characteristics, the responses of the models also differ. From the Table 4.3, we can note the following important points.

- *Smaller the document, and smaller the number of paragraphs, the better is the classification performance*. ISOT is a balanced dataset curated from real sources with a small number of paragraphs and sentences. All the models provide comparable performances with a small margin.
- For the dataset with embedded noise (NELA dataset) where dominant contents are congruent with headline, both the feature and encoding-based methods perform poorly, but encoding-based methods have an edge over feature-based methods. Relatively poor performance of classifiers with explicit features (except for XGBoost) is due to high overlapping features between congruent and incongruent classes.
- *For the dataset with coherent content (ISOT and FNC datasets), both the feature and encoding-based methods respond effectively.* Poor responses of encoding methods over

FNC in Table 4.3 is due to difference of domain for the training and testing samples, which is further verified in section 4.7.3.

Motivated by the above contradicting responses over NELA and FNC datasets, we propose two methods, *GraSHE* for embedded noise (NELA) and *RaSHE* (non-gated version of *GraSHE* as shown in Figure 4.2) for coherent content (FNC). As both the *GraSHE* and *RaSHE* are deep hierarchical encoding-based models; therefore we mainly compare their performances with hierarchical encoding-based counterparts (namely, AHDE [28], GHDE [1], HeLSTM, RoBERT, and HoBERT). We also try the following forms of the proposed models to study the models from different perspectives.

$-$*GraSHE*: The proposed weighted model, as shown in Figure 4.1.

$-$*GraSHE*$^{(=)}$: Equal-weighted model of *GraSHE*.

$-$*GraSHE*$^{(Ui)}$: *GraSHE* with unidirectional LSTM, instead of BiLSTM.

$-$*GraSHE*$^{(Ui,=)}$: *GraSHE*$^{(=)}$ with unidirectional LSTM, instead of BiLSTM.

$-$*RaSHE*: Proposed non-gated model as shown in Figure 4.2.

$-$*RaSHE*$^{(Ui)}$: *RaSHE* with unidirectional LSTM, instead of BiLSTM.

From Table 4.3, it is evident that the proposed encoding methods outperform almost all the baselines methods (hierarchical, non-hierarchical) for all the datasets (except for FNC feature models). Further, *GraSHE*$^{(=)}$ dominates others over NELA dataset (embedded noise), and *RaSHE* dominates others for FNC dataset (coherent content). The strength of the proposed models lies with the following three key points; (i) *recursive encoding of sentences*, (ii) *hierarchical encoding of the document* and (iii) *gated normalization of the body encoding*. We further investigate the importance of these points below.

4.7.1 Effect of Recursive Encoding of Sentences

It is evident from Table 4.4 that both the *GraSHE* and, *RaSHE* outperform their hierarchical counterparts *AHDE* [28] and *GHDE* [1] over all datasets. It may be noted that *AHDE* and *GHDE* use sequential encoding of sentences, where *RaSHE* uses recursive encoding. Further, we also developed a sequential version of *RaSHE* i.e., $HeLSTM$, where $HeLSTM$ uses LSTM, and *RaSHE*$^{(Ui)}$ uses child-sum Tree LSTM. *RaSHE*$^{(Ui)}$ outperforms $HeLSTM$ by 8.12% and 3.33% over NELA and FNC datasets. Syntactic tree-based recursive models prove particularly beneficial for tasks demanding the representation of long-distance relations among words, such as semantic connections between nouns. Even though two tokens may appear distant in word sequence, they can still be structurally close to each other.The sentences within the body of news articles are often lengthy, both in terms of the

number of words and the complexity of their syntactic structure. Additionally, the writing style and syntactic structure of true news articles typically differ from those of fake news articles [27]. Hence, representing sentences from news articles in terms of syntactic trees and subsequently applying recursive encoding to these syntactic tree structures aids in capturing long-term dependencies between sentences and also captures the syntactic structure between words within sentences. Comparable findings have been reported in prior studies [78] [41] examining semantic relatedness and sentiment classification applications, demonstrating that recursive encoding-based models surpass sequential encoding-based models.

4.7.2 Effect of Hierarchical Encoding

It is evident from Table 4.4 that recursive encoding of sentences helps to capture a better representation of the body. It can also be validated from the comparison between $HeLSTM$ and $RaSHE^{(Ui)}$, where these models differ only in sentence encoding ($HeLSTM$ uses LSTM, and $RaSHE^{(Ui)}$ uses child-sum Tree LSTM). To confirm these observations, we further compare the performance of the proposed models with $HeLSTM$. It could be observed that our proposed models $GraSHE$, $GraSHE^{(=)}$ and $RaSHE$ outperform the $HeLSTM$. The superior performance exhibited by the proposed deep hierarchical encoding models, namely $GraSHE^{(=)}$ and $RaSHE$, compared to the sequential encoding-based model $HeLSTM$, establishes that deep hierarchical encoding-based models are effective in capturing incongruent news articles. Every sentence in the paragraph relates to the previous and next sentences in the paragraph, and every paragraph in the news body relates to the next and previous paragraph. Our proposed model $GraSHE^{(=)}$ and $RaSHE$ are designed to capture the above hierarchical properties of incongruent news articles. On the other hand, non-hierarchical encoding-based models treat the news body as a sequence of words, consequently failing to capture the hierarchical properties inherent in news articles. Hence, we can conclude that hierarchical encoding and deep hierarchical encoding methods are more effective for incongruent news classification.

Importance of Select Gate Comparing the performance of the proposed models $GraSHE^{(=)}$ and $RaSHE$ over the NELA dataset, as shown in Table 4.3 it is observed that the performance of $GraSHE^{(=,)}$ is superior to that of $RaSHE$. It is because the noise added in incongruent samples may be present in any part of the news article. Since the select gate also considers the hidden state of the intermediate's paragraph instead of considering only the concatenated hidden state of the last and first paragraph. Due to such consideration of select gate, the noise padded in the middle of news article also contributes while obtaining an encoded

representation of news article body. Hence, the performance of the $GraSHE^{(=)}$ is superior to *RaSHE* over the NELA dataset. For FNC dataset *RaSHE* is superior to $GraSHE^{(=)}$. This is due to the fact that there is no noise present in the FNC dataset. The samples of FNC dataset are true to nature. Only the actual news article headline and news article body are mismatched to create an unrelated class sample. Hence, considering the last and the first paragraph's hidden state may be sufficient while obtaining an encoded representation of the news article body. Further, comparing the performance of the proposed model with BERT and hierarchical transformer models, it can also be observed from Table 4.3, our proposed models outperforms *BERT* and its hierarchical models namely *RoBERT* and *HoBERT* over NELA, ISOT and FNC datasets. Our proposed model $GraSHE^{(=)}$ outperforms all other baseline systems over NELA and ISOT datasets.

Table 4.4 Performance of sequential encoding of sentence versus recursive encoding of sentence structure by exploiting the hierarchical structure of news article.

	NELA-17		**ISOT**		**FNC**	
Model	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**
GHDE[1]	0.550	0.330	-	-	-	-
AHDE[1]	**0.606**	**0.606**	0.913	0.913	0.666	0.487
HeLSTM	0.603	0.603	**0.997**	**0.997**	**0.689**	**0.597**
$GraSHE^{(Ui,=)}$	0.664	0.663	**0.999**	**0.999**	0.715	0.629
$GraSHE^{(=)}$	**0.70**	**0.699**	**0.999**	**0.999**	0.723	0.528
$GraSHE^{(Ui)}$	0.63	0.63	0.998	0.998	0.718	0.505
GraSHE	0.676	0.675	0.998	0.998	0.715	0.514
$RaSHE^{(Ui)}$	0.652	0.652	**0.999**	**0.999**	0.712	0.624
RaSHE	0.677	0.652	**0.999**	**0.999**	**0.743**	**0.668**

Table 4.5 Comparison performance of hierarchical structure-based model versus non-hierarchical sequential model.

		NELA-17		**ISOT**		**FNC**
Model	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**
LSTM	0.555	0.55	0.991	0.99	0.616	0.504
HeLSTM	0.603	0.603	**0.997**	**0.997**	**0.689**	**0.597**
AHDE[28]	**0.606**	**0.606**	0.913	0.913	0.666	0.487

4.7.3 Domain Dependency

The domain of the news article gives insight into the news article. So, it becomes crucial to study the impact of the domain on incongruent news article detection tasks. *Is incongruent news article detection domain-independent or domain-dependent? What happens if models for incongruent news article detection are trained over news articles from a certain domain and tested over news articles from a different domain.* We conduct an empirical ablation study over the FNC dataset to answer such questions. The FNC dataset provided by the FNC-1 contest has different domain distributions in the training and test set. The training set has news articles from 200 domains, and the test has news articles from 100 domains, and there is no common domain in the training and test set. From table 4.3, it can be observed that the performance of deep learning-based models is inferior compared to feature-based models. The main reason behind the poor performance of the deep learning-based model over FNC data sets is that the news articles from the train and test sets belong to different domains. Another potential reason for this could be that feature-based models heavily depend on lexical overlap between the headline and news body. While training and test sets may encompass different topics, feature-based models still need to distinguish between the headline and news body based on lexical overlap features. In contrast, deep learning-based models learn parameters based on underlying patterns within the training data. Consequently, they may struggle to adapt to topic drift between the training and test sets for incongruent news detection. To confirm this observation, we created another dataset as follows: *we merged the train and test set provided by the FNC organizer and randomly permuted the sample in the merged FNC datasets. We then create, a train and test set with the same distribution as the original FNC dataset.* We called this newly created data set FNC domain overlap dataset FNC^R. From table 4.6, it can be observed that the performance of every deep learning-based model significantly improved over the FNC^R data set compared to the performance of these models over the FNC dataset. From table 4.6 it can be clearly observed that our proposed model $RaSHE$ outperformed $AHDE$[28], $GSFD$[143], $HeLSTM$, $RoBERT$ and $HoBERT$. This further validates our claim that extending the hierarchical structure of the news body from the body to the word level (deep hierarchical encoding of the news body) significantly improves the model's performance for incongruent news article detection tasks. Hence, we can conclude that incongruent news article detection is domain-dependent. So training and test sets should have news articles from the same domain.

Table 4.6 Comparison of performance of models over FNC dataset with different domain distribution in train and test versus FNC with similar domain distribution ($\boldsymbol{FNC^R}$) in train and test. ↑ denote the improvement of performance of model over ($\boldsymbol{FNC^R}$) compared to performance of same model over **FNC**. Here, (Acc) and (F) indicate accuracy and F-measure, respectively. Similarly, (Cong.) and (Incong.) indicate F-measure of congruent and incongruent class, respectively.

	FNC				$\boldsymbol{FNC^R}$			
Model	**Acc**	**F**	**Cong.**	**Incong.**	**Acc**	**F**	**Cong.**	**Incong.**
AHDE[28]	0.673	0.4477	0.158	0.797	0.681↑	0.472↑	0.140	0.804
GSFD [143]	0.615	0.493	0.244	0.742	0.829↑	0.752↑	0.614	0.890
HeLSTM	**0.689**	0.597	0.402	0.79	0.809 ↑	0.764 ↑	0.661	0.867
RoBERT	0.664	0.583	0.4	0.79	0.828↑	0.755 ↑	0.622	0.888
HoBERT	0.686	**0.632**	0.491	0.773	**0.861**↑	**0.823**↑	0.741	0.905
$GraSHE(Ui,=)$	0.715	0.59	0.366	0.814	0.847 ↑	0.809 ↑	0.722	0.894
$GraSHE(=)$	0.723	0.528	0.226	0.831	0.842 ↑	0.804 ↑	0.718	0.89
$RaSHE(Ui)$	0.715	0.629	0.442	0.805	0.842 ↑	0.805 ↑	0.72	0.89
RaSHE	**0.743**	**0.668**	0.51	0.826	**0.876** ↑	**0.844**↑	0.775	0.914
GraSHE	0.715	0.514	0.203	0.826	0.798 ↑	0.748 ↑	0.637	0.86

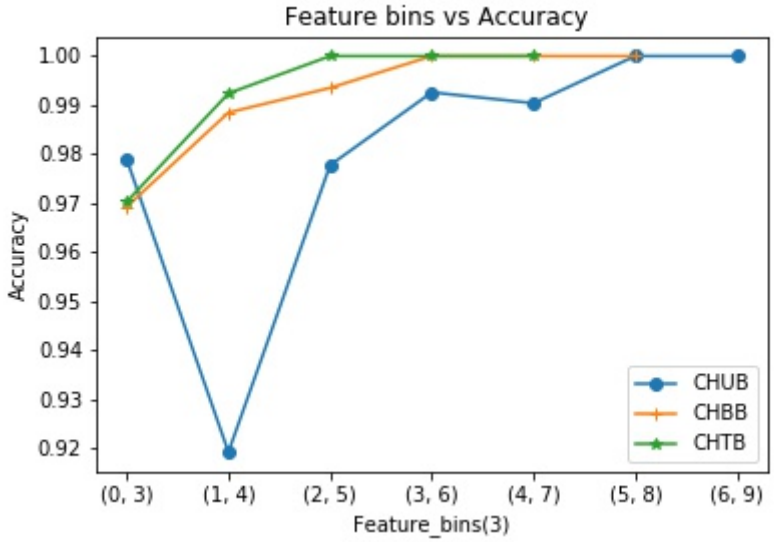


Fig. 4.3 The response of **RaSHE** model versus n-grams overlapping between headline and body over FNC dataset. Present the response of models versus n-grams overlapping between headline and body over FNC dataset. Here *CHUB* denotes the count of headline unigrams present in the body, *CHBB* denotes the count of headline bigrams present in the body, *CHTB* denotes the count of headline trigrams present in the body. Features bins indicate the count of overlapping n-grams.

Table 4.7 Empirical Study of Different Feature Sets: (i) **G1**: Denotes count overlapping features, (ii) **G2**: Denotes TF-IDF similarity between headline and body and SVD of headline, body along with similarity between them, (iii) **G3**: Denotes sentiment features extracted from headline and body, (iv) $\oplus$ denotes concatenation operation between features.

	NELA-17		**FNC**	
Model	**Acc**	**F**	**Acc**	**F**
$G1$	0.57	0.57	**0.962**	**0.951**
$G2$	0.475	0.443	0.653	0.649
$G3$	0.557	0.544	0.715	0.450
$G1 \oplus G2$	0.494	0.371	0.721	0.419
$G1 \oplus G3$	0.5	0.333	0.721	0.837
$G2 \oplus G3$	0.477	0.774	0.636	0.615
$G1 \oplus G2 \oplus G3$	**0.629**	**0.629**	0.926	0.906

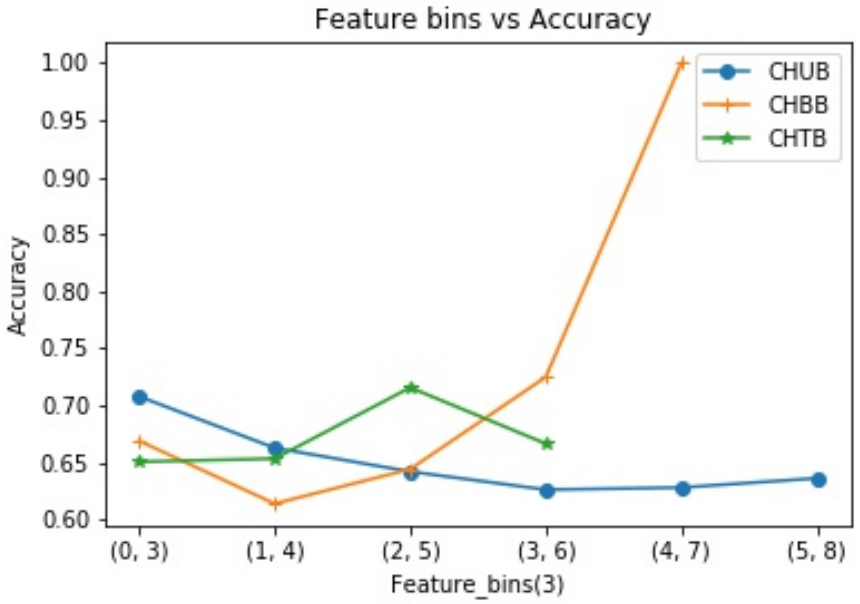


Fig. 4.4 The response of **RaSHE** model versus n-grams overlapping between headline and body over NELA dataset. Present the response of models versus n-grams overlapping between headline and body over NELA dataet. Here *CHUB* denotes the count of headline unigrams present in the body, *CHBB* denotes the count of headline bigrams present in the body, *CHTB* denotes the count of headline trigrams present in the body. Features bins indicate the count of overlapping n-grams.

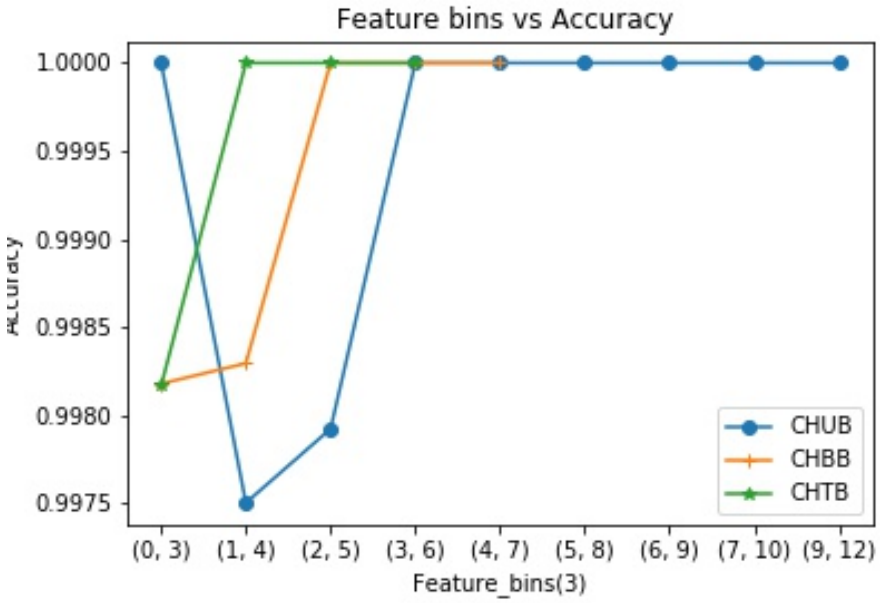


Fig. 4.5 The response of **RaSHE** model versus n-grams overlapping between headline and body over ISOT dataset. Present the response of models versus n-grams overlapping between headline and body over ISOT dataset. Here *CHUB* denotes the count of headline unigrams present in the body, *CHBB* denotes the count of headline bigrams present in the body, *CHTB* denotes the count of headline trigrams present in the body. Features bins indicate the count of overlapping n-grams.

4.7.4 Effect of Explicit Features

From Table 4.3, it is evident that feature-based models are domain-independent. This chapter conducts an empirical study of feature sets to find a minimal feature set for the incongruent news article detection task. The primary motivation for conducting the empirical study is to identify the minimal feature set that facilitates the detection of incongruent news and determine the key features that contribute to incongruent news article detection. The feature set is divided into three groups. The first group feature set contains the count of n-grams features and overlapping count and ratio of n-grams between headline and body. The second group feature set includes TF-IDF similarity between headline and body, top k SVD headline and body and cosine similarity between SVD of headline and body. Similarly, the third group feature consists of headline and body sentiment scores. Positive, negative, average, and compound sentiment scores of headline and body were considered. Given the pair of headline $\mathcal{H}$ and body $\mathcal{B}$, we estimate count overlap, TF-IDF and SVD features and sentiment features as follows:

- **Count overlap features :** This feature estimates the similarity between the headline $\mathcal{H}$ and body $\mathcal{B}$ in terms of lexicon overlap. Count overlap features concatenate the following count overlapping information between the headline and body.
  1. Count of unigrams, bigrams, and trigrams in the headline and body.
  2. Count of overlapping unigrams, bigrams, and trigrams between the headline and body.
  3. Estimates n-grams overlap ratio by dividing the count of overlapping n-grams by n-grams in news body; here, n-gram refers to unigrams, bigrams and trigrams separately.

- **Term Frequency–Inverse Document Frequency (TF-IDF) and Singular Value Decomposition (SVD):** To learn the relation between the headline and body in terms of influence of words, co-occurrence of words and latent representations. We estimate TF-IDF and SVD similarities as follows:
  1. *TF-IDF :* First, we obtain the TF-IDF vector of both headline $\mathcal{H}$ and body $\mathcal{B}$ by estimating each unigrams term frequency and normalizing it by its inverse document frequency. Then we estimate cosine similarity between the TF-IDF vector of $\mathcal{H}$ and body $\mathcal{B}$.
  2. *SVD :* Singular Value Decomposition (SVD) [146] are popular methods for latent topic representation of documents. We obtain SVD features by concatenating the SVD of the headline, body and cosine similarity between the SVD of the headline and body. To obtain the SVD of the headline and body, we construct a headline-to-words matrix and body-to-words matrix, where an entry in the body-to-words and the headline-to-words matrix is the TF-IDF of each word. SVD is then applied over both headline word and body-to-word matrix. We retained the top 50 dimensions from both matrices in their decomposition. Subsequently, we estimate the cosine similarity between the SVD of the headline and the SVD of the body.
  3. **Sentiment features:** Fake news is created with various stylistic tricks to excite and exploit readers' sentiments [147][148][149]. To capture the sentiment relation between headline and body, we obtain positive, negative, neutral and compound polarity scores of the headline and body separately using the NLTK sentiment analyzer[14]. Next, we obtain the sentiment feature of the news article by concatenating the sentiment polarity score of the headline and body.

Table 4.7 presents the empirical study of the feature sets. From Table 4.7, it is evident that the count feature outperformed all other feature groups over the FNC dataset. However, a combination of all three feature groups outperformed all other feature groups for the NELA-17 dataset, which indicate that feature set and feature combinations are not uniform for every dataset for the incongruent news article detection task. Our empirical study suggested that

[14] NLTK sentiment analyzer

similarity between headline and body based on TF-IDF, SVD top-k vectors, and sentiment features play an essential role in incongruent news article detection. This study identifies 151 features (TF-IDF and SVD similarity between headline and body, SVD of headline and body, count overlapping and sentiment features) from the ten million-plus features used in Xgboost. We concatenate these 151 features with the concatenated feature vector obtained in equation 4.20 and pass to a fully connected layer. We merge the feature with following models namely $HeLSTM$, $RaSHE^{(Ui)}$, $GraSHE^{(=,Ui)}$ and named them $HeLSTM^{(F)}$, $RaSHE^{(Ui,F)}$ and $GraSHE^{(=,Ui,F)}$ respectively. Here, F indicate that concatenated features are added to models. From table 4.3, it is evident that by incorporating important features from a different domain, our proposed models provide comparable results over ISOT and FNC datasets. From table 4.3 by comparing the performance of $HeLSTM$, $RaSHE^{(Ui)}$ and $GraSHE^{(Ui)}$ with $HeLSTM^{(F)}$, $GraSHE^{(=,Ui,F)}$ and $RaSHE^{(Ui,F)}$. We can claim that incorporating explicit features makes models domain-independent for incongruent news article detection tasks. However, for the NELA-17 dataset, there is a reduction in performance after adding handcrafted features. It indicates that feature engineering needs an understanding of the underlying datasets. We further build another multi-layer perceptron-based classifier *MLP* over the 151 manual features to investigate the response of the features. From table 4.3, it is observed that MLP+Feature outperforms several other baselines. Dataset domain-specific feature engineering and adaptation of the proposed model are not included in this study but are left as a future task.

4.7.5 Effect of Overlapping n-grams Between Headline and Body

We further investigate the performance of models with respect to dataset characteristics. We consider the response of the *RaSHE* model over test set across FNC, NELA and ISOT datasets. We counted the number of overlapping unigrams, bigrams and trigrams between headline and body. We divided the samples into bins, and considered the accuracy of samples within the bins of 3. From Figure 4.3, it is evident that the performance of the *RaSHE* model increases as the count of unigrams, bigrams and trigrams overlap increases. This observation further validates the performance of models over the FNC dataset in Table 4.7 where the performance of *MLP* models is high for the count overlap feature ($G1$). From Figure 4.4, it is evident that the performance of the model is average irrespective of the unigrams, bigrams and trigrams overlap between headline and body. This could be a possible reason behind the downgrade in performance of models in Table 4.3 over the NELA dataset after adding explicit features. Such observation further validates the performance of the models over the NELA dataset presented in Table 4.3 and Table 4.7. From Figure 4.5, it is evident that

performance of *RASHE* is always high irrespective of n-grams overlapping between headline body. Such observations suggest that performance of models over ISOT dataset does not depend on overlapping between headline and body. This could be a possible reason behind the high performance of models over the ISOT dataset.

Summary

This study proposed *Gated Recursive And Sequential Deep Hierarchical Encoding* model, namely *GraSHE*, to detect incongruent news articles. The proposed models capture long-term dependencies between words and syntactic structures of sentences at the sentence level and sequential structures at the body and paragraph level. From various experiments over three datasets, it is observed that capturing structural properties at the sentence level improved the performance of incongruent news article detection tasks. The key observations from different empirical studies are as follows: i) Incongruent news articles is a domain-dependent task in case of encoding-based models, i.e., the models perform inferior if news articles in training and testing datasets are from different domains or topics (refer to subsection 4.7.3) ii) Encoding sentence structure instead of sequential encoding of sentence enhanced the performance (refer to subsection 4.7.2), and iii) Performance of hierarchical structure-based models are superior to non-hierarchical structure-based models (refer to Table 4.5). We identify the following four potential future research directions; (i) Incorporating features of different nature with the hierarchical modeling, (ii) identifying appropriate feature engineering for the datasets of different nature, and (iii) Devising appropriate document summarization to reduce document size for incongruent news article detection, and (iv) Devise a domain-independent deep learning model for incongruent news article detection.

Chapter 5

Dual Summarization for Detecting Incongruent News

With the increasing concerns of disinformation shared over digital platforms, detecting incongruent news articles has become an important research challenge. Earlier research on incongruent news article detection predominantly concentrates on globally encoding the entire body [23] [1] [28] or summarizing the body to align with the headline [30] [29]. However, these approaches often struggle to identify partially incongruent news articles, where only a few sentences or paragraphs are incongruent with the headline. In such cases, certain sentences or paragraphs within the news body may be incongruent with the headline, and global encoding methods fail to differentiate between paragraphs congruent to the headline and paragraphs incongruent to the headline and produce a unified encoding of the entire news body. As the summarization in these studies is biased towards the dominant content of the body, such summarization may fail to capture the incongruent sentences or paragraphs present in partially incongruent news articles. Motivated by the above observation, this thesis proposes two dual summarization-based methods : (i) *DuSum*, which splits news article body sentences into two sets of highly congruent sentences and poorly congruent sentences based on their relationship with the headline. Subsequently, generate two different summaries of both sets separately using convolution and contextualized LSTM. (ii) *Multi-head Attention Dual Summary MADS* based method, which also generates two types of summaries that capture the congruent and incongruent parts in the body separately. We conducted our experiments and several ablation studies over three publicly available benchmark datasets, namely FNC, ISOT and NELA datasets. Our experimental results and ablation studies show that our proposed models *MADS* and *DuSum* outperform the state-of-the-art baseline models

in literature and more effectively identify incongruent news articles of different natures, including partially incongruent news articles.

5.1 Introduction

Amid growing apprehensions regarding the proliferation of disinformation across digital platforms, the identification of incongruent news articles has emerged as a pivotal research endeavor[3, 2, 18–20]. A news article is said to be incongruent[1,2] if its headline misrepresents the claim made in its body [3, 4]. As reported in the studies [11, 9, 12], the influence of the news headline is persistent, and approximately 60% of the news circulated on social media is shared without reading the whole article [11] which leads to the spread of misinformation over digital platforms. Hence, it has become an essential task to detect incongruent news articles circulated on digital platforms.

Incongruent news detection is a sub-problem of fake news detection. An incongruent news article may be of different forms; such as (i) The headline is not related to its body, (ii) Both the headline and its body are related, but the claim made in the headline is different from its body, (iii) Both the headline and its body report a genuine event/incident, but the dates, numeric values or name entities are manipulated, (iv) Textual noise (paragraphs/sentences) extracted from other sources is inserted into a genuine news article (referred as partially incongruent news). As the proportion of the inserted noise as compared to the original content may be small, partially incongruent news articles are often fail to be detected by fact-checkers, and often used to spread misleading content or malicious propaganda[3].

Earlier studies on incongruent detection can be grouped into two categories, namely *similarity-based* and *summarization-based*. In a similarity-based approach, the idea is to learn the encoding of a news headline and its body and check their similarity. Initial studies [22, 23] on similarity-based approach utilize bag-of-words based features like n-gram, latent topic, TF-IDF, etc. to classify the news articles. Realizing the importance of contextual information present in the news articles, the studies in [98, 25] utilize both the contextual information and the above bag-of-words-features. Further, the studies in [26–28] exploit the hierarchical structure of news articles (body-paragraph-sentence relation) along with contextual and sequential information present in the news article.

[1] Fake news inform of misleading headline
[2] Fake news inform of misleading headline about WHO
[3] Misinformation matrix and types of misinformation

Our preliminary experiments on similarity-based methods, and also in the study [1], suggest that these methods perform well for small articles of few paragraphs but perform relatively poorly for large articles with the large number of paragraphs. To overcome such limitations, [1] propose graph-based methods to capture the similarity between news headlines and long news bodies. However, this method needs supervised information for every headline and paragraph in a pair of news articles. Creating such a fine-grained dataset is expensive. Unlike similarity-based methods, summarization-based studies [29–31] summarize the body of a news article to a synthetic headline. Then, the synthesized headline and the actual headline are matched. The synthetically generated headline from the news article body may not be a faithful or good representation of the news article body [32–34]. The above summarization-based methods for incongruent news article detection are biased towards the dominant content of the body and often fail to detect partially incongruent news articles.

Chapter 4 of this thesis introduces the Gated Recursive And Sequential Deep Hierarchical Encoding (*GraSHE*) method for detecting incongruent news articles. This approach extends the hierarchy structure of the news body from the body to the word level and incorporates incongruent weights. The *GraSHE* model captures long-term dependencies and syntactic structure by integrating sequential information at the paragraph and body levels (using BiLSTM), and syntactic structure at the sentence level (child-sum Tree LSTM [41]). Furthermore, unlike headline-guided attention models [28][27], *GraSHE* integrates incongruity weights to highlight non-dominant textual segments that are incongruent with other parts of the news body. The `Gated Recursive And Sequential Deep Hierarchical Encoding` (*GraSHE*) method aims to estimate the similarity between the encoding of the headline and the deep hierarchical encoding of the news body. Hence, `Gated Recursive And Sequential Deep Hierarchical Encoding` (*GraSHE*) is a similarity-based approach and may also fail to detect partially incongruent news. This is evident from the performance of *GraSHE* over the NELA dataset, which represents characteristics of partially incongruent news articles, where it performs averagely. Motivated by the aforementioned limitation of similarity-based methods and the `Gated Recursive And Sequential Deep Hierarchical Encoding` (*GraSHE*) approach (refer to Section 4.5 of Chapter 4), this chapter of the thesis proposes a dual summarization method for detecting incongruent news articles.

5.2 Motivation

As discussed in Section 5.1, existing similarity and summarization-based methods in the literature often fail to detect partially incongruent news articles. Motivated by the above limitations, this Chapter proposes dual summarization-based methods *Multi-head Attention Dual Summarization MADS* and *dual-summarization based approach,* namely **DuSum**, which divides the news article body into two sets, positive and negative set. Then, generate two different summaries of both the positive and negative sets. Next, match the headline with a summary of positive and negative for incongruent news article detection. To divide the sentences of the news body into positive and negative sets, this chapter explores two distinct approaches: (i) *Similarity Threshold*: Initially, the method assesses the similarity between the headline and each sentence. Subsequently, sentences with a similarity score exceeding θ are allocated to the positive set, while those below the threshold are assigned to the negative set. (ii) *Top-K and Least-k*: Initially, the similarity between the headline and each sentence is computed. Following this, the top-K sentences exhibiting the highest similarity to the headline are placed in the positive set, while the bottom-K sentences with the least similarity to the headline are allocated to the negative set, where k is a positive integer number.

5.3 Research Objective

1. This Chapter proposes a *Multi-head Attention Dual Summarization* **MADS** based summarization method which is capable of handling partially incongruent news by summarizing both the congruent and incongruent parts of the article body. The proposed method divides the body of the news article into two sets - *positive: highly congruent sentences with headline* and *negative: highly incongruent sentences with headline*. Further, for each set, different forms of representation are captured using multi-head attention and convolution. From various experiments over three publicly available benchmark datasets, it is observed that the proposed method outperforms the existing state-of-the-art baseline counterparts, including the dataset with partially incongruent news article.

2. This Chapter also proposes a *dual-summarization based approach,* namely **DuSum**, which splits the sentences in a news article into two sets; *highly congruent sentences with headline* and *poorly congruent sentences with headline*, and generates dual summaries using *headline guided attention*. If part of the news article is incongruent with the headline, then it is moved to a poorly congruent set, otherwise to a highly congruent set. The intuition is that while generating encoding of the body, one should incorporate both the congruent and incongruent parts of the body to enable capturing minority incongruent content as well. It then extracts summaries of the highly and

poorly congruent sets separately and performs matching between summaries and the headline.

5.4 Contributions

- This Chapter introduces two novel dual summary-based approaches denoted as **MADS** and **DuSum**. This method addresses the constraints observed in existing similarity and summarization-based models documented in the literature. Our proposed approach effectively identifies incongruent news articles across various characteristics.
- We investigate the effect of two methods to split the news body into two sets based on its contextual similarity with the news headline, i.e., *(i) Similarity threshold value* and *(ii) Top k and Least k*.
- We propose a sequence-weighted summation-based summarization method to find a summary vector for a set of sentences using contextualized LSTM. We further perform several ablation studies to establish the superiority of the proposed model.
- We conduct extensive experiments over three publicly available benchmark datasets to establish that the proposed dual summary-based methods **MADS** and **DuSum** are superior to similarity and summarization-based state-of-the-art models in the literature. We also analyzed the significance of different components of **DuSum** through several ablation studies.

5.5 Literature Review

In the literature, studies [130, 132–136, 150–154] have briefly reviewed and analyzed the works related to misinformation. Studies related to misinformation and disinformation detection can be grouped into several categories, such as evidence claim verification, clickbait detection, rumour detection [150], fake news detection in multi-modal data, fake news detection in social media networks, and incongruent news article detection. Among various forms of misinformation detection tasks, *clickbait* and *incongruent* news detection problems look similar and are often misunderstood, as both the problems relate to news headlines and news articles. Studies [6, 3, 141] briefly discussed the difference between incongruent news articles detection and clickbait detection. A clickbait can be detected based on the headline only, but an incongruent headline is determined by the connection between the headline and the news article body [6]. Clickbait attempts to attract the reader's attention, but incongruent news articles do not force readers to click some link and follow up [3]. This study retrospects work related to incongruent news article detection only.

As mentioned in the previous section, earlier studies on incongruent detection can be grouped into two categories; namely *similarity-based* and *summarization-based* methods. Initial studies [22, 23, 35, 138, 155] on similarity-based approaches utilise bag-of-word-based features like *n*-gram, latent topic, TF-IDF, etc. to classify the news articles. Realising the importance of contextual information present in the articles, the studies in [98, 25] utilise both the contextual information and the above bag-of-words based features. Further, the studies in [26–28, 141] exploit the hierarchical structure of news articles (body-paragraph-sentence relation) along with contextual and sequential information present in news articles. However, the above approaches perform poorly because the headline of the news article is usually short and concise, and the body of the news article is usually large. As the above methods do not perform well on large articles with large numbers of paragraphs, the study [1] proposes a graph-based method to capture the similarity between news headlines and individual paragraphs in the body by using an expensive paragraph-level annotation process. In the summarization-based methods, the study [30] applies an abstractive-based summarization to generate synthetic news headlines using generative adversarial network [156]. Next, estimate the similarity between synthesized headlines from the news body and actual headlines for incongruent news article detection. However, as reported in [157, 34], abstractive summarization may not generate a faithful or good representation of the news article body, especially for large news articles. The study [29] uses pretrained language models [158] to extract sentences related to the headline and generate a summary of the news body, then apply a match between the headline and extracted summary of the news body. Further, the study [143, 31] under summarization-based approach applies a graph-based extractive summarization to rank the sentences in the news article with reference to the headline and then generate a rank-weighted summation summary of news body for incongruent news article detection. Study [37] apply a part-of-speech-based hierarchical attention network to detect incongruent news articles. Recent study [159] proposed an explainable decision system and study [160] proposed a heterogeneous graph-based model for detecting fake accounts to counter the spread of fake news over social media.

5.5.1 Research Gaps: Summary

While a news article may become incongruent with the presence of a small incongruent text in the body, the majority of the above methods may be biased towards the dominant content (for instance, partially incongruent news articles). Unlike the above studies, our proposed method splits sentences in body into two sets based on their relationship with the headline and generates a summary of each set. Subsequently, we apply a match between the summary

of both sets to check consistency and similarity within body content. Similarly, we also apply a match between the headline and summaries of sets to verify the similarity and consistency between the headline and body contents. In this way, our proposed model **DuSum** checks consistency and similarity within body contents and consistency and similarity between headlines and body contents, which is important for incongruent news articles and fake news article detection.

5.6 Proposed Framework: Dual Summarization

This section discussed our dual summarization-based proposed models, *Multi-head Attention Dual Summarization* **MADS** and *Dual-Summarization* **DuSum**. Figure 5.1 present th wroking diagram of *Dual-summarization* **DuSum** and Figure 5.2 presents working diagram of *Multi-head Attention Dual Summarization* **MADS**.

5.6.1 Dual-Summarization based Approach DuSum

As mentioned above, the proposed method generates a *dual-summary* of a given news article : summary of the highly congruent sentences with the headline, and summary of the poorly congruent sentences with the headline. Given a news article $a =< h, b >$ with its headline h and body b, we first split the sentences in b into two sets $\mathcal{S}^+$ and $\mathcal{S}^-$ represent the set of highly congruent sentences, and the set of poorly congruent sentences. The primary rationale for partitioning body sentences into highly congruent sets ($\mathcal{S}^+$) and poorly congruent sets ($\mathcal{S}^-$) stems from addressing the complexities of partially incongruent news articles. In such instances, sentences highly contextually similar to the headline are assigned to the highly congruent sets ($\mathcal{S}^+$). In contrast, those least contextually similar to the headline are categorized into the poorly congruent set ($\mathcal{S}^-$). Likewise, in cases of fully congruent news articles, the majority of body sentences are anticipated to belong to the ($\mathcal{S}^+$) set, with only one few sentences placed into the ($\mathcal{S}^-$) set. Conversely, in the case of a fully incongruent news article, where every sentence in the body contradicts the headline, all sentences would be placed under the ($\mathcal{S}^-$) set, except one or a few sentences will be placed in ($\mathcal{S}^-$) set. Hence, splitting news bodies into highly congruent sets $\mathcal{S}^+$ and poorly congruent sets $\mathcal{S}^-$ helps in detecting partially incongruent news articles.

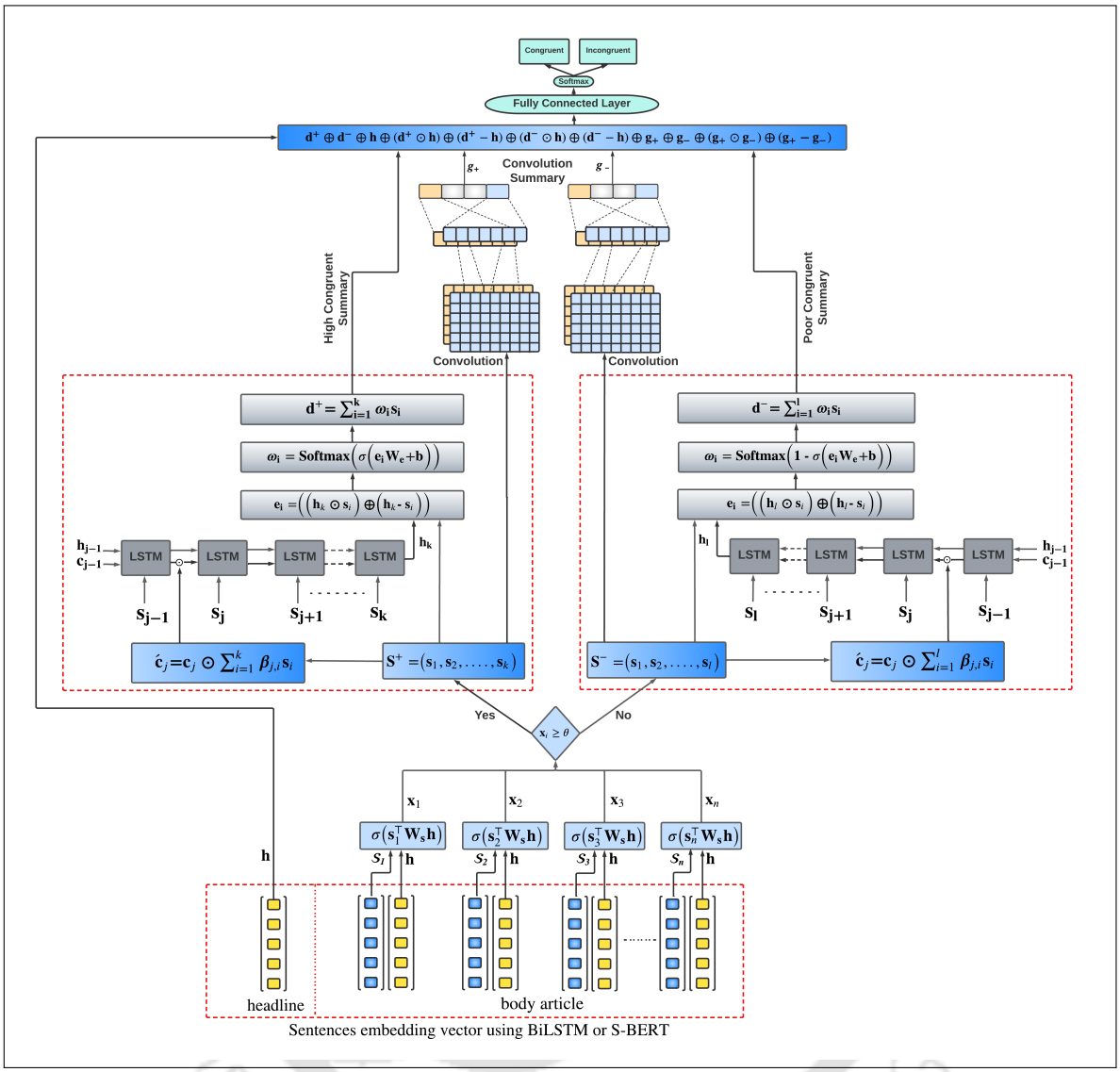


Fig. 5.1 The proposed *DuSum* model is depicted in this diagram. Initial sentence $\mathbf{s}_i$ and headline $\mathbf{h}$ encoding are obtained using BiLSTM or a sentence transformer. $\mathbf{x}_i$ defines the similarity between headline encoding $\mathbf{h}$ and sentence encoding $\mathbf{s}_i$. If $\mathbf{x_i} \geq \theta$, the sentence is placed in $\mathcal{S}^+$ set; otherwise, it is placed in $\mathcal{S}^-$ set. A sequential weighted summation summary $\mathbf{d}^+$ and $\mathbf{d}^-$, and a convolution summary $\mathbf{g}^+$, $\mathbf{g}^-$ are obtained using contextualize LSTM and CNN respectively for the sets $\mathcal{S}^+$ and $\mathcal{S}^-$. Thereafter, the similarity feature between the generated summaries and headline is obtained and passed to a fully connected layer for classification.

5.6.1.1 Headline Guided Attention

Every sentence in a news article's body has a different contextual relation, with the headline, in terms of contextual similarity. Sentences of the highly congruent $\mathcal{S}^+$ set will have high contextual similarity with the headline, and sentences of the poorly congruent $\mathcal{S}^-$ set will have low contextual similarity with the headline. To exploit such a relationship between the headline and sentences of the body, we apply headline-guided attention between headline and body sentences. Intuitively, headline-guided attention will give higher attention weight to sentences with high similarity with the headline and low attention weight to sentences with low similarity with the headline.

Given a sentence $s_i \in b$ and the headline h, the contextual similarity between h and s_i is estimated using a parametric similarity score [161, 162] as defined below.

$$x_i = \sigma\left(\mathbf{s}_i^\top \mathbf{W_s h}\right) \tag{5.1}$$

where $\mathbf{W_s}$ is a learnable parameter matrix, and $\mathbf{s}_i$ and $\mathbf{h}$ are the encoding of the sentence s_i and headline h respectively obtained using either with BiLSTM [123] or sentence transformer[4][163]. Given the similarity score x_i between headline h and sentence, s_i attention weight between headline h and sentence s_i is estimated by the following equation 5.2.

$$\mathbf{b_i} = \frac{exp\left(\mathbf{x}_i\right)}{\sum_{i=0}^{n} exp\left(\mathbf{x}_i\right)} \tag{5.2}$$

Now, based on the attention weight b_i, sentence encoding s_i is modified by equation 5.3.

$$\mathbf{s_i} = \mathbf{b_i s_i} \; \forall i \in [1\ ,\ \mathrm{n}] \tag{5.3}$$

Next, If $\mathbf{x_i} \geq \theta$ where θ is a user-defined similarity threshold, then the sentence $\mathbf{s}_i$ is added to $\mathcal{S}^+$, otherwise $\mathcal{S}^-$. Before placing the sentence encoding $\mathbf{s}_i$ in respective sets, i.e. $\mathcal{S}^+$ or $\mathcal{S}^-$, sentence encoding $\mathbf{s}_i$ is multiplied by attention weight $\mathbf{b}_i$ as defined in equation 5.3.

[4] Pretrained S-BERT

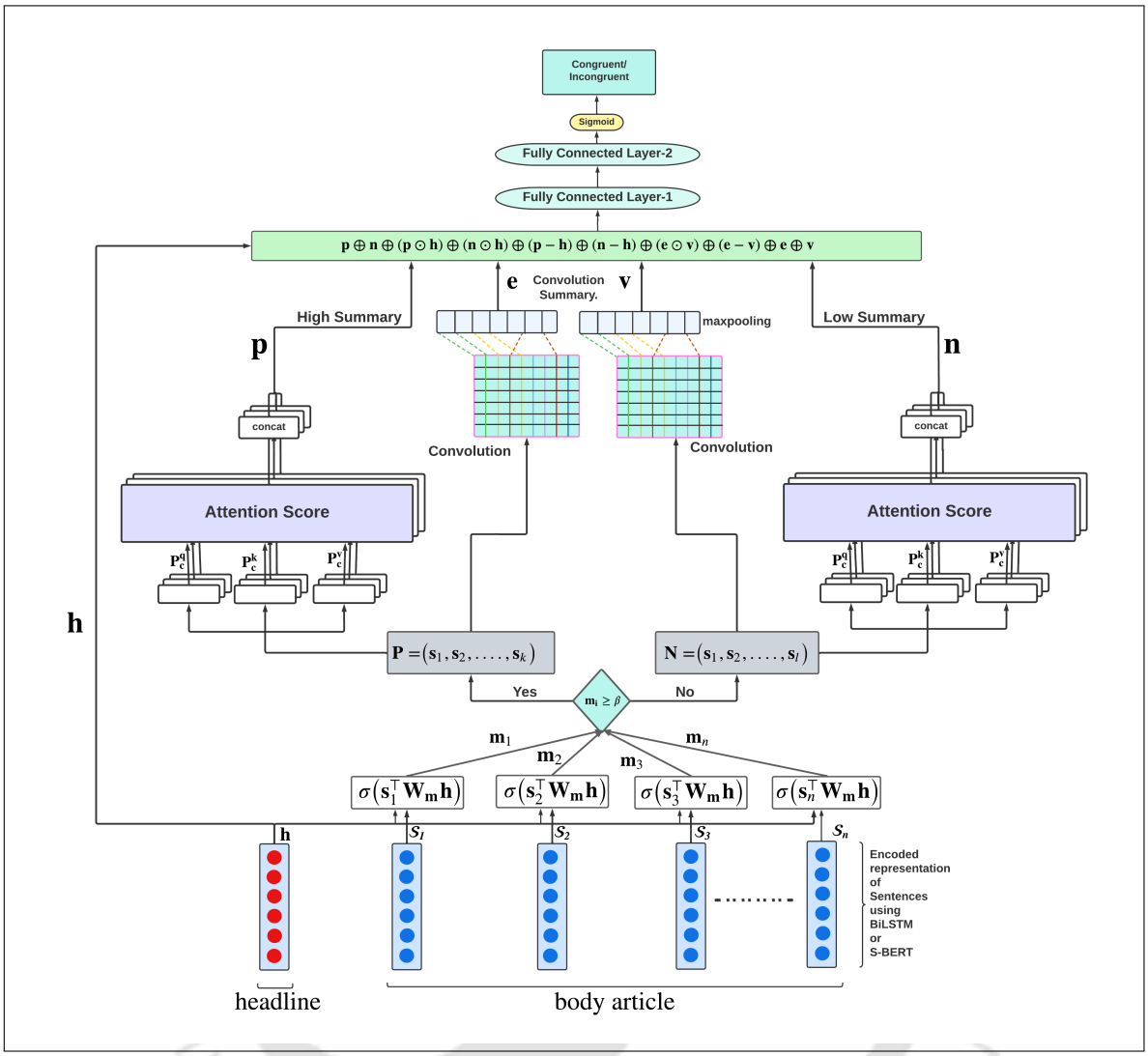


Fig. 5.2 The proposed model *MADS* is represented in the diagram. First, sentence encoding is obtained using BiLSTM or S-BERT. Then, a similarity score m_i between $\mathbf{h}$ and $\mathbf{s}_i$ is estimated. If $m_i \geq \beta$ is true, the sentence is placed in the positive set otherwise, it is placed in the negative set. Then we generate a summary of these positive and negative sets using multi-head attention and convolution. Thereafter, text matching features between headline and representative summary generated from multi-head attention and convolution is obtained and passed to the two fully connected layers for the classification.

5.6.1.2 Dual Summarization

The *dual summarization* means, given an article $a =< h, b >$, b is summarized into highly congruent representation from $\mathcal{S}^+$ and poorly congruent representation from $\mathcal{S}^-$. For a

given set $\mathcal{S}^+$ or $\mathcal{S}^-$, we extract two forms of summarization vectors; (i) *sequential weighted summation*, and (ii) *convolution*.

5.6.1.3 Summary using Sequential Weighted Summation

Given the set of sentences $\mathcal{S} = \{s_1, s_2, ..., s_k\}$, a simple way of summarization can be summation of the encoding of these sentences, i.e., $\sum_{i=0}^{k} \mathbf{s}_i$. However, different sentences may have different influence; thus, one may incorporate the influential weight β_i of the sentences as follows, $\sum_{i=0}^{k} \beta_i \mathbf{s}_i$. Further, in a text document where the position of the sentences may play an influencing role in summarization, the sequential model like LSTM may be further used to summarize a sequence of sentences, i.e., *the hidden state of the last LSTM unit* $\mathbf{h}_k$. In this study, we integrate the influential importance of a sentence along with a sequential summarization. The characteristics of sequence weighted summation summary over highly congruent set $\mathcal{S}^+$ and poorly congruent set $\mathcal{S}^-$ is defined as follows: *(i)* Sequentially encode sentence of set using contextualize LSTM to obtain a global representative vector $\mathbf{h}_k$ of set. *(ii)* If sentences $s_i \in \mathcal{S}^+$ are highly similar to the global representative vector $\mathbf{h}_k$ of the set, $\mathcal{S}^+$ then s_i should be given high weight while generating a summary of the highly congruent set, $\mathcal{S}^+$. *(iii)* Similarly, if the sentence $s_i \in \mathcal{S}^-$ is least similar to the global representative vector $\mathbf{h}_k$ of the set, $\mathcal{S}^-$ then s_i should be given high weight while generating a summary of poorly congruent set $\mathcal{S}^-$. The prime motivation behind the characteristics mentioned above is that if the headline is similar to the summary obtained from the high-influenced (sentence highly similar to $\mathbf{h}_k$ in the highly congruent set) and low-influenced (sentence least similar to $\mathbf{h}_k$ in the poorly congruent set) sentences, then it is congruent. Considering the performance of context-aware LSTM in sentiment analysis task in the study [164], we consider explicit contextualized the LSTM [123] to obtain a global representative of the set by sequentially encoding sentences of the set. Explicit contextualized LSTM learns a mapping between input sentence, weighted summation representation of the sentence in the set and the current hidden state of LSTM recursively at every time stamp. Explicit context of contextualize LSTM is defined by weighted summation of sentences in the set. The weight of each sentence is defined by the similarity between the current hidden state of contextualize LSTM at a time stamp j and sentences of the set. If this weight is high with all sentences in the set, then it indicates that the current hidden state is a good representative of the set, and if the weight is low with all sentences of the set, then it indicates that the current hidden state is not a good representative of the set. So, based on the similarity between the current hidden state and sentences of the set, the context of contextualized LSTM is updated by taking the weighted summation of sentences encoding of the set. Learning additional context enforced that the

last hidden state of contextualized LSTM will be a good representative of sentences of the set. Given a sequence of sentences $\mathcal{S} = \{s_1, s_2, ..., s_k\}$, at a given sequential instance $j, i \leq k$, $\mathbf{h}_j$ represents the current hidden state vector after encoding sentences $s_{1..j}$. Now, we estimate the similarity score $\mathbf{z}_{j,i}$ between i^{th} sentence $\mathbf{s}_i$ and the current hidden state vector $\mathbf{h}_j$ as follows.

$$\mathbf{z}_{j,i} = \sigma\left(\mathbf{s}_i \mathbf{W}_z + \mathbf{h}_j \mathbf{U}_z\right) \tag{5.4}$$

where $\mathbf{W}_z$ and $\mathbf{U}_z$ is a learnable matrix. Next, we convert the similarity score $\mathbf{z}_{j,i}$ into sentence weight $\beta_{\mathbf{j,i}}$ using equation 5.5.

$$\beta_{\mathbf{j,i}} = \frac{exp\left(\mathbf{z}_{j,i}\right)}{\sum_{l=0}^{k} exp\left(\mathbf{z}_{j,l}\right)} \tag{5.5}$$

At the given instance $j, j \leq k$, the context vector $\mathbf{c\acute{}}_j$ of the sentences in $\mathcal{S}^+$ is defined by the weighted summation of sentence $s_i \in \mathcal{S}$, as follows.

$$\mathbf{c\acute{}}_j = \sum_{i=1}^{k} \beta_{j,i} \mathbf{s}_i \tag{5.6}$$

We now modify the current context cell state $\mathbf{c}_j$ of LSTM unit as follows, and given as the context state for the next unit.

$$\mathbf{c}_j = \left(\mathbf{c}_j \odot \mathbf{c\acute{}}_j\right) \tag{5.7}$$

After inputting the last sentence in set $\mathcal{S} = \{s_1, s_2, ..., s_k\}$, the last hidden state vector $\mathbf{h_k}$ of contextualized LSTM is considered as a global representative vector for of set, $\mathcal{S}$. Now, as discussed above, we want to give high weight to a sentence which is highly similar to the global representative vector $\mathbf{h}_k$ while generating a summary for a highly congruent set $\mathcal{S}^+$. Similarly, high weight to a sentence which is least similar to the global representative vector $\mathbf{h}_k$ while generating the summary for a poorly congruent set $\mathcal{S}^-$. Therefore, we estimate similarity and difference of sentences $s_i \in \mathcal{S}$ with $\mathbf{h}_k$ as follows.

$$\mathbf{e\acute{}}_i = \left(\left(\mathbf{h}_k \odot \mathbf{s}_i\right) \oplus \left(\mathbf{h}_k \text{-} \mathbf{s}_i\right)\right) \tag{5.8}$$

Subsequently, we pass the above vector $\acute{\mathbf{e}}_\mathbf{i}$ to a dense layer to obtain a similarity score $\acute{\mathbf{z}}_i$ between the sentence s_i and global representative vector $\mathbf{h}_k$.

$$\acute{\mathbf{z}}_i = \sigma\left(\acute{\mathbf{e}}_\mathbf{i}\mathbf{W}_e+\mathbf{b}\right) \tag{5.9}$$

where $\mathbf{W}_e$ is a learnable parameter. Next, to obtain weight for sentences, we pass the similarity score vector $\acute{\mathbf{z}}$ to Softmax function defined by equation 5.10.

$$\omega_i=\frac{exp\left(\acute{\mathbf{z}}_i\right)}{\sum_{i=1}^{k} exp\left(\acute{\mathbf{z}}_i\right)} \tag{5.10}$$

Now, the final summary representative vector $\mathbf{d}^+$ of highly congruent set $\mathcal{S}^+$ is defined as

$$\mathbf{d}^+=\sum_{\mathbf{i=1}}^{\mathbf{k}} \omega_\mathbf{i}\mathbf{s}_\mathbf{i} \tag{5.11}$$

Similarly, to extract the summary representative vector $\mathbf{d}^-$ of poorly congruent set $\mathcal{S}^-$, the equation 5.9 is replaced by equation 5.12. The idea is that sentences with the least similarity score should be given high weighted while generating sequential weighted summation summary vector of poorly congruent set $\mathcal{S}^-$.

$$\acute{\mathbf{z}}_i = 1-\sigma\left(\acute{\mathbf{e}}_\mathbf{i}\mathbf{W}_\mathbf{e}+\mathbf{b}\right) \tag{5.12}$$

Though the proposed model focuses on generating a weighted summation of sentence encoding as the summary representative vector (summary), the extractive summary of the congruent set and the incongruent set can be obtained from equations 5.10.

5.6.1.4 Convolution over Highly and Poorly Congruent Sets

Further, as mentioned above, we want to capture local patterns and important n-grams substructure between sentences of the highly congruent and poorly congruent set. Hence, we apply convolution over the encoded vectors of the sentences, i.e., over $\mathcal{S}^+$ and $\mathcal{S}^-$. The main motivation behind convolution summary is to extract local patterns and important n-grams substructure from sentence encoding matrix of respective $\mathcal{S}^+$ and $\mathcal{S}^-$ sets. We applied convolution to obtain convolution summary $\mathbf{g}^+$ and $\mathbf{g}^-$ of $\mathcal{S}^+$ and $\mathcal{S}^-$ set respectively. Our convolution setting is similar to the study [165]. To obtain convolution summary, $\mathbf{g}^+$ we concatenate the convolution summary [165] with unigrams, bigrams up to 7*grams* over

sentence encoding matrix of set $\mathcal{S}^+$, where n-grams refer to the number of sentences in a filter at a time. Similarly, we also obtained convolution summary $\mathbf{g}^-$ over sentence encoding matrix of set $\mathcal{S}^-$.

5.6.1.5 Aggregation and Classification

Once we obtain the summarized encoding of $\mathcal{S}^+$ and $\mathcal{S}^-$, we generate the final feature vector considering the following two aspects.

1. **Relation between the headline and the summaries:** To verify the similarity between headline and body, we estimate an angle and difference between summaries and headline encoding as follows:
   (a) Estimate angle m^+ and difference p^+ between headline encoding h and sequence weighted summation summary of highly congruent set d^+ as defined in equation 5.13 and 5.14.
   $$\mathbf{m}^+ = \mathbf{d}^+ \odot \mathbf{h} \tag{5.13}$$
   $$\mathbf{p}^+ = \mathbf{d}^+ - \mathbf{h} \tag{5.14}$$
   (b) Estimate angle m^- and difference p^- between headline encoding h and sequence weighted summation summary of poorly congruent set d^- as defined in equation 5.15 and 5.16.
   $$\mathbf{m}^- = \mathbf{d}^- \odot \mathbf{h} \tag{5.15}$$
   $$\mathbf{p}^- = \mathbf{d}^- - \mathbf{h} \tag{5.16}$$
2. **Relation between the summaries:** We estimate the similarity between the summary of the highly congruent set and the summary of the poorly congruent set to verify the consistency within the new body. with such motivation, we estimate an angle m and difference p between convolution summaries of highly congruent set $\mathcal{S}^+$ and poorly congruent set $\mathcal{S}^-$ as defined in equation 5.17 and 5.18.
   $$\mathbf{m} = \mathbf{g}^+ \odot \mathbf{g}^- \tag{5.17}$$
   $$\mathbf{p} = \mathbf{g}^+ - \mathbf{g}^- \tag{5.18}$$

Where $\odot$ denotes element-wise multiplication between vectors and $-$ denotes element-wise difference between vectors. Now, we define the final feature for the classification as follows. Though many permutations are possible between sequential summarization and convolution summarization, empirically we have noted that the features estimated using

equations 5.13 to 5.18 provide superior performance.

Intuitively, considering the relation between headline and body helps in detecting incongruent news articles of characteristics (i) and (ii) (discussed in section 5.1) because both the characteristics decide incongruity based on the relation between headline and body. Similarly, considering the relation between the summaries of highly and poorly congruent sets helps in detecting partially incongruent news articles (characteristics (iii) discussed in section 5.1). The relationship between headline and summaries is estimated as follows:

$$\mathbf{f}=\left(\mathbf{d}^{+}\oplus\mathbf{d}^{-}\oplus\mathbf{h}\oplus\mathbf{m}^{+}\oplus\mathbf{p}^{+}\oplus\mathbf{m}^{-}\oplus\mathbf{p}^{-}\oplus\mathbf{m}\oplus\mathbf{p}\oplus\mathbf{g}^{+}\oplus\mathbf{g}^{-}\right) \tag{5.19}$$

Where $\oplus$ denotes concatenation of vectors. The feature vector **f** is then passed through a two layer fully connected neural network with a *Softmax* output layer. We apply cross entropy loss to learn the parameters.

5.6.2 Multi-head Attention Dual Summarization MADS:

Given a news article $\mathcal{I} = \left(\mathcal{H}, \mathcal{B}\right)$ with a pair of its headlines $\mathcal{H}$ and its body $\mathcal{B}$, *MADS* divides the sentences in the body $\mathcal{B}$ into positive $\mathcal{P}$ and negative $\mathcal{N}$ sets based on the matching scores between the sentence $\mathcal{S}_i$ and the headline $\mathcal{H}$. The main motivation behind splitting body sentences into positive $\mathcal{P}$ and negative $\mathcal{N}$ sets is that if a news article is partially incongruent, then sentences congruent with the headline will be in the positive set $\mathcal{P}$ and sentences incongruent with a headline will be in the negative set $\mathcal{N}$. Similarly, in the case of a full congruent news article, most of the sentences of the body should be in $\mathcal{P}$ set, and only few sentences will be in $\mathcal{N}$ set. However, if a news article is fully incongruent, then all the sentences in the body should be incongruent with the headline; hence it should be in $\mathcal{N}$ except one or few sentences in $\mathcal{P}$. Next, summary of $\mathcal{P}$ and $\mathcal{N}$ are obtained separately to match with headline for incongruent news article detection.

5.6.2.1 Similarity Between Headline and Body:

This study uses bidirectional LSTM (BiLSTM) to obtain encoded representation **h** and $\mathbf{s}_i$ of headline $\mathcal{H}$ and sentence $\mathcal{S}_i$, respectively. However, considering the effectiveness of sentence embeddings generated by sentence-BERT (S-BERT) [76] in different NLP tasks[5], we have

[5] Why S-BERT

also used S-BERT to encode headlines and sentences, in this study. Like in [162] [84], the similarity score m_i between $\mathbf{h}$ and $\mathbf{s}_i$ is estimated using the following expression 5.20

$$m_i = \sigma\left(\mathbf{s}_i^{\top} \mathbf{W_m} \mathbf{h}\right) \tag{5.20}$$

where $\mathbf{W_m}$ is a learnable parameter matrix, σ is the sigmoid function and $\top$ is a transpose operation over a vector. Given the similarity score m_i between headline h and sentence, s_i attention weight between headline h and sentence s_i is estimated by the following equation 5.21.

$$\mathbf{b_i} = \frac{exp\left(\mathbf{m}_i\right)}{\sum_{i=0}^{n} exp\left(\mathbf{m}_i\right)} \tag{5.21}$$

Now, based on the attention weight b_i, sentence encoding s_i is modified by equation 5.22.

$$\mathbf{s_i} = \mathbf{b_i s_i} \; \forall i \in [1 \; , \mathrm{n}] \tag{5.22}$$

If $m_i \geq \beta$, then the sentence $\mathbf{s}_i$ is added to set $\mathcal{P}$, otherwise it is added to set $\mathcal{N}$.

5.6.2.2 Summarization

Given two sets of sentences, $\mathcal{P}$ and $\mathcal{N}$, we extract two different types of summaries - *multi-head attention-based summary* and *convolution summary* for each set separately.

Summary using Multi-head Attention: The characteristics of dual summary over positive $\mathcal{P}$ and negative $\mathcal{N}$ sets are defined as follows: *(i)* a sentence which is highly similar to other sentences in the set $\mathcal{P}$ should be given high priority while generating a summary of a positive set $\mathcal{P}$. *(ii)* A sentence which is not similar or least similar to other sentences in the set $\mathcal{N}$ should be given high importance while generating a summary of $\mathcal{N}$. The main motivation behind such a dual summary is that if a summary generated by a highly influenced (sentence with high similarity with all other sentences in the set) sentence from a positive set and a summary generated by the least influenced (a sentence which is either not similar or least similar with other sentences in the set) sentence from $\mathcal{N}$ are congruent with the headline, then the news article is congruent, otherwise incongruent. To capture representation of sentences from different aspects, we apply multi-head attention [58]. As shown in Figure 5.2, given a sequence of sentences $(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_k)$, we define a matrix $\mathbf{P}$ (each row representing a

sentence encoding) to obtain the query $\mathbf{P}^q$, key $\mathbf{P}^k$ and value $\mathbf{P}^v$ matrices using the following expression.

$$\mathbf{P}_c^q, \mathbf{P}_c^k, \mathbf{P}_c^v = \mathbf{P} \cdot \mathbf{W}_c^q, \mathbf{P} \cdot \mathbf{W}_c^k, \mathbf{P} \cdot \mathbf{W}_c^v \tag{5.23}$$

where $\mathbf{W}_c^q$, $\mathbf{W}_c^k$ and $\mathbf{W}_c^v$ are learnable parameter matrices of query, key and value projections respectively, for c^{th} attention head of multi-head self attention and $\cdot$ is the dot product between matrix. Subsequently, attention weigh $\mathbf{A}_c$ is defined as follows:

$$\mathbf{M} = \left(\frac{\mathbf{P}_c^q \, (\mathbf{P}_c^k)^\top}{\sqrt{\mathbf{z}}} \right) \tag{5.24}$$

$$\mathbf{A}_{c,i,j} = \left(\frac{\exp(\mathbf{M_{ij}})}{\sum_{k,l} \exp(\mathbf{M_{k,l}})} \right) \tag{5.25}$$

Here M is matching matrix and $\mathbf{A}_c$ is attention weight matrix of c^{th} attention head. $\mathbf{A_c}[\mathbf{i},\mathbf{j}]$ entry represents the similarity probability between i^{th} and j^{th} sentence of set $\mathcal{P}$. $\mathbf{z}$ is the dimension of $\mathbf{P}_c^q$. Next, weighted summation is applied over encoding of sentences $\mathbf{s}_i$ based on similarity with other sentences in the set.

$$\mathbf{u}_{c,i} = \left(\sum_{j=1, i \neq j}^{k} \mathbf{A}_{c,ij} \mathbf{P}_{c,i}^v \right) \tag{5.26}$$

Where $\mathbf{u}_{c,i}$ is the sentence representation obtained after weighted summation between i^{th} sentence of $\mathbf{P}_c^v$ and attention weight $\mathbf{A}_{c,ij}$ between i^{th} sentence with all other sentences j in $\mathbf{P}_c^v$ of attention head c. Similarly, by following equation 5.26, representation of other sentences in a respective set are also obtained to form a sentence representation matrix $\mathbf{U}_c = \{\mathbf{u}_{c,1}, \mathbf{u}_{c,2}, ..., \mathbf{u}_{c,k}\}$ of attention head c. Now we concatenate the sentence representation obtained by different attention heads and pass it to a dense layer to obtain the final sentence representation $\mathbf{U}$.

$$\mathbf{U} = \Big(\mathbf{U}_1 \oplus \mathbf{U}_2 \oplus .. \mathbf{U}_c \oplus . \oplus \mathbf{U}_l \Big) \mathbf{W}_u \tag{5.27}$$

Where $\mathbf{W}_u$ is the trainable parameter matrix and $\mathbf{U}_c$ is c^{th} attention head. $\mathbf{U}$ is a sentence representation matrix obtained by concatenating representation of i^{th} sentence obtained by l attention head. Now we concatenate representations of sentences $\mathbf{u}_i$ in the sentence representation matrix $\mathbf{U}$ and pass to the dense layer to obtain a summary $\mathbf{p}$ of positive set $\mathcal{P}$.

$$\mathbf{p} = \Big(\mathbf{u}_1 \oplus \mathbf{u}_2 \oplus .. \oplus \mathbf{u}_i \oplus . \oplus \mathbf{u}_k \Big) \mathbf{W}_m \tag{5.28}$$

Where $\mathbf{u}_i$ is a row vector of the matrix U and $\mathbf{W}_m$ is the learnable parameter matrix. Similarly, to extract a summary $\mathbf{n}$ of a negative set, $\mathcal{N}$ equation 5.25 is replaced by equation 5.29. The reason behind this is that the sentence with the least similarity score with other sentences in the set $\mathcal{N}$ should be given high importance while generating a summary $\mathbf{n}$ of the set $\mathcal{N}$.

$$\mathbf{A}_{c,i,j} = \left(\frac{\exp(\mathbf{1} - \mathbf{M_{ij}})}{\sum_{k,l}\exp(\mathbf{1} - \mathbf{M_{k,l}})}\right) \tag{5.29}$$

5.6.2.3 Local Patterns Summary

We also extract a summary by extracting meaningful n-grams substructure and local patterns within sentence encoding matrix $\mathbf{P}$ and $\mathbf{N}$ of positive set $\mathcal{P}$ and negative $\mathcal{N}$ sets respectively. To extract summary $\mathbf{e}$ and $\mathbf{v}$ based on the local structure and meaningful n-grams substructure, we employ convolution [165] over positive $\mathcal{P}$ and negative $\mathcal{N}$ sets. Our convolution settings over sentence encoding matrix $\mathbf{P}$ and $\mathbf{N}$ of positive $\mathcal{P}$ and negative $\mathcal{N}$ sets are similar to convolution setting discussed in study [165][6]. We concatenate the summary obtained by unigrams, bigrams, trigrams upto 7-grams convolution operations to generate summary $\mathbf{e}$ and $\mathbf{v}$ of positive $\mathcal{P}$ and negative $\mathcal{N}$ sets respectively.

Subsequently, we further estimate feature vectors to measure similarity and contradiction between headline encoding $\mathbf{h}$ and summary obtained using multi-head attention $\mathbf{p}$, $\mathbf{n}$. The main objective behind estimating similarity and contradiction between headline and summary of the positive and negative set is that if a news article is fully congruent, then the similarity between the headline and summary of positive and negative sets should be high. Similarly, in the case of fully incongruent news article, the similarity of headline encoding $\mathbf{h}$ with both summaries $\mathbf{p}$ and $\mathbf{n}$ should be low. Intuitively, in the case of a partially incongruent news article, the similarity between headline encoding $\mathbf{h}$ and summary $\mathbf{p}$ of the positive set may be high. Still, the similarity between headline encoding $\mathbf{h}$ and summary $\mathbf{n}$ of negative set should be low. With the above-mentioned objectives, we estimated similarity and contradiction between headline and summary of positive and negative set as follows:

$$\mathbf{a}^{+} = \mathbf{p} \odot \mathbf{h} \tag{5.30}$$

[6] Convolutional Neural Networks Implementation GitHub Link

$$\mathbf{a}^{-} = \mathbf{n} \odot \mathbf{h} \tag{5.31}$$

$$\mathbf{b}^{+} = \mathbf{p} - \mathbf{h} \tag{5.32}$$

$$\mathbf{b}^{-} = \mathbf{n} - \mathbf{h} \tag{5.33}$$

$$\acute{\mathbf{f}} = \left(\mathbf{a}^{+} \oplus \mathbf{a}^{-} \oplus \mathbf{b}^{+} \oplus \mathbf{b}^{-} \oplus \mathbf{p} \oplus \mathbf{n}\right) \tag{5.34}$$

Where $\odot$ denotes element-wise multiplication and $\oplus$ denotes concatenation of vectors. $\mathbf{a}^{+}$ and $\mathbf{b}^{+}$ is angle and difference (similarity measure features) between summary of positive set and headline. Similarly, $\mathbf{a}^{-}$ and $\mathbf{b}^{-}$ are similarity features between headline and summary of negative set. Next, we also estimate the similarity between $\mathbf{e}$ and $\mathbf{v}$ convolution summary of positive set $\mathcal{P}$ and negative set, $\mathcal{N}$ respectively. The key motivation behind estimating similarity between $\mathbf{e}$ and $\mathbf{v}$ is that if a news article is congruent, then similarity between the summary of positive set $\mathcal{P}$ and negative set $\mathcal{N}$ should be high because sentences in the body of a congruent news article are related to each other and similar in topics. Whereas in the case of a partially incongruent or fully incongruent article, there must be some sentences in the body content which does not correlate with the headline and other sentences of the body. Hence, in case of incongruent news article, dissimilarity between summary of positive set $\mathcal{P}$ and negative set $\mathcal{N}$ should be high. With such motivation, we estimate similarity between $\mathbf{e}$ and $\mathbf{v}$ convolution summary of positive set $\mathcal{P}$ and negative set $\mathcal{N}$ as follows:

$$\mathbf{c}^{+} = \mathbf{e} \odot \mathbf{v} \tag{5.35}$$

$$\mathbf{c}^{-} = \mathbf{e} - \mathbf{v} \tag{5.36}$$

$$\mathbf{f} = \left(\acute{\mathbf{f}} \oplus \mathbf{c}^{+} \oplus \mathbf{c}^{-} \oplus \mathbf{e} \oplus \mathbf{v}\right) \tag{5.37}$$

Finally, the feature vector $\mathbf{f}$ is passed to a two-layer fully connected neural network followed by softmax for incongruent news article classification.

5.7 Evaluation Methodology

In this section, we describe the experiment settings for accessing the performance of the candidate methods. We also provide reproducibility information for the shown results and analyses.

5.8 Experimental Results and Discussions

This chapter validates three datasets from the literature to assess the proposed models' performance: ISOT for true and fake news, FNC for fully congruent and incongruent articles, and NELA-17 for partially incongruent articles. Each dataset poses unique challenges in incongruent news detection, facilitating a thorough evaluation of the proposed model. Detailed descriptions of these datasets are provided in the Section 3.2 of Chapter 3.

5.8.1 Baseline Models

This study considers several existing state-of-the-art models from the literature to compare the performance of the proposed model. We consider existing state-of-the-art feature-based models `FNC-1` *(Fake News Challenge)*[7] [22], `UCLMR` *(UCL Machine Reading)* [35][8] and `StackLSTM`[9] [98] from literature to compare the response of proposed model against bag-of-words based feature-based models. Similarly, we also consider similarity-based state-of-the-art models `HDSF` *(Hierarchical Discourse level Structure Learning)*[10] [26], `AHDE` *(Attentive Hierarchical Dual Encoder)*[11] [28] and `GHDE` *(Graph-based Hierarchical Dual Encoder)*[12] [1] as baseline models. Since `GHDE` model requires annotation between every pair of headline and paragraph; hence, we reproduce the results of `GHDE` model for only the NELA dataset. Besides feature-based and similarity-based baseline, we consider recently published summarization-based model `GSFD` *(Fake News Detection Using Summarization)* [38, 31] as summarization-based baseline models. In addition to similarity and summarization-based baseline models from literature, we consider five models as the baseline.

`MLP`: A multilayer perceptron classifier is trained over the following features: *(i)* the count of unigram, bigram, trigram overlap between headline and body, *(ii)* TF-IDF similarity between headline and body, *(iv)* SVD of headline and body and cosine similarity between SVD of headline and body, *(v)* the sentiment of headline and body. `BiLSTM`: It finds contradiction and similarity between the headline and body. The headline and the body are encoded separately using BiLSTM [123]. An entailment and similarity feature vector is formed by concatenating the encoding of the headline, the body, and the difference and similarity vectors between them. The concatenated vector is passed to a two-layer multilayer perceptron for incongruent

[7] FNC-1 baseline by organizer code
[8] UCLMR implementation code
[9] stackLSTM based model code repository
[10] HDSF code repository
[11] Attentive Hierarchical Dual Encoder(AHDE) code
[12] GHDE model code repository

Table 5.1 Present details of hyperparameters used to produce results.

Hyperparameters	**Values**
Epoch	40
Threshold Value	0.25, **0.5**, 0.75
K Value	**5**, 7 ,10
Number of Attention Head	1, 2, 8
Batch Size	50
Word Embedding Dimension	200
Learning Rate	0.01
Loss Function	Cross Entropy
Memory Dimension of LSTM	100
Hidden Dimension of LSTM	100
Number of Layer in MLP	2
Hidden State Dimension of Sentence Transformer	384
Filter Size for Sentence Transformer Models	n-gram $\times$ (384)
n-gram values	1 to 7
No. of Attention Head	1, 2, 8

news classifications.

`BERT`: It follows a similar approach of *BiLSTM*, first, encode headline and body separately using pre-trained BERT[13] [80]. Next, an angle and difference feature vector is formed by concatenating the encoding of the headline, the body, and the angle and differenced vectors between them. The concatenated vector is passed to a two-layer multilayer perceptron for incongruent news classifications.

`RoBERT` *(Recurrence over BERT)* [125]: Considering the length of the news article body, which is large, the study proposed the *RoBERT* model [125] for large document classifications. Motivated by study, we consider *RoBERT* [125] models as a baseline in this study. It splits the news article body into sentences and gets BERT encoding of the sentences and headline. It then applies LSTM over the BERT encoding headline and sentences. The last hidden state of LSTM is passed to a two layer fully connected neural network to classify the news article.

For all the experiments reported in this study, the pre-trained Google's word2vec [166] embedding is used for word level embedding. The F-measure (F), classwise F-measure, Accuracy (Acc) have been used as evaluation metrics. Table 5.1 presents the details of hyperparameters used to produce the results. Though we have experimented with different hyperparameters, we have presented only the optimal hyperparameters value in Table 5.1.

[13] Huggingface pre-trained BERT

Table 5.2 Comparison of the performances of different models over three benchmark datasets. Here, (Acc) and (F) indicate accuracy and F-measure, respectively. Similarly, (Cong.) and (Incong.) indicate the F-measure of congruent and incongruent classes, respectively. Similarly, **S.** indicates similarly based methods. Color indicates the best performance across models over a dataset.

		Models	NELA-17				ISOT				FNC			
			Acc	F	Cong.	Incong.	Acc	F	Cong.	Incong.	Acc	F	Cong.	Incong.
Baseline	Features	**FNC-1** [22]	0.586	0.586	0.564	0.608	0.844	0.844	0.847	0.842	0.586	0.496	0.282	0.709
		UCLMR [35]	0.589	0.588	**0.608**	0.569	**0.997**	**0.997**	**0.997**	**0.997**	0.964	0.955	0.934	0.975
		MLP	**0.629**	**0.629**	0.574	0.628	0.985	0.985	0.985	0.985	0.926	0.909	0.939	0.909
		StackLSTM [98]	0.597	0.591	0.541	**0.641**	0.992	0.992	0.992	0.992	**0.971**	**0.963**	**0.946**	**0.982**
	Encoding	**AHDE** [28]	0.606	0.606	**0.614**	0.598	0.913	0.913	0.909	0.909	0.691	0.454	0.094	0.8143
		GHDE [1]	0.55	0.331	0.331	0.332	-	-	-	-	-	-	-	-
		HDSF [26]	0.517	0.494	0.602	0.386	0.720	0.712	0.665	0.759	**0.758**	**0.666**	**0.492**	**0.841**
		BiLSTM	0.555	0.55	0.563	0.547	0.99	0.99	0.99	0.99	0.616	0.504	0.269	0.740
		BERT	0.572	0.563	0.624	0.503	0.894	0.894	0.894	0.891	0.722	0.419	0.21	0.838
		RoBERT	**0.615**	**0.613**	0.54	**0.642**	**0.996**	**0.996**	**0.996**	**0.996**	0.664	0.583	0.4	0.767
	S.	**GSFD** [38, 31]	0.533	0.532	0.550	0.515	0.998	0.998	0.998	0.998	0.878	0.837	0.755	0.918
Proposed	Dual Summary	**MADS(BiLSTM, $\beta = 0.5$, H = 8)** [21]	0.581	0.575	0.527	0.623	**0.999**	**0.999**	**0.999**	**0.999**	**0.971**	**0.963**	**0.947**	**0.98**
		MADS(BiLSTM, $\beta = 0.5$, H = 2) [21]	0.624	0.623	0.637	0.609	0.998	0.998	0.998	0.998	0.966	0.958	0.939	0.977
		MADS(BiLSTM, $\beta = 0.5$, H = 1) [21]	**0.641**	**0.640**	**0.652**	**0.629**	0.998	0.998	0.998	0.998	0.969	0.960	0.942	0.978
		MADS(ST, $\beta = 0.5$, H = 1) [21]	0.63	0.628	0.603	0.654	0.984	0.984	0.984	0.984	**0.971**	**0.963**	**0.947**	**0.98**
		MADS(ST, $\beta = 0.5$, H = 2) [21]	0.625	0.62	0.579	0.662	0.972	0.972	0.972	0.972	0.968	0.959	0.94	0.978
		MADS(ST, $\beta = 0.5$, H = 8) [21]	0.568	0.562	0.514	0.593	0.978	0.977	0.977	0.978	0.962	0.952	0.93	0.974
		DuSum(BiLSTM,θ)	0.668	0.668	0.673	0.664	**0.998**	**0.998**	**0.998**	**0.998**	0.986	0.982	0.974	0.990
		DuSum(ST,θ)	**0.701**	**0.701**	**0.703**	**0.7**	0.99	0.99	0.99	0.99	0.99	0.987	0.981	0.993
		DuSum(BiLSTM,k)	0.627	0.625	0.62	0.63	0.997	0.997	0.997	0.997	0.985	0.981	0.973	0.990
		DuSum(ST,k)	0.647	0.643	0.66	0.626	**0.998**	**0.998**	**0.998**	**0.998**	**0.991**	**0.988**	**0.983**	**0.993**

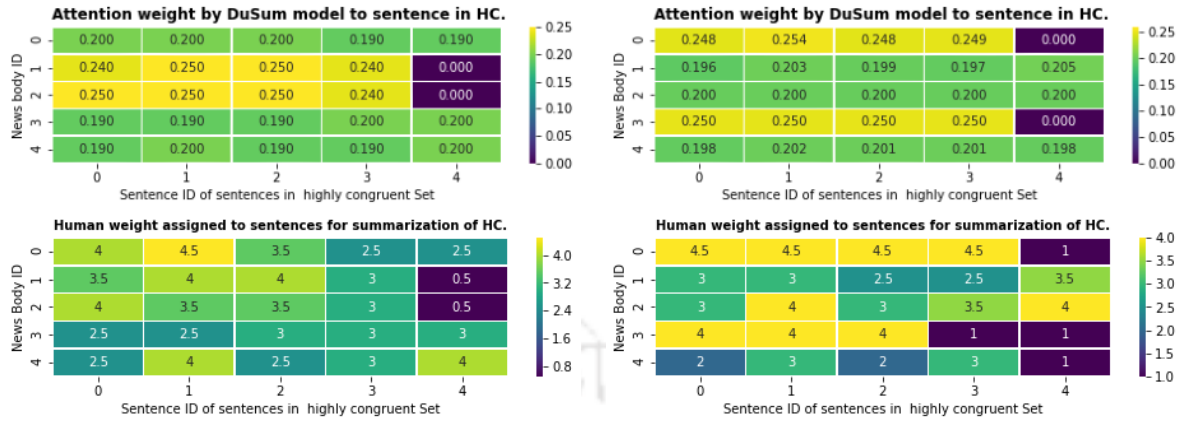
5.9 Results

5.9.1 Results and Discussions

Table 5.2 presents the performance of different models over three benchmark datasets, namely NELA, ISOT and FNC. The seventeen baseline models are further grouped into *features, encoding* and *dual summary* depending on whether a model uses explicit features, neural encoded vectors or summarization. Among the baseline models (explicit features and encoding), *MLP* shows superior performance over all other models for the NELA dataset, and *stackLSTM* outperforms all other baseline models over the FNC dataset. It may be because *MLP* and *stackLSTM* consider bag-of-words based features of different nature such as n-grams, latent topics, sentiments, etc. Interestingly, many of the feature-based models outperform many of the state-of-the-art encoding-based models. It indicates that carefully chosen features may still dominate complex encoding models. However, extracting appropriate features is a challenging task. While inspecting the performance of the baseline models across the three datasets, the following interesting observations were made.

- All the models (both explicit features and encoding) provide comparable performance over ISOT dataset. It is because the samples in the ISOT dataset are small, and class labels are balanced. Similar performance of models over the ISOT dataset has been observed in studies [167, 168].
- All the models provide relatively poor performance over NELA dataset, as compared to FNC and ISOT datasets. It is because samples in NELA dataset represent the characteristics of a partially incongruent news article, but ISOT and FNC samples are fully congruent and incongruent. As discussed in section 3.2.1, in the case of a fully incongruent news article, all sentences and paragraphs of the body are unrelated to the headline. Hence, a classifier needs to estimate the similarity between the headline and body to detect fully incongruent or congruent news articles. In the case of a partially incongruent news article, there are a few sentences in the body related to the headline and a few sentences in the body unrelated to the headline. So, a part of content in the body is contradictory to each other in terms of relation to the corresponding headline. Hence, detecting partially incongruent news articles is more challenging than detecting fully incongruent news articles.
- The encoded models perform relatively poorer than the explicit feature-based models. It indicates that neural encoding methods which are influenced by dominant characteristics may fail to capture minority incongruity indicators.
- The GSFD, a summarization-based method, performs superior among the neural encoding methods over the FNC dataset (dataset of full congruent or incongruent characteristics) compared to that of the NELA dataset (dataset of partially incongruent or fully congruent characteristics). This is because, in the case of FNC dataset, though the headline is incongruent with the body, sentences within the body correlate with each other. Hence, a summary of the body is the true representative of the body, so comparing the headline with a summary of the body enhances the performance over the FNC dataset. In contrast, a partially incongruent news article, sentences in the body do not correlate with each other. This implies that either congruent or incongruent parts will dominate the summary. So summary may not be a true representative of the body always. Hence, summarization-based model *GSFD* performance is inferior compared to other neural encoding-based models over the NELA dataset.

The above observations indicate that the global encoding of the entire body captures a general representation of the document and likely fails to capture the specific incongruity indicators. In contrast, the summarization-based model is good for fully incongruent news article detection but ineffective for partially incongruent news article detection. Our proposed model *DuSum* addresses the problem by generating a dual summary. The proposed model **DuSum** has been implemented using two setups; (i) generating sentence encoding using BiLSTM [123], or sentence transformer (ST) [163], and (ii) generating , and $\mathcal{S}^-$ using threshold θ, or considering most congruent and incongruent top $k > 0$ $\mathcal{S}^+$, and $\mathcal{S}^-$. We name



(*a*) NELA Dataset (*b*) FNC Dataset

Fig. 5.3 Present comparison between heatmap of human assigned weights and heatmap of *DuSum* model weight ω (estimated in equation 5.10) for summarization of sentences in the highly congruent set. Figure 5.3 (a) presents the heatmap comparison between the human assigned weights and heatmap of $DuSum(ST,\theta)$ model weight ω to sentences in the highly congruent set of the NELA dataset. Figure 5.3 (b) presents the heatmap comparison between the human assigned weights and heatmap of $DuSum(ST,\theta)$ model weight ω to sentences in the highly congruent sentences of the FNC dataset. Here, **HC** indicates the highly congruent set.

them as **DuSum**$(BiLSTM,\theta)$, **DuSum**$(BiLSTM,k)$, **DuSum**(ST,θ), and **DuSum**(ST,k) in the Table 5.2. From Table 5.2, it is evident that all variants of the proposed model outperform the majority of the baseline methods across all datasets. To be specific, $DuSum(ST,\theta)$ outperforms all the baseline models and all other proposed model setups over the NELA dataset. $DuSum(BiLSTM,\theta)$ and $DuSum(ST,k)$ jointly outperform other proposed model setup over ISOT dataset. Similarly, $DuSum(ST,k)$ outperforms other proposed model setups over the FNC dataset. Considering the performance of **DuSum** models compared to summarization-based baseline model *GSFD* [38, 31]. It is established that obtaining a summary of the highly congruent and poorly congruent set separately is more effective than obtaining a summary of the entire body for the incongruent news article detection task. On further comparing the performance of proposed model variations, we can claim the sentence transformer *ST* is more suitable as an encoder than *BiLSTM*. One of the possible reasons for getting improved performance with sentence transformer is due to the usage of pre-trained sentence transformer embedding, which in turn reduces the number of trainable parameters. Comparing the performance of $DuSum(ST,k)$ and $DuSum(ST,\theta)$ from Table 5.2 we observe that $DuSum(ST,k)$ outperform the $DuSum(ST,\theta)$ with a significant margin over FNC dataset. The possible reason behind this could be that though the news article is incongruent or congruent in the FNC dataset, the sentences in the news body are coherent with each other, hence

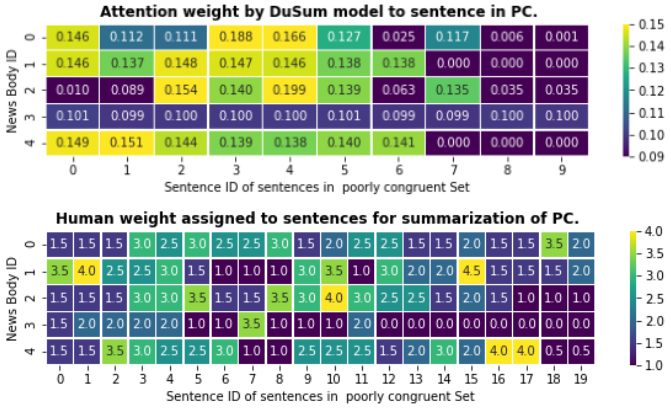


Fig. 5.4 Present comparison between heatmap of human assigned weights and heatmap of $DuSum(ST,\theta)$ model weights ω (estimated in equation 5.10) for summarization of sentences in the poorly congruent set of the FNC dataset. Here, **PC** indicates poorly congruent set.

considering the summary of top k and least k similar sentences represents the summary of the highly congruent and poorly congruent set. From the above observation, it is apparent that considering top k and least k similar sentences with the headline as $\mathcal{S}^+$ and $\mathcal{S}^-$ is sufficient for incongruent news article detection in case sentences of the news article body is coherent with each other in news body. However, the performance of $DuSum(ST,\theta)$ is superior over $DuSum(ST,k)$ over the NELA dataset. We investigate the degradation in performance of $DuSum(ST,k)$ over the NELA dataset and observe that the number of sentences in NELA is much higher than the number of sentences in the FNC dataset and sentences in the news body of the NELA dataset are not coherent. So, the summary of the least k and top k sentences does not represent the summary of the highly and poorly congruent set, and since the sentences in the news body are not coherent with each other in the NELA dataset number of incongruent sentences may be higher than the least k sentence. On another side, the $DuSum(ST,\theta)$ model processes the entire news body by generating a summary of highly and poorly congruent sets; hence, the performance $DuSum(ST,\theta)$ of the model is high over the NELA dataset. Further, using top k most congruent and most incongruent sentences for summarization are computationally efficient and also performs relatively well because only least and top sentences need to be summarized instead of summarizing full news article body in case of splitting based on threshold value θ. Further, comparing the performance proposed model *DuSum* and *MADS* [21] from Table 5.2, it is evident that different setups of *DuSum* outperform all setups of *MADS* [21] over FNC and NELA datasets and comparable performance over ISOT dataset. Such observation establishes that sequence-weighted summation-based summarization (defined in subsection 5.6.1.3) is more effective than multi-head attention-

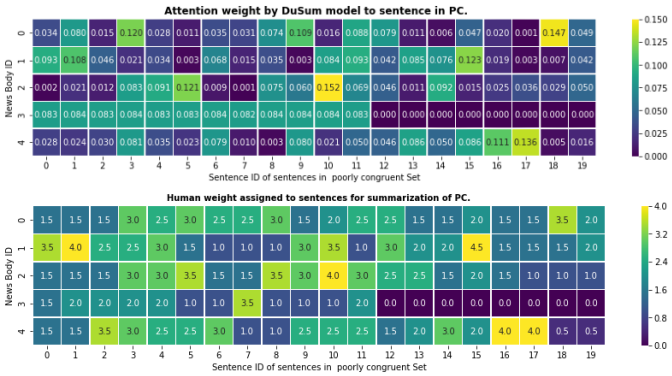


Fig. 5.5 Present comparison between heatmap of human assigned weights and heatmap of $DuSum(ST, \theta)$ model weights ω (estimated in equation 5.10) for summarization of sentences in the poorly congruent set of the NELA dataset. Here, **PC** indicates poorly congruent set.

based summarization for incongruent news article detection. A possible reason behind such performance of *DuSUM* over *MADS* could be that considering the sequence of sentences in the highly and poorly congruent set is also important for incongruent news article detection. But *MADS* only considers multiple representations of the highly and poorly congruent set using multi-head attention, and ignores the sequential information between sentences of the highly and poorly congruent set. Also, *DuSum* considers headline-guided attention between the headline and sentence of the news body, which helps the *DuSum* model to highlight the sentence in the news article body based on its contextual similarity with the headline. Which further boosts the performance of the *DuSum* model. Next, comparing the performance of BiLSTM encoding and BERT encoding-based models following observations can be made.

- *RoBERT* model, which use BERT to encode the headline and sentences of the body, outperformed all other encodings-based baseline models over the NELA dataset. Similarly, the *BERT* and *RoBERT* models outperform *BiLSTM* models over the FNC dataset.
- Comparing the performance of **DuSum**$(BiLSTM, -)$ and **DuSum**$(ST, -)$ setup of **DuSum** from Table 5.2 it is evident that $DuSum(ST, -)$ setup outperforms $DuSum(BiLSTM, -)$ setup over NELA and FNC dataset, but similar performance is observed over ISOT dataset.

From the above observations, it is apparent that the performance of `BERT` is superior over `BiLSTM` models for encoding-based baseline and the performance of **DuSum**$(ST, -)$ setup is superior over **DuSum**$(BiLSTM, -)$ setup of *DuSum*. Hence, it can be concluded that encoding the sentence and headline with *ST* or *BERT* is better than encoding the headline

Table 5.3 Comparison of the performance between dual summary based proposed model *DuSum*$(ST,-)$, summary of only poorly congruent set **PcSum**(**ST**,θ) and summary of only least k sentences **LkSum**(**ST**,**k**).

	NELA		ISOT		FNC	
Model	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**
DuSum(**ST**,θ)	**0.701**	**0.701**	**0.99**	**0.99**	**0.99**	**0.987**
PcSum(**ST**,θ)	0.663	0.662	0.951	0.951	0.808	0.704
DuSum(**ST**,**k**)	**0.647**	**0.643**	**0.998**	**0.998**	**0.991**	**0.988**
LkSum(**ST**,**k**)	0.618	0.615	0.966	0.966	0.956	0.944

and sentences using BiLSTM. The possible reason behind this is that BERT or ST captures long-term dependency between words and reduces the number of trainable parameters in case pre-trained BERT is considered for encoding sentences.

5.9.2 Effect of Poorly Congruent Set and Least k Sentence Summary

DuSum$(-,-)$ detect incongruent news articles by comparing headlines with summaries of the highly congruent and poorly congruent set. Because sentences in poorly congruent set hold the least similarity with the headline, intuitively, it is interesting to estimate the similarity between the headline and summaries of only poorly congruent set and ignore the summaries of the highly congruent set. Table 5.3 present performance comparison between **DuSum**(**ST**,θ), **PcSum**(**ST**,θ) (Partially congruent set summaries only), **DuSum**(**ST**,**k**) and **LkSum**(**ST**,**k**) (Least k sentence summaries only). Comparing the performance of models from Table 5.3, it is established that matching headlines with summaries of both the highly congruent set and the poorly congruent set is more effective than matching headlines with summaries of only poorly congruent set or summaries of only least k sentences. We further compare the performance of **PcSum**(**ST**,θ) and **LkSum**(**ST**,**k**) from Table 5.3 with feature and neural encoding-based baseline models from Table 5.2. The above comparison establishes that **PcSum**(**ST**,θ) outperform all bag-of-words feature-based, encoding and dual summarization-based state-of-the-art baseline models from literature over the NELA dataset. Similarly, by comparing the performance of **LkSum**(**ST**,**k**) with feature and neural encoding baseline models in Table 5.2 it can be concluded that **LkSum**(**ST**,**k**) outperform neural encoding and feature baseline models except for *MLP*. Similarly, by comparing the performance of **LkSum**(**ST**,**k**) with feature and neural encoding baseline models from Table 5.2 over the FNC dataset, it is evident that **LkSum**(**ST**,**k**) outperform $FNC-1$ [22], *MLP*,

Table 5.4 Comparison of the performances between Multi-head Attention Dual summarization *MADS* and Multi-headed Attention and convolution-based Negative set Summarization *MANS*. Results are obtained using attention head $H = 1$ for NELA dataset and $H = 8$ for FNC and ISOT datasets.

	NELA		FNC		ISOT	
Model	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**
MADS(BiLSTM , β = 0.5)	**0.641**	**0.640**	**0.970**	**0.963**	**0.999**	**0.999**
MANS(BiLSTM , β = 0.5)	0.619	0.618	0.927	0.907	0.997	0.997

AHDE [28], *HDSF* [26], *GSFD* [38] [31], *BiLSTM*, *BERT* and *RoBERT* models. From such observations, it can be concluded that matching a headline with summaries of the only poorly congruent set is more effective than matching a headline with the encoded body representation or summary of the full news body for incongruent news article detection. However, **DuSum** is still superior for incongruent news article detection tasks.

5.9.3 Dual Summary Versus Summary of Negative Set

MADS estimates similarity between the headline and a summary of positive and negative set. Considering the essential characteristics of the negative set as discussed in section 5.6.2, It is intuitive to ignore the positive set summary and match the headline with the summary of the only negative set for incongruent news article detection. Table 5.4 presents performance comparison between $MADS(BiLSTM, \beta = 0.5)$ and $MANS(BiLSTM, \beta = 0.5)$. *MANS* (Multi-headed Attention and convolution-based Negative set Summarization) discard the positive set and consider only the negative set for summarization, all other settings are similar to $MADS(BiLSTM, \beta = 0.5)$. From table 5.4 it is evident that $MADS(BiLSTM, \beta = 0.5)$ outperform $MANS(BiLSTM, \beta = 0.5)$. Consequently, it establishes that matching a headline with a summary of a positive and the negative set together is more effective. We further compare $MANS(BiLSTM, \beta = 0.5)$ from Table 5.4 and baseline models from Table 5.2. It is evident that $MANS(BiLSTM, \beta = 0.5)$ outperform both *Feature* and *Encoding* baseline models over NELA dataset. Similarly, $MANS(BiLSTM, \beta = 0.5)$ outperform baseline models *FNC* [22], *AHDE* [28], *HDSF* [36], *FEDS* [38] [31], *BiLSTM*, *BERT* and *RoBERT* over FNC dataset. From such observations, it is apparent that dual summarization is more effective than considering individual summary of the negative set for the underlying task. But matching a headline with a summary of the only negative set is more effective than

Table 5.5 Comparison of performance between proposed model $DuSum(-,-)$, $DuSum(-,-)^*$ (*DuSum* with only sequential weighted summation summary component (discussed in 5.6.1.3) and without convolution summary component) and $CDuSum(-,-)$ (*DuSum* with only convolution summary and without sequential weighted summation summary component (discussed in 5.6.1.3)) over NELA and FNC datasets. Similarly, comparison of the performances between $MADS(BiLSTM, \beta = 0.5)$ and CDS: Convolution Dual Summary. Here $*$ in $MADS(BiLSTM, \beta = 0.5)$ indicate that $MADS(BiLSTM, \beta = 0.5)$ without convolution summary component and $CDS(BiLSTM, \beta = 0.5)$ is similar to $MADS(BiLSTM, \beta = 0.5)$ without multi-head attention summary component. Results are obtained using attention head $H = 1$ for NELA dataset and $H = 8$ for FNC datasets. Here $*$ indicates the experimental result of the model without convolution summary.

	NELA		FNC	
Model	**Acc**	**F**	**Acc**	**F**
DuSum(**BiLSTM**,θ)	**0.668**	**0.668**	**0.986**	**0.982**
DuSum(**BiLSTM**,θ)*	0.61	0.609	0.985	0.981
CDuSum(**BiLSTM**,θ)	0.637	0.637	0.965	0.956
DuSum(**ST**,θ)	0.701	0.701	**0.99**	**0.987**
DuSum(**ST**,θ)*	**0.703**	**0.703**	0.982	0.976
CDuSum(**ST**,θ)	0.688	0.688	0.981	0.975
DuSum(**ST**,k)	0.641	0.64	**0.991**	**0.988**
DuSum(**ST**,k)*	0.634	0.633	0.987	0.9843
CDuSum(**ST**,k)	**0.637**	**0.637**	0.983	0.979
MADS(**BiLSTM** , β **= 0.5**)	**0.641**	**0.64**	**0.971**	**0.963**
MADS(**BiLSTM** , β **= 0.5**)*	0.629	0.605	0.958	0.947
CDS(**BiLSTM** , β **= 0.5**)	**0.637**	**0.637**	**0.965**	**0.956**

summarization-based baseline *FEDS* [38] [31] and other state-of-the-art similarity-based baseline models for incongruent news article detection.

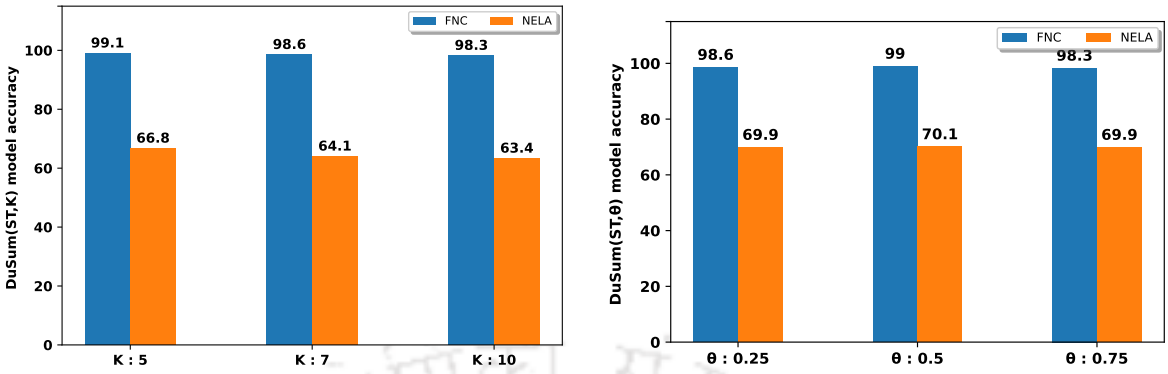
5.9.4 Effect of Local Summary (Convolution)

To understand the importance of considering two different forms of summary, sequence weighted summation (discussed in subsection 5.6.1.3) and convolution summary (discussed in subsection 5.6.1.4) in the proposed model **DuSum**. We conduct an empirical study over the

proposed model setup $DuSum(-,-)$ and proposed model setup with sequence weighted summation summary (discussed in subsection 5.6.1.3) only $DuSum(-,-)^*$, and proposed model setup with only convolution summary (discussed in subsection5.6.1.4) $CDuSum(-,-)$. Table 5.5 presents the performance comparison of the proposed model setup and proposed models with only sequence-weighted summation and only convolution summary. From Table 5.5, it can be clearly observed that $DuSum(BiLSTM,\theta)$ and $DuSum(ST,k)$ outperform their counterparts over NELA and FNC datasets. However, the performance of $DuSum(ST,\theta)^*$ without convolution summary is superior to $DuSum(ST,\theta)$ over the NELA dataset. In contrast, again, the performance of $DuSum(ST,\theta)^*$ is less than the performance of $DuSum(ST,\theta)$ over the FNC dataset. Hence, it can be concluded that considering a convolution summary along with a sequential weighted summation summary improves the performance of the proposed model. Similarly, we also study the importance of different summarization components of *MADS*, we compare the performance of $MADS(BiLSTM,\beta=0.5)$ with *MADS* without convolution summary component $MADS(BiLSTM,\beta=0.5)^*$ and *CDS* (Convolution Dual Summary) differ from $MADS(BiLSTM,\beta=0.5)$ in considering convolution summary only. From table 5.5 it is apparent that *MADS* outperform *MADS* without convolution summary component $MADS(BiLSTM,\beta=0.5)^*$ and $CDS(BiLSTM,\beta=0.5)$. Similarly, the superiority of convolution-based summary over multi-head attention-based summary is apparent on comparing the performance of $MADS(BiLSTM,\beta=0.5)^*$ and $CDS(BiLSTM,\beta=0.5)$ in table 5.5

5.9.5 Selection of θ and k Parameters

The threshold values β and θ are used to split the sentences into positive and negative sets for *MADS* and *DuSum*, respectively. This study considers three different threshold values of β 0.25, 0.5 and 0.75 to produce the results of $MADS(-,\beta,H)$ Similarly, considers three different threshold values of θ 0.25, 0.5 and 0.75 to produce the results of $DuSum(-,\theta)$. From Figure 5.7 (a) it is apparent that the proposed model $MADS(BiLSTM,\beta,H)$ performs better on threshold value $\beta=0.5$ across datasets. Similarly, Figure 5.7 (b)presents the result of $MADS(S-BERT,\beta,H)$ for a different value of β. From Figure 5.7 (b) it is evident that $MADS(S-BERT,\beta,H)$ performance is superior on $\beta=0.5$. Hence, $\beta=0.5$ could be considered as the optimal threshold value for both models $MADS(BiLSTM,\beta,H)$ and $MADS(S-BERT,\beta,H)$. Similarly, from Figure 5.6 (a) presents the performance of the proposed model with three different values k i.e., 5, 7, and 10 using $DuSum(ST,k)$ over the NELA and FNC datasets. it can be clearly observed that the performance of $DuSum(ST,k)$ model is superior with $k=5$ as compared to $k=7$ and 10. Therefore, all the experiments of



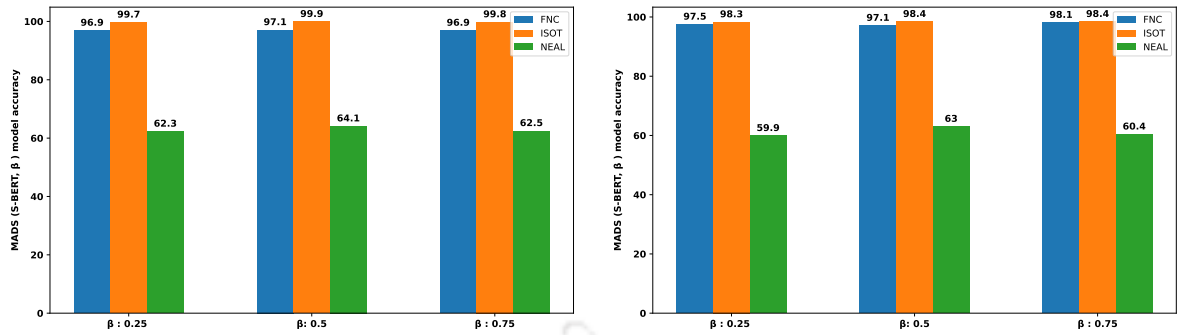
(a) $MADS\Big(\mathbf{BiLSTM}\ ,\ \beta\ ,\ \mathbf{H}\Big)$ (b) $DuSum(ST,\theta)$

Fig. 5.6 Performance of $DuSum(ST,k)$ with different values of k over NELA and FNC dataset, and performance of $DuSum(ST,\theta)$ with different value of θ with the model over NELA and FNC dataset.

DuSum$(-,\mathbf{k})$ in Table 5.2, 5.4, 5.5 and 5.6 used the value of $k=5$. Further, Figure 5.6 (b) shows the performance of $DuSum(ST,\theta)$ over different values of, θ i.e., $0.25, 0.5$ and 0.75. Considering the comparable performance of $DuSum(ST,\theta)$ over different values of θ, we have considered $\theta = 0.5$ for all the experiments of the $DuSum(-,\theta)$, reported in Table 5.2, 5.4, 5.5 and 5.6.

5.9.6 Contextualized LSTM Versus BiLSTM

To investigate the usefulness of contextualized LSTM (discussed in subsection 5.6.1.3) in performance of the proposed model **DuSum**, we build another variant of the proposed model **DuSum** by replacing equations 5.4, 5.5, 5.6 and 5.7 with a plain BiLSTM [123] to obtain a global representative $\mathbf{h}_k$ of sentences in a given set. The BiLSTM in **DuSum**(**-**,**-**,**BiLSTM**) indicates that plain BiLSTM has been used to obtain global representative $\mathbf{h}_k$ of sentences in set $\mathcal{S}$. Table 5.6 presents the performance comparison between $DuSum(-,-)$ and $DuSum(-,-,BiLSTM)$. From Table 5.6 it is evident that the performance of $DuSum(-,-)$ is superior in comparison with the performance of $DuSum(-,-,BiLSTM)$ overall datasets. This observation further validates our intuition (discussed in subsection 5.6.1.3) behind considering contextualized LSTM over plain BiLSTM.



$(a)\ MADS\Big(\mathbf{BiLSTM}\ ,\beta\ ,\mathbf{H}\Big)$ $(b)\ MADS\Big(\mathbf{S-BERT}\ ,\beta\ ,\mathbf{H}\Big)$

Fig. 5.7 Performance of $MADS\Big(\mathbf{BiLSTM}\ ,\beta\ ,\mathbf{H}\Big)$ with different values of β over NELA, FNC and ISOT dataset and performance of $MADS\Big(\mathbf{S-BERT}\ ,\beta\ ,\mathbf{H}\Big)$ with different value of β with the model over NELA, FNC and ISOT dataset.

5.9.7 Split of Body Sentence into Highly Congruent and Poorly Congruent Sets

The performance of **DuSum** depends on the quality of highly congruent $\mathcal{S}^+$ and poorly congruent $\mathcal{S}^-$ sets. Further, the quality of highly congruent $\mathcal{S}^+$ and poorly congruent $\mathcal{S}^-$ sets depends on the effective split of the body sentence into $\mathcal{S}^-$ and $\mathcal{S}^+$. Figure 5.8 presents a sentence heatmap based on the similarity score between headline and sentences of an incongruent news article from the FNC dataset. Sentences in the dark belong to a highly congruent set, and sentence in light colored belongs to poorly congruent sentences. As seen in the figure, only one sentence is colored dark indicating high congruity, and all other sentences are colored light indicating poorly congruity. This indicates that our split criteria based on the similarity between the headline and sentence (equation 5.20) is effective. To further understand the distribution of sentences present in highly and poorly congruent sets in terms of the number of sentences, we investigate the number of sentences present in both highly and poorly congruent sets for the test samples in both the NELA and FNC datasets. Figure 5.9 (a) presents the distribution of sentences presents in the highly and poorly congruent sets of test samples in FNC dataset. Similarly, Figure 5.9 (b) presents the distribution of sentences present in highly and poorly congruent sets in NELA dataset. The following conclusion can be drawn from Figure 5.9: *(i)* The number of sentences in the poorly congruent set is higher than the number of sentences in the highly congruent set for both *True* and *Fake* classes across datasets. This clearly indicates that the *DuSum* model effectively places the least contextually similar sentence from the news body into the poorly congruent set and the high

Table 5.6 Comparison of the performances between **DuSum** with contextualized LSTM and **DuSum** with BiLSTM over NELA, FNC and ISOT datasets.

Model	NELA		ISOT		FNC	
	Acc	**F**	**Acc**	**F**	**Acc**	**F**
DuSum(**BiLSTM**,θ)	**0.668**	**0.688**	0.998	0.998	**0.986**	**0.982**
DuSum(**BiLSTM**,θ,**BiLSTM**)	0.647	0.646	0.998	0.998	0.984	0.98
DuSum(**ST**,θ)	**0.701**	**0.701**	**0.99**	**0.99**	**0.99**	**0.987**
DuSum(**ST**,θ,**BiLSTM**)	0.689	0.689	0.986	0.986	0.981	0.975
DuSum(**BiLSTM**,**k**)	**0.627**	**0.625**	0.997	0.997	0.985	0.981
DuSum(**BiLSTM**,**k**,**BiLSTM**)	0. 604	0.601	**0.998**	**0.998**	0.981	0.976
DuSum(**ST**,**k**)	**0.647**	**0.643**	**0.998**	**0.998**	**0.991**	**0.988**
DuSum(**ST**,**k**,**BiLSTM**)	0.629	0.627	0.976	0.975	0.985	0.981

contextual sentence from the news body into the highly congruent set. *(ii)* The distribution of sentences in the highly and poorly congruent sets for the NELA dataset is different from the distribution of sentences in the highly and poorly congruent sets of the FNC datasets. Considering the fact that the NELA dataset represents the characteristics of partially. It can be concluded that the distribution of sentences in highly congruent and poorly congruent sets also depends on the nature of the dataset, and the proposed model *DuSum* efficiently splits the sentences into highly and poorly congruent sets irrespective of the nature of the dataset. Such observations from Figure 5.9 clearly validate our intuition behind splitting the into poorly and highly congruent sets (discussed in Section 5.6)).

Table 5.7 Presents the Pearson correlation score between the weight assigned by the human annotator (HW) and the weight assigned by the *Dusum* model (MW) for the summarization of sentences in the highly congruent (HC) and poorly congruent (PC) sets of NELA and FNC dataset.

Samples	NELA.		FNC.	
	Highly Congruent	**Poorly Congruent**	**Highly Congruent**	**Poorly Congruent**
	HW. Vs MW.	**HW. Vs MW.**	**HW. Vs MW.**	**HW. Vs MW.**
0	0.918	0.514	0.999	0.813
1	0.969	0.512	0.674	0.745
2	0.974	0.733	0.624	0.117
3	0.666	0.823	0.611	0.755
4	0.963	0.717	0.715	0.958

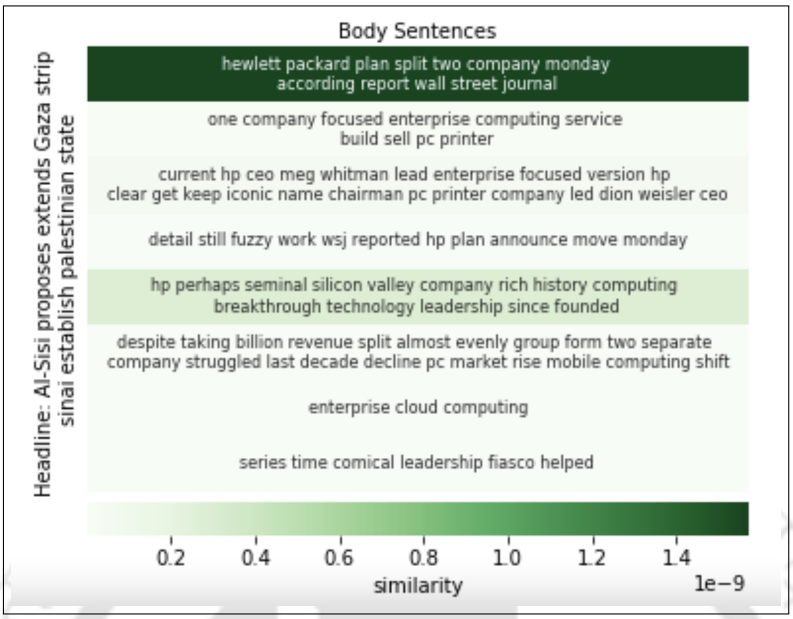
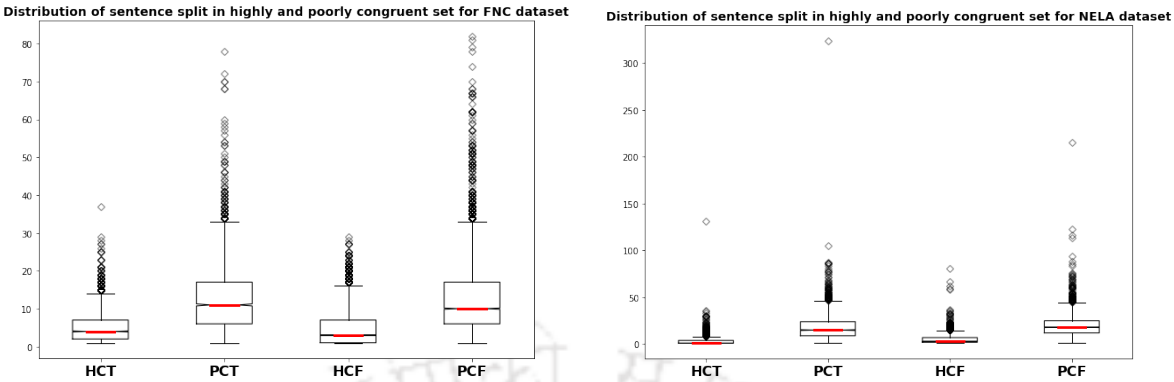


Fig. 5.8 Heatmap represents the relation of the headline with different sentences of the news article. News article body considered in this heatmap belongs to an incongruent class. The Heatmap is based on similarity score between headlines and sentences. It can be visualized that most of the sentence is light colored, which indicates it belongs to the poorly congruent set. And the sentence in the dark belongs to the highly congruent set.

5.9.8 Quality assessment of Sequential Weighted Summary

Though the underlying task in this study is incongruent news article detection, we also assess the quality of summary produced by sequence weighted summation summary (discussed in subsection 5.6.1.3). We randomly selected five news articles from both the NELA and FNC datasets to assess the quality of the sequence-weighted summation summary. Next, for selected five samples from the NELA and FNC datasets, we manually assign weight between a scale of 0 to 5 to sentences in highly congruent and poorly congruent sets based on the importance of the sentence in the summarization of the respective set. Table 5.7 presents the Pearson correlation score between human assign weight and model weight ω. From Table 5.7, it is evident that the Pearson correlation score between human assigned weight and model weight ω (estimated in equation 5.10) is high. From the above observations, it is apparent that our proposed summarization method (sequential weighted summation summarization) is effective in generating summaries of highly and poorly congruent sets for the incongruent news article detection task. Figure 5.3 present the heatmap comparison between heatmap based on human assign weight and $DuSum(ST, \theta)$ model weight ω for NELA (Figure 5.3



(*a*) Split distribution FNC dataset (*b*) Split distribution NELA dataset

Fig. 5.9 Present the distribution of sentences in the highly congruent and poorly congruent sets after the split of the news body into highly and poorly congruent sets. Figure 5.9 (a) presents the distribution of sentences in highly and poorly congruent sets over the test set of the FNC dataset. Similarly, Figure 5.9 (b) presents the distribution of sentences in highly and poorly congruent sets over a test set of the NELA dataset. Here *(i)* **HCT** indicates: highly congruent set of the True class, *(ii)* **PCT** indicates: poorly Congruent set of the True class, *(iii)* **HCF** indicates: highly congruent set of the Fake class and *(iv)* **PCF** indicates: poorly congruent set of the Fake class.

(a)) and FNC (Figure 5.3 (b)) datasets. From Figure 5.3, it is apparent that the heatmap generated on the ω value assigned to sentences by $DuSum(ST, \theta)$ the model is almost similar to the heatmap generated on the human weight assigned to sentences. Similarly, comparing the heatmap presented in Figure 5.4 and 5.5, it is evident that the heatmap generated on the ω value assigned to the sentence by $DuSum(ST, \theta)$ the model is almost similar to the heatmap generated on the human weight assigned to sentences. Such observation indicates that our proposed summarization methods sequence weighted summation summary (discussed in subsection 5.6.1.3) for summarization of news article body for incongruent news article detection.

Human weight annotation to sentences: One of the authors of this study manually read the news body and assigned weights to sentences based on their role in the summarized text of the news body. The weights to the sentence have been assigned between the scale of 0 to 5 as follows: **scores between $\geq$0 and $<$1 :**if the sentence has no effect in the summary of the news body, or it should not be present in the summary of news body. In Otherworld, it gives redundant information to other sentences, **scores between $\geq$1 and $<$2 :** summary of news body will not have any effect irrespective of whether the sentence is present or not present in the summary of news body, **scores between $\geq$2 and $<$3 :** if some information from sentence must be present in the summary of news body, $\geq$3 and $<$4 if the sentence should

be present in the summary of news body, **scores between $\geq$4 and $<$5 :** If this sentence is sufficient to represent a paragraph or whole news body as a summary of a paragraph or news body.

5.9.9 Implications of Dual Summarization

As discussed in section 5.1 random fake news in the form of incongruent news article may be of different forms; such as (i) The headline is not related to its body, (ii) Both the headline and its body are related, but the claim made in the headline is different from its body, (iii) Both the headline and its body report a genuine event/incident, but the dates, numeric values or name entities are manipulated, (iv) Textual noise (paragraphs/sentences) extracted from other sources is inserted into a genuine news article (referred as partially incongruent). Considering the characteristics of datasets discussed in section 3.2.1, the incongruent news sample in NELA dataset is partially incongruent in nature and represents characteristics (iv) of incongruent news. Similarly, as discussed in section 3.2.1 unrelated class samples of the FNC dataset represent fully incongruent news articles and represent the characteristics (i) of incongruent news. Fake news in ISOT datasets represents the characteristics (ii) of incongruent news. From table 5.2, it is evident that proposed methods $DuSum(ST,\theta)$ outperform both similarity and summarization-based models with a significantly high margin over the NELA dataset. Similarly, from table 5.2, it is also evident that proposed methods $DuSum(ST,k)$ outperform both similarity and summarization-based models with a significantly high margin over FNC and comparable high performance over ISOT datasets. From such observations, it is apparent that our proposed model is more effective in detecting incongruent news articles of different characteristics. Considering the performance of *DuSum* over three publicly available benchmark datasets of different characteristics, it can be claimed that *DuSum* is applicable in real misinformation detection such as partially incongruent news, fake news detection, verifying the consistency within a document, the false connection between headline and body, the false context between headline and body, manipulated and fabricated content in news article[14].

5.9.10 Comparison of Models over English and Hindi Datasets

To study the response of state-of-the-art models in literature over Hindi and English datasets and the difference in the response over English and Hindi datasets, we conduct an empirical

[14] https://firstdraftnews.org/long-form-article/understanding-information-disorder/

Table 5.8 Comparing the performance of models over proposed training datasets for the Hindi language over BBC corpus and the performance of models over the English datasets. (i) **Acc** : indicates the accuracy, (ii) **F** indicates F-measure score.

Models		BBC corpus								English					
		SM		NE-R		POS-R		POS-NE-R		NELA - 17		ISOT		FNC	
		Acc	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**
Feature	**BiLSTM**	0.840	0.839	0.926	0.926	0.991	0.991	0.991	0.991	0.555	0.550	0.990	0.990	0.616	0.504
	RoBERT	0.894	0.893	0.983	0.982	**0.998**	**0.998**	0.996	0.995	0.615	0.613	0.996	0.996	0.664	0.583
Hierarchical	**AHDE** [28]	0.691	0.671	0.869	0.869	0.963	0.963	0.948	0.949	0.606	0.606	0.913	0.913	0.691	0.454
	HDSF [36]	0.889	0.888	0.983	0.983	0.994	0.994	0.994	0.994	0.517	0.494	0.720	0.712	0.758	0.666
	HoBERT	0.899	0.898	**0.985**	**0.984**	0.997	0.997	0.997	0.997	0.635	0.634	0.991	0.991	0.686	0.632
	HeLSTM	0.909	0.908	**0.985**	**0.984**	0.995	0.995	**0.998**	**0.998**	0.602	0.602	0.997	0.997	0.689	0.597
Summarization	**FEDS** [38] [31]	0.514	0.502	0.505	0.504	0.490	0.488	0.498	0.490	0.533	0.532	0.998	0.998	0.878	0.837
	MADS(**BiLSTM**, $\beta = 0.5$, **H = 8**)	0.898	0.897	0.934	0.934	0.949	0.949	0.943	0.943	0.581	0.575	0.999	0.999	0.971	0.963
	MADS(**S-BERT**, $\beta = 0.5$, **H = 8**)	0.851	0.850	0.505	0.496	0.509	0.500	0.949	0.949	0.568	0.562	0.978	0.977	0.962	0.952
	DuSum(**BERT**, $\beta = 0.5$)	0.905	0.914	0.866	0.866	0.996	0.966	0.996	0996	0.701	0.701	0.99	0.99	0.991	0.998

study over different models from the literature. We consider our baseline model *BiLSTM* and *RoBERT* along with state-of-the-art models from literature **AHDE** [28], **HDSF** [36] and **FEDS** [38] [31] as models to study the difference in response over Hindi and English datasets. We also consider proposed models from this thesis **HoBERT** [39], **HeLSTM** [39] and dual summarization-based models *MADS* and *DuSum* to study the difference in response over Hindi and English datasets. Table 5.8 presents the performance over proposed Hindi dataset and existing English datasets in literature. From Table 5.8, it is evident that the response of models on the split and merge approach (SM) of the Hindi dataset and the NELA-17 dataset for the English language are similar. The performance of the dual summarization-based method *DuSum* is superior on both datasets. Such similar performances of models are not surprising considering the dataset curation process of the split and merge approach (SM) of the Hindi dataset and NELA-17 dataset for the English language. Similarly, from Table 5.8, it is also evident that existing models in literature which are proposed to detect incongruent news and fake news for the English language are also effective in detecting incongruent and fake news articles for the Hindi language. From Table 5.8, it is apparent that performance of models in literature over English datasets and performance of models over the Hindi datasets correlated to each other; hence it can be concluded that proposed Hindi datasets are similar in nature to English datasets.

5.9.11 Response of Models over Hindi Datasets

We conducted an empirical study of different models from the literature to study the response of state-of-the-art models in literature over the proposed Hindi datasets for fake and incongru-

Table 5.9 Comparing the performance of models trained over synthetic datasets and tested over real fake news datasets. (i) **Acc**: indicates the accuracy, (ii) **T** and **F** indicates F-measure score for *True* news and *Fake* news class respectively.

		SM			NE-R			POS-R			POS-NE-R		
	Model	**Acc**	**T**	**F**	**Acc**	**T**	**F**	**Acc**	**T**	**F**	**Acc**	**T**	**F**
BBC	**BiLSTM**	0.603	0.703	0.402	0.688	0.730	0.630	0.465	0.622	0.087	0.551	0.657	0.352
	RoBERT	0.775	0.722	0.811	0.705	0.673	0.732	0.746	0.669	0.794	0.743	0.666	0.792
	AHDE [28]	0.699	0.572	0.768	0.691	0.723	0.650	**0.762**	**0.630**	**0.750**	0.754	0.690	0.796
	HDSF [36]	0.473	0.640	0.022	0.673	0.694	0.649	0.750	0.680	0.794	**0.765**	**0.714**	**0.851**
	HoBERT	**0.783**	**0.734**	**0.817**	0.753	0.691	0.794	0.747	0.672	0.794	0.744	0.668	0.791
	HeLSTM	0.640	0.631	0.648	**0.756**	**0.692**	**0.799**	0.741	0.662	0.791	0.741	0.663	0.790
	FEDS [38] [31]	0.500	0.667	0.502	0.500	0.666	0.498	0.500	0.667	0.511	0.504	0.542	0.671
	MADS(BiLSTM, $\beta = 0.5$, H = 8)	0.691	0.590	0.752	0.516	0.643	0.249	0.737	0.659	0.787	0.739	0.658	0.786
	MADS(BERT, $\beta = 0.5$, H = 8)	0.716	0.643	0.765	0.500	0.666	0.498	0.504	0.667	0.502	0.740	0.665	0.788
	DuSum(BERT, $\beta = 0.5$, H = 8)	0.755	0.684	0.799	0.51	0.666	0.21	0.741	0.663	0.79	0.739	0.666	0.786

ent news article detection. We consider our baseline model *BiLSTM* and *RoBERT* along with hierarchical encoding based state-of-the-art models from literature **AHDE** [28], **HDSF** [36] and **FEDS** [38] [31] as models to study the response of models over the proposed datasets for Hindi. We also consider proposed models from this thesis **HoBERT** [39], **HeLSTM** [39] and dual summarization-based models *MADS* and *DuSum* to study the response of dual summarization-based models over the over Hindi and English datasets. Table 5.9 presented the response of models trained using proposed Hindi datasets and tested using real fake news datasets. Table 5.9 shows that hierarchical encoding-based methods are more suitable for fake news detection in the Hindi language compared to similarity and dual summarization methods.

Summary

This thesis chapter proposed dual summarization-based methods, a Multi-head Attention Dual Summarization model, *MADS* and dual summary-based method *DuSum*, for detecting incongruent news articles of different characteristics. *MADS* extract two types of summary, viz. multi-head attention and convolution summary over positive and negative set separately. Similarly, *DuSum* also extract two different forms of summaries: *(i)* sequential weighted summation summaries and *(ii)* convolution summaries. Subsequently, estimate similarity features to check the similarity between headline and body and consistency within news body for incongruent news article detection. From various experimental observations, it is evident that the proposed models outperform all of the baseline models for all three datasets of different characteristics. It is further observed that the proposed model is capable of capturing

not only incongruent news articles, but also the partially incongruent news articles. We also conducted several ablation studies to study the strengths and weaknesses of the proposed models *MADS* and *DuSum*. The key observations from our ablation studies are as follows: *(i)* considering the summaries of both highly and poorly set is more effective than considering the summaries of only poorly congruent set or least k sentences; however, considering the summaries of only poorly congruent set or least k sentence is more effective than state-of-the art similarity and summarization-based approach for incongruent news article detection (refer subsection 5.9.2). Similarly, in the case of *MADS*, we also observe that considering the summaries of both positive and negative sets is more effective than considering the summaries of only the negative set; however, considering the summaries of only the negative set is more effective than state-of-the-art similarity and summarization-based approach for incongruent news article detection (refer subsection 5.9.3). *(ii)* Considering sequence-weighted summation and convolution summaries together boosts the performance of both the proposed *MADS* and *DuSUM* (refer to subsection 5.9.4). *(iii)* performance of *DuSUM* models is superior for θ and K value 0.5 and 5 respectively (refer to subsection 5.9.5). *(iv)* The performance of *DuSum* with contextualized LSTM is superior over *DuSum* with BiLSTM (refer subsection 5.9.6). This study can be extended in multiple directions; one significant extension of this study could be to study the effect of considering triplet loss function, contrastive learning, or considering attention between summaries, instead of estimating similarity between summaries by estimating similar features in subsection 5.6.1.5. Further, this study discerns the following potential future works. (i) *topic aware summarization* - identify the topic present in the header and generate the topic-specific summarization, (ii) *Knowledge-based summarization* - most of the news articles are related to some real-world event or entities. Summarization can be enhanced by considering knowledge based such as Wikipedia or news achieves.

Chapter 6

Headline-Centric Approaches for Incongruent News Detection Using Graph-Based Context Matching

The prevalence of deceptive and incongruent news headlines underscores their pivotal role in propagating fake news, exacerbating the spread of misinformation and disinformation. In the literature, researchers have explored various bag-of-word-based features [22][35] [23], news body-centric [1] [28] [30] [29] and news headline-centric encoding methods [37] [21] for incongruent news article detection. However, headline-centric and body-centric approaches in the literature fail to detect partially incongruent articles efficiently. Motivated by the above limitations, this study proposes *Graph-based Dual Context Matching* **GDCM** and *Graph-based Context Matching* **GCM**, both the proposed methods first represent headlines and news bodies as a bigram network to capture contextual relations between words and document structure. Then, for every word in the headline, both methods extract dual context from the news body Bigram Network. Next, it estimates the similarity between the extracted context and the headline for incongruent news detection.

6.1 Introduction

Detecting and incongruent news headlines have proven to be a potent catalyst in disseminating fake news, leading to a dual impact that exacerbates the spread of misinformation and disinformation [7]. First, when incorporated into trending news articles, these headlines lure

readers into engaging with the content, thereby amplifying the reach of the false information. Second, they perpetuate a cycle of misinformation by distorting facts and manipulating readers' perceptions, undermining the credibility of legitimate news sources. These two-fold harms significantly contribute to the spread of misinformation and disinformation in today's media landscape [3, 8] [1,2]. A news article is said to be incongruent if the news headline misrepresents its body through fabrication, manipulation, false connections[3], or incorrect context[4] [2–4]. As delineated in studies [10, 11, 4], misleading headlines play a significant role in making a news viral on any social media and also influence [9] the opinion of readers. Incongruent news can negatively affect readers, such as false beliefs and wrong opinions [5,6] [2, 13, 14]. Once such misleading information spreads, the study [17] found that rectification measures do not have much impact, and in certain instances, readers may even backfire by reinforcing individuals' misconceptions instead. Consequently, detecting deceitful and incongruent news articles [3, 2, 18–20] is becoming an important research problem to counter the spreading of misinformation over digital media. As reported in studies [3, 21] there are four key characteristics of incongruent news articles. (i) The claim made in the headline is unrelated to or contradicts the claim made in the news body. (ii) News headlines and body are from the same topic and about the same event, but the content in the headline and body are not related to each other. (iii) The headline and body communicate or describe an authentic event or incident, but the dates or names of entities in the news headline and body are manipulated. (iv) one or more paragraphs of the news body are congruent to the headline, and one or more paragraphs of the news body are not congruent to the headline (referred to as *partially incongruent*). As per the studies [1, 21], based on the above characteristics, there are two types of incongruent news articles, namely *(i)* fully incongruent and *(ii)* partially incongruent. Incongruent news which follows the above (i), (ii), or (iii) characteristics may be referred to as fully incongruent news. Similarly, a news article is considered partially incongruent if some portions of the news body are incongruent with the news headline and the rest are congruent with the new headline. Intuitively, partially incongruent news follows (iv) characteristics of incongruent news mentioned above.

Studies on incongruent news detection can be broadly grouped into three groups: *bag-of-word feature*, *body centric* and *headline centric*. Bag-of-word-based studies [22, 23, 35] identified various types of features, such as n-grams and topic modeling, latent features

[1] Misleading headline fake news regarding WHO
[2] Misleading headlines as fake news
[3] When the caption of the image does not align with its image or the headline does not support its content.
[4] Legitimate information is presented in the wrong context
[5] Effects of the misleading headline on health
[6] Impact of misleading headlines related to economy

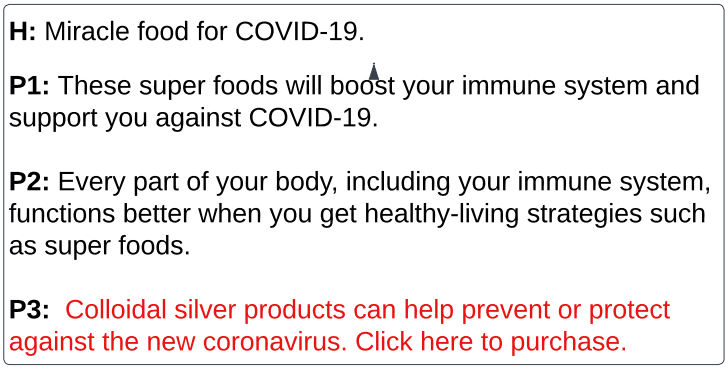
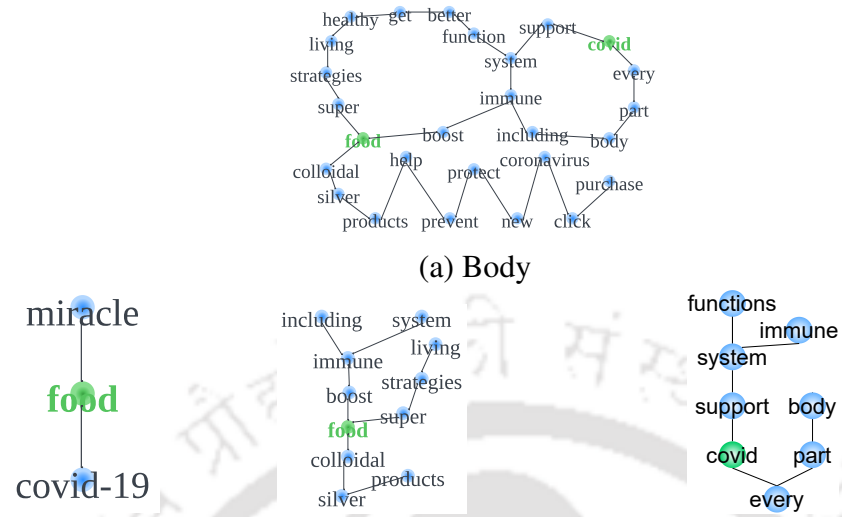


Fig. 6.1 Present example of partially incongruent presented in the study [1]. Where **H** represents the headline, $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ indicates the new body's first, second, and third paragraphs. It is evident that paragraphs $\mathbf{P}_1$ and $\mathbf{P}_2$ are congruent to headline H because they are in the context of food and health. But **H** are not congruent because the headline is in the context of food and health, but P3 talks about the advertisement for some Colloidal silver products.

and sentiment, etc., and built a classifier. However, as reported in the study [98] bag-of-word feature-based models failed to detect incongruent news articles, which requires the understanding of propositional content, complex negation, and deep semantic relationships. In body centric approaches, the idea is to generate an encoding of the news body and compare its similarity with that of the headline. Representation of the body is obtained either by applying sequential encodings [98, 25], hierarchical encodings [36, 27, 28, 39] or summarization [29–31]. Whereas in the headline-centric approach, encoding of the news body is guided by considering the information in the headline [37, 21]. A recent headline-centric dual summarization-based study [21] also strives to detect partially incongruent news by generating a dual summary (positive set and negative set) of the news body. Chapter 5 focuses on detecting partially incongruent news through dual summarization, where they generate positive and negative summaries of news articles. However, it does not consider contextual information from headlines while generating summary representative vectors of positive and negative sets. Consequently, it fails to capture the positive (segment of news body where discussion regarding news headline is present) and negative context (segment of news body where discussion regarding news headline is not present) in different segments of the news body.



(a) Body

(b) Headline (c) 3-hop sub-graph of *food* (d) 3-hop sub-graph of *covid*

Fig. 6.2 Presents *Bigram* network of (**a**) news body *Bigram* network of the example document in Figure 6.1, **(b)** headline *Bigram* network of the example document in Figure 6.1, and (c) 3-hop subgraph of the word *food* extracted from news body *Bigram* network **(a)**. From the subgraph (Figure (b)) of the headline *Bigram* network, it is evident that the context of the keyword *food* in the headline is *Miracle* and *COVID*, but the neighbours (context) of the subgraph with *food* (Figure (c)) constructed by extracting 3-hop neighbour of *food* from body *Bigram* network (Figure (a)) is *super*, *strategies*, *living*, *boost* and *immune* which from congruent part of news body (Paragraph $\mathbf{P}_1$ and $\mathbf{P}_2$). But other neighbours in the subgraph of *food* in the body *Bigram* network are *colloidal*, *silver*, and *product*, which is from incongruent paragraphs ($\mathbf{P}_3$) of the new body. So, the contexts of the subgraph with *food* are similar to paragraphs P1 and P2, but the context of *food* in the headline and paragraph P3 do not match. When we generate the encoding of the subgraph with *food*, due to the presence of **P3**, its similarity with the encoding of the headline will be less than the scenario where **P3** is absent. Similarly, In the subgraph (presented in Figure 6.1) of the headline word *COVID* (Figure (d)), constructed by extracting 3-hop neighbours from the body Bigram network, it is evident that the context of *COVID* in the subgraph and headline is slightly similar. When we generate the encoding of the subgraph with *COVID*, its similarity with the encoding of the headline will be high. Further, no node with the headline keyword *miracle* or nodes with the close meaning of *miracle* is present in the body Bigram network. Therefore, considering the context of the headline and the context of *food* and *miracle* in the news body, it can be concluded that the given news article is incongruent. Because some context of food in the body matches with the headline, and some contexts do not match, and no context or presence of a miracle is found in the body *Bigram* network. The proposed **GCM** model is designed to capture the above intuitions.

H: 'Start Here': Trump demands to meet whistleblower as Schiff confirms testimony

1. Whistleblower testimony.
President Donald Trump said he wants to meet the whistleblower whose anonymous complaint spurred an impeachment probe.

....

2. **Hong Kong chaos.**
There were violent clashes between pro-democracy demonstrators and police in the streets of Hong Kong over the weekend.

....

3. Child arrests.
The two 6-year -old-students in Orlando, Florida, whose arrests sparked nationwide outrage and the firing of a school resource officer, are counted among thousands of child arrests in the U.S. annually, raising questions about how school officials discipline misbehavior.

Fig. 6.3 Present example of partially incongruent presented in the study [1]. Where **H** represents the headline and 1. *Whistleblower testimony*, 2. *Hong Kong chaos* and 3. *Child arrests* indicates the news body's first, second and third paragraphs. It is evident that paragraph 1. *of Whistleblower testimony* is congruent to the headline **H**. But the paragraphs 2. *Hong Kong chaos* and 3. *Child arrests* are not congruent with the headline **H**. Because the headline is in the context of President Donald Trump wanting to meet a Whistleblower, paragraphs 2. and 3. are about *Hong Kong chaos* and *Child arrests*

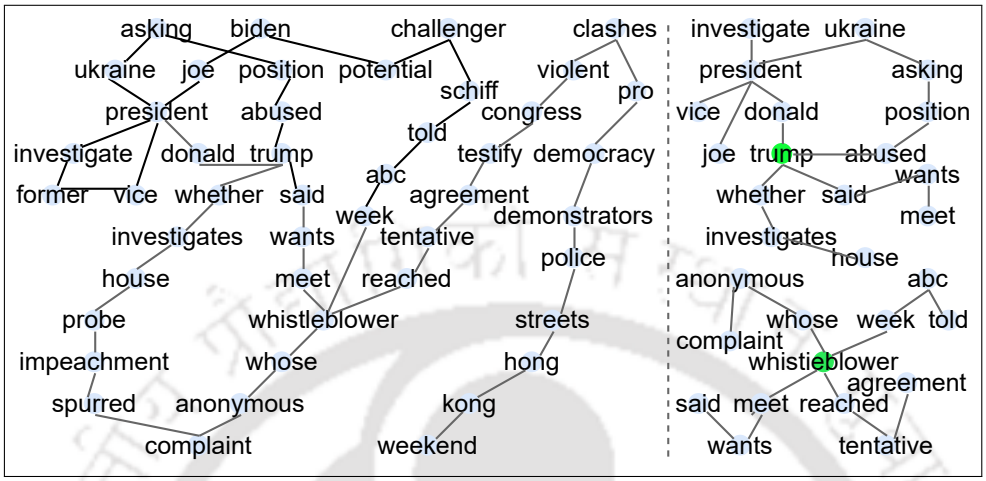


Fig. 6.4 Present the bigram network of news body (left) and subgraph (right) related to keywords in headlines "trump" and "whistleblower". This Figure considers the news body and headline from Figure 6.3. Here, **"start trump demands meet whistleblower Schiff confirms testimony agreement"** is the headline. We can observe that neither the headline keywords "start", "demands", and "testimony" are directly present, nor the words semantically closer to the headline keywords "start", "demands", and "testimony" are present in the body bigram network (left). Suppose neither the words from the headline are directly present in the body bigram network nor the words semantically similar to the words in the headline are present in the body bigram network. In that case, it indicates the headline does not represent the news body, and possibly, the given news article is incongruent news. Further, On the right side, this Figure also presents the subgraph for the headline keywords "trump" and "whistleblower", which is constructed by extracting three hop neighbour of nodes with "trump" and "whistleblower" in the body bigram network (presented on the left of this Figure). Now we can see the context of "trump" and "whistleblower" in the subgraph and headline are similar. Therefore, the given news article appears to be congruent news, but if we match paragraphs 2. and 3. of the news body and headline of the news article in Figure 6.3, we can confirm that this news article is incongruent. This observation suggests that matching the positive context only is insufficient for incongruent news detection. Accordingly, it is essential to match the headline with both the positive context of the headline extracted from the news body and the negative context of the headline extracted from news body for effective incongruent news article detection. The proposed model **GDCM** is based on the above observations and intuitions.

6.2 Motivation

Initial studies [22, 23, 35] on incongruent news articles mainly focused on bag-of-word features to estimate the similarity between headline and body. Further, studies on incongruent news detection can be broadly categorized into two groups – *headline centric* and *body centric*. Body centric approaches aim to encode the news body using sequential [98, 25], hierarchical [36, 27, 28, 39] or summarization [29–31] methods. Then, the similarity between the headline and body is estimated. Whereas in the headline-centric approach, encoding of the body centric approach is guided by considering the information in the headline [37, 21]. Headline-centric dual summarization-based method [21] also attempted to detect partially incongruent news with dual summarization (congruent and incongruent summary) by splitting the sentences of the body into two sets. But study [21] does not consider the contextual similarity of headlines while generating the summary. Consequently, it fails to capture the context of headline present in different segments of the news body. Motivated by the above limitations, this thesis proposes two methods based on graph context matching – *Graph-based Context Matching* **GCM** and *Graph-based Dual Context Matching* (**GDCM**).

6.3 Research Objective

1. This Chapter also proposes *Graph-based Context Matching* **GCM**, which extracts contexts of the words in the headline from different segments of the body and then estimates the similarity between the headline and extracted contexts. As reported in the study, [169, 170] n-gram network is more suitable to represent the document's structure and contextual relation between words in a text document. We represent the headline and body using the bigram network, then extract context related to the headline by constructing several subgraphs related to words in the headline, and then estimate the similarity between the headline bigram network, subgraph, and the body bigram network for incongruent news detection. Figure 6.2 presents the intuition behind our proposed model **GCM**, considering an example of partial incongruent news in Figure 6.1.

2. This Chapter proposes **GDCM** method, which extracts both positive context (paragraphs of news body where discussion regarding headline is present) and negative context (paragraphs of news body where discussions regarding headline is not present) of the words in headlines from different segments of news body. Next, we estimate the similarity between the headline and extracted context (positive context and negative context) from the news body. The key intuition behind our proposed model **GDCM** is that if news is congruent, then both positive context and negative context regarding headlines in different segments of the news body will be similar to the headline, and

similarly, in the case of partially incongruent news the negative context regarding headline in different segments of news body will not be similar to the headline. So, extracting both the positive and negative context of headlines from the news body is important for efficient incongruent new detection. Figure 6.4 presents the above observations and intuitions in detail by considering the partially incongruent news presented in Figure 6.3.

6.4 Contributions

1. We propose *Graph-based Context Matching* **GCM**, which extracts contexts of the words in the headline from different segments of the body and then estimates the similarity between the headline and extracted contexts.

2. We also propose a Graph-based Dual Context Matching **GDCM** model, which extracts the context of a headline from paragraphs of the news body where discussion regarding the headline is present (positive context) and also extracts the context of a headline from paragraphs of the news body where discussion regarding the headline is not present (negative context). Then estimate the similarity between the headline, positive context, and negative context for incongruent news detection.

3. We conduct experiments over three benchmark datasets to establish the superiority of the proposed **GDCM** model over state-of-the-art methods in the literature for incongruent news detection.

4. We also perform various ablation studies to understand and demonstrate the importance of extracting both positive and negative context regarding headlines from different segments of the news bodies for partial incongruent news article detection.

6.5 Literature Review

In the literature, studies [130–136] have briefly assessed and evaluated the existing work related to misinformation and disinformation detection. This study focuses on retrospectively reviewing works related to incongruent news article detection. As reported in the study [3], incongruent news is different from other forms of misinformation, and its dissemination contributes to the spread of fake news articles across digital platforms. While incongruent news detection and clickbait detection are related to news headlines, it is crucial to note that incongruent news detection differs entirely from clickbait detection [6, 3]. Incongruent news article detection involves identifying the relationship between the headline and news body and identifying news articles presenting information that contradicts or conflicts with credible

sources or facts. On the other hand, clickbait detection focuses on recognizing content with sensational or misleading headlines designed to attract clicks but may not necessarily involve misinformation or incongruent information [3]. Therefore, despite incongruent news and clickbait detection being related to news headlines, the incongruent news detection and clickbait detection methods are utterly different in their approaches and objectives [21, 39].

Initial studies on incongruent news article detection utilize bag-of-words-based features such as n-grams, TF-IDF and topic modeling features [22][24][137][35][138]. Considering the importance of contextual and sequential information present in headlines and bodies, the studies [24][25][40] combine sequential encoding of news headline and body with bag-of-words based feature for incongruent news detection. Further studies on incongruent news can be grouped into two groups, namely *Headline-centric encoding* and *Body-centric encoding*. Body centric encoding approach aims to encode or summarize news bodies to obtain an efficient representation of news headlines and body, which helps estimate similarity with headlines for incongruent news detection. Body centric encoding approach can be further categorized into four categories, namely (i) sequential encoding, (ii) hierarchical encoding, (iii) deep hierarchical encodings, and (iv) summarization. Sequential encoding-based methods [98, 25] apply sequential encoding over news body and headline to obtain an encoded representation of news body, and headline then estimates the similarity between encoded representation of news headline and body. Considering the hierarchical structure and hierarchical information presented in the news article, studies [36, 27, 28] exploit the hierarchical structure of the news article while encoding the news headline and news body. However, the above hierarchical encoding-based studies consider the hierarchical structure of news body up to paragraph level only, consequently failing to capture long-term dependencies between words of sentences [39]. Considering such limitation of hierarchical encoding-based methods [36, 27, 28], study [21] proposed deep hierarchical encoding based which extends the hierarchical structure of news article from news body to paragraphs, paragraphs to sentences and sentences to words while encoding the news body. Then estimates the similarity between the headline and news body to detect incongruent news articles. However, the above-mentioned sequential encoding, hierarchical encoding, and deep hierarchical encodings-based methods fail to detect incongruent news in the case of text length (concise headline with a very high number of sentences of the paragraph of sentences in news body) mismatch [30, 21]. To address the above limitations, [29–31] summarize the news body to generate a small representation headline and then estimate the similarity between the representative headline and real news headline for incongruent news detection.

6.5.1 Research Gaps: Summary

However, two key limitations exist with the above-discussed body-centric encoding-based (sequential encoding, hierarchical encoding, deep hierarchical encodings, and summarization) studies. (i) Fails to detect incongruent news articles that are incongruent due to manipulation of numeric figures or named entities (incongruent news with (iii) characteristics) [37]. (ii) Fails to detect partially incongruent news articles (incongruent news with (iv) characteristics) [21]. Headline-centric encoding studies [37, 21] attempt to overcome the above limitations by encoding the news body based on information in the news headline. Headline-centric encoding-based study [37] applies cardinal Part-of-Speech Tag patterns-based attention to focus on cardinal and number value matching between headline and body and also uses headline guided attention between headline and news body to highlight sentences of news body which are more relevant to the news headline while encoding the news body. Similarly, a headline-centric encoding-based study [21] proposed a dual summarization-based method to detect partially incongruent news articles. Study [21] first, constructs a congruent set and incongruent set by placing the sentences into the congruent set and incongruent set based on the contextual similarity between sentence and headline, then obtains a summary separately for both congruent and incongruent set. However, study [21] considers the contextual similarity between headlines and sentences just to split the sentences into congruent and incongruent sets and does not consider the contextual similarity between headline and sentences of news body while generating the summary of the congruent and incongruent set. So, it fails to capture the context of headlines in different paragraphs or sentences of the news body. Consequently, study [21] also fails to detect partially incongruent news articles efficiently. Motivated by the above challenges and limitations in the literature, we proposed two methods, *Graph-based Context Matching* **GCM** and *Graph-based Dual Context Matching* **GDCM**. Our proposed model, *GDCM* first represents headline and news body in a bigram network to capture the global context of words within news headline and body. Next, it extracts the positive context (segment of the news body where discussion regarding headline keywords is present) and negative context (segment of news body where discussion regarding headline keywords is not presented) for words in the headline from the body Bigram network. Subsequently, it estimates the similarity between headlines and positive and negative contexts for incongruent news detection. Similarly, the *Graph-based Context Matching* **GCM** model aggregates the context (segment of news body where discussion regarding headline is present) of a headline from different segments of the news body and then applies to match between aggregated context from different segments of news and the headline.

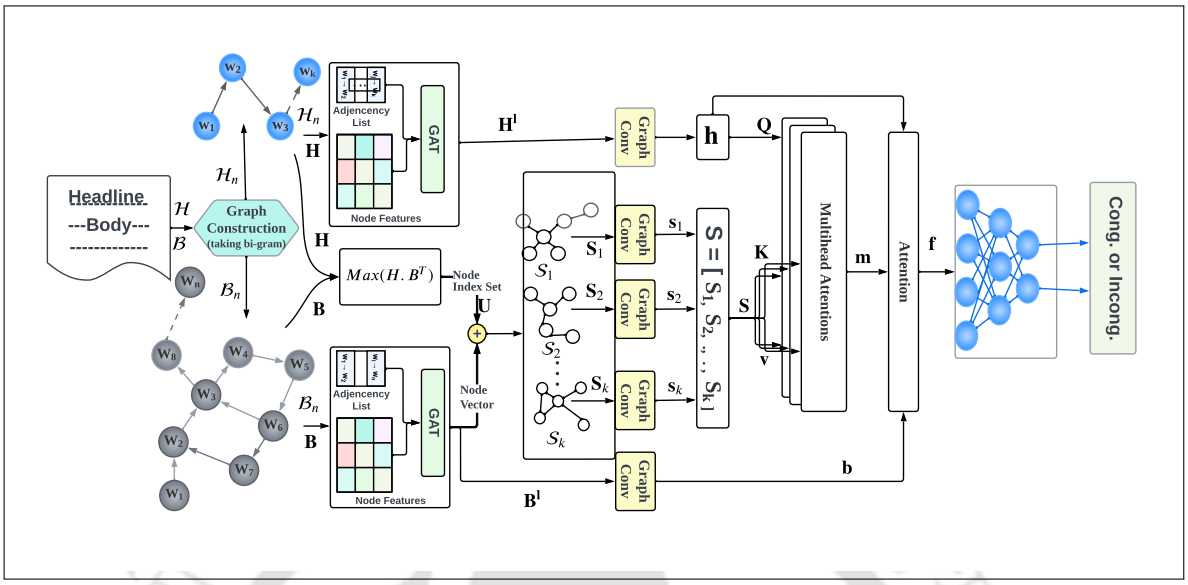


Fig. 6.5 The Working of proposed model **GCM** models is presented in the diagram.

6.6 Proposed Framework: Graph Context Matching

In this section, we present our proposed model, *Graph-based Context Matching* **GCM** and *Graph-based Dual Context Matching* **GDCM**. Figure 6.5 presents the working diagram of *Graph-based Context Matching* **GCM** and Figure 6.6 presents a working diagram of *Graph-based Dual Context Matching* **GDCM** model.

6.6.1 Graph Context Matching based Approach GCM

Figure 6.5 presents a block diagram of the proposed model, `Graph-based Context Matching` **GCM**. Given a news article with a pair of headline $\mathcal{H}$ and body B, **GCM** model, constructs a bigram network [170] for $\mathcal{H}_n$ and $\mathcal{B}_n$ by considering unigram as node and edge between nodes if two words co-occur, respectively. The prime instinct behind representing news headlines and news bodies in the form of bigram networks is that ngram networks are very efficient in representing document structure and contextual relationships among the words within documents [169, 170]. Next, initialize the nodes of $\mathcal{H}_n$ and $\mathcal{B}_n$ with the fastText embedding [171] of the word associated with nodes and form an embedding matrix **H** and **B** for $\mathcal{H}_n$ and $\mathcal{B}_n$, respectively. Now, we obtain a matching node index set **U** by considering the column index of row-wise maximum in every row of the matching matrix **M** and the

matching matrix **M** is obtained by the equation defined below,

$$\mathbf{M} = \mathbf{H}^{\mathbf{T}} \cdot \mathbf{B} \tag{6.1}$$

where ' · ' is a matrix multiplication operation. Each entry **M**[**i**, **j**] in the matching matrix **M** indicates the similarity between i^{th} node of the headline bigram network $\mathcal{H}_n$ and j^{th} node of the body ngram network $\mathcal{B}_n$. The prime instinct behind obtaining a node index set **U** is that if a news article is congruent, then either exactly the same, synonyms or contextually similar to nodes in the headline will be present in $\mathcal{B}_n$ and the index of that node will be added in **U**. Whereas in case an incongruent news node index is added in, **U** from $\mathcal{B}_n$ will not be exactly the same, synonyms or less contextually similar to nodes in $\mathcal{B}_n$. Similarly, in the case of partial incongruent, some node indexes in **U** from $\mathcal{B}_n$ will be exactly the same, synonyms or contextually similar to nodes in $\mathcal{B}_n$ and some node indexes in **U** from $\mathcal{B}_n$ will not be exactly the same, synonyms or less contextually similar to nodes in $\mathcal{B}_n$. Subsequently, with the motivation to learn local neighbourhood structure and update the embedding of nodes based on the context of their neighbour nodes, we apply **t** layer graph attention network **GAT**, similar to the study [93], over $\mathcal{H}_n$ and $\mathcal{B}_n$, and obtain updated node embedding matrix $\mathbf{H^l}$ and $\mathbf{B^l}$ for $\mathcal{H}_n$ and $\mathcal{B}_n$, respectively. Next, we apply convolution over $\mathbf{H^l}$ and $\mathbf{B^l}$ to obtain a global representation vector **h** and **b** of headline and body bigram network, respectively.Here **h** is the encoded representation headline bigram network $\mathcal{H}_n$ and **b** is the encoded representation body bigram network $\mathcal{B}_n$. The further details to obtain global representation vector **h** and **b** are presented in subsection 6.6.1.1. Next, we form a subgraph $\mathcal{S}_i$ for each node $\mathbf{u}_i$ in the node index set **U** by sampling x hop neighbour (radius) of a node $\mathbf{u}_i$ from the body bigram network $\mathcal{B}_n$. Now, we obtain a representation vector $\mathbf{s}_i$ for each subgraph $\mathcal{S}_i$ and pass it to multi-head attention [58] as value and key with representation vector of headline h as a query and obtain a representation vector **m**. The prime motivation behind multi-head attention between the representation of headline **h** and representation of subgraphs $\mathbf{s}_i$ is that if a news article is congruent, then the headline and subgraph should be close; consequently, multi-head attention should give high attention weight to subgraph representation $\mathbf{s}_i$. Whereas in the case of an incongruent news article context, the headline and subgraph should be dissimilar; accordingly, multi-head attention should give low attention weight to subgraph representation $\mathbf{s}_i$. The further details of headline subgraph context matching are presented in subsection 6.6.1.2. Finally, we aggregate and combine the headline global representation vector **h**, body global representation vector **b** and context machining vector **m** using scaled attention [58], additive attention [83] or similarity feature-based approach [39, 21] to form a feature vector f and then the feature vector f is passed through a neural network for classification. The prime intuition behind applying scaled attention [58], additive attention [83] or similarity feature [39, 21]

based aggregation is that if the news is congruent, then the representation of the headline **h** should be highly similar to both representations of subgraph **m** and representation of body bigram network **b**. But in the case of an incongruent news article, the representation of headline **h** should be least similar to both representations of subgraph **m** and representation of body bigram network **b**. Similarly, in the case of partial incongruent news, the representation of headline **h** may or may not be similar to representations of subgraph **m** but **h** will be the least similar to the representation of body bigram network **b**. Subsection 6.6.1.3 presents a detailed description of aggregate and combine steps.

6.6.1.1 Global Representative Body and Headline Bigram Network

Given the bigram networks $\mathcal{H}_n$ and $\mathcal{B}_n$ along with their node embedding matrix **H** and **B**, the attention coefficient to nodes in the neighbourhood of $\mathbf{j}^{th}$ a node in the body bigram network $\mathcal{B}_n$ is computed as defined below.

$$\alpha^l_{j,r} = \left(\frac{\exp(\mathbf{e}^l_{j,r})}{\sum_r \exp(\mathbf{e}^l_{j,r})}\right) \tag{6.2}$$

Where $\mathbf{e}^l_{j,r}$ is estimated as defined below.

$$\mathbf{e}^l_{j,r} = \mathbf{LeakyReLU}(\mathbf{a}^{\mathbf{T}}[\mathbf{W}\mathbf{b}_j || \mathbf{W}\mathbf{b}_r]) \tag{6.3}$$

Where $\alpha^l_{j,r}$ denotes the importance of the node j to r. Now node embedding of j^{th} node is updated as defined below.

$$\mathbf{b}^{l+1}_j = \blacksquare_{\mathbf{r}\in\mathbf{N}(\mathbf{j})}\ \alpha_{j,r}\mathbf{W}^l\mathbf{b}^l_r \tag{6.4}$$

Where $\mathbf{N}(\mathbf{j})$ contains the index of neighbour nodes of $\mathbf{j}^{th}$ node and $\mathbf{W}^l$ is a learnable parameter. Similarly, we also obtain $\mathbf{H}^l$ by following equations 6.2, 6.3 and 6.4. Next, we obtain the global representation vector **h** and **b** of $\mathcal{H}_n$ and $\mathcal{B}_n$ by concatenating the min, max and average pulling over $\mathbf{H^l}$ and $\mathbf{B^l}$ respectively.

$$\mathbf{h} = \left(\mathbf{Max}\left(\mathbf{H}^l\right) \oplus \mathbf{Min}\left(\mathbf{H}^l\right) \oplus \mathbf{Avg}\left(\mathbf{H}^l\right)\right) \tag{6.5}$$

$$\mathbf{b} = \left(\mathbf{Max}\left(\mathbf{B}^l\right) \oplus \mathbf{Min}\left(\mathbf{B}^l\right) \oplus \mathbf{Avg}\left(\mathbf{B}^l\right)\right) \tag{6.6}$$

Where **Min**, **Max** and **Avg** indicate min pooling, max pooling and average pooling operation.

6.6.1.2 Headline and Subgraphs Context Matching

Given a matching index node set $\mathbf{U}$ and node embeddings matrix $\mathbf{B}^l$ of the body bigram network, we construct subgraph $\mathcal{S}_i$ for each node $\mathbf{u}_i$ in the best matching node set $\mathbf{U}$ by sampling their x hop neighbour (radius) in the body bigram network $\mathcal{B}_n$. Next, considering $\mathbf{N}\left(\mathbf{i}\right)$ as neighbour node index for nodes in $\mathcal{S}_i$, we obtain subgraph embedding matrix $\mathbf{S}_i$ by extracting embedding of nodes in $\mathbf{N}\left(\mathbf{i}\right)$ from bigram body embedding $\mathbf{B}^l$ as defined below.

$$\mathbf{S}_i = \left(\mathbf{B}^l\left[\mathbf{u}(\mathbf{i})\right] \oplus \mathbf{B}^l\left[\mathbf{N}(\mathbf{i})\right]\right) \tag{6.7}$$

Where $\mathbf{S}_i$ is the embedding matrix of subgraph $\mathcal{S}_i$. Subsequently, we obtain subgraph graph representation vector $\mathbf{s}_i$ from subgraph matrix $\mathbf{S}^i$ by following the equation defined below.

$$\mathbf{s}_i = \left(\mathbf{Max}\left(\mathbf{S}_i\right) \oplus \mathbf{Min}\left(\mathbf{S}_i\right) \oplus \mathbf{Avg}\left(\mathbf{S}_i\right)\right) \tag{6.8}$$

Next, form an embedding matrix $\mathbf{S}$ by concatenating the embedding of subgraphs $\mathbf{S}_i \; \forall i \in [1,k]$. Where k is the number of words in the headline, accordingly, there will be k subgraphs. Next, we pass headline representation vector $\mathbf{h}$ as query and subgraph representation matrix $\mathbf{S}$ as key and value to multi-head attention for context matching between headline and subgraph. Then the query $\mathbf{h}^q$, key $\mathbf{S}^k$ and value $\mathbf{S}^v$ matrices for $\mathbf{a}^{th}$ attention head are defined as follows.

$$\mathbf{h}_a^q, \mathbf{S}_a^k, \mathbf{S}_a^v = \mathbf{h} \cdot \mathbf{W}_a^q, \mathbf{S} \cdot \mathbf{W}_c^a, \mathbf{P} \cdot \mathbf{W}_a^v \tag{6.9}$$

Where $\mathbf{W}_a^q$, $\mathbf{W}_a^k$ and $\mathbf{W}_a^v$ are learnable parameter matrices of query, key and value projections respectively, for $\mathbf{a}^{th}$ attention head of multi-head attention, and $\cdot$ represents the dot product between matrix. Subsequently, attention weight $\mathbf{d}_a$ for $\mathbf{a}^{th}$ attention head is defined as follows:

$$\mathbf{d}_a = \sigma\left(\frac{\mathbf{h}_a^q \left(\mathbf{S}_a^k\right)^\top}{\sqrt{\mathbf{z}}}\right) \tag{6.10}$$

Where $\mathbf{d}_a$ is weight vector of a^{th} attention head. $\mathbf{d_a}[\mathbf{i}]$ entry represents the similarity probability between $\mathbf{h}$ and subgraph representation of $\mathbf{i}^{th}$ subgraph representation matrix $\mathbf{S}$. $\mathbf{z}$ is the dimension of $\mathbf{S}_a^q$. Next, the weighted summation is applied over subgraph representation $\mathbf{s}_i$ based on similarity with headline representation $\mathbf{h}$.

$$\mathbf{m}_a = \left(\sum_{j=1, i \neq j}^{k} \mathbf{d}_{a,i}\, \mathbf{S}_{a,j}^v\right) \tag{6.11}$$

Where $\mathbf{m}_a$ is the representation of subgraph after weighted summation between $\mathbf{d}_a$, and subgraph representation matrix $\mathbf{S}_a^v$. Next, we concatenate the representation $\mathbf{m}_a$ obtained for each attention and pass through a dense layer to obtain a final representation $\mathbf{m}$.

$$\mathbf{m} = \Big(\mathbf{m}_1 \oplus \mathbf{m}_2 \oplus .. \oplus \mathbf{m}_i \oplus . \oplus \mathbf{m}_a\Big)\mathbf{W}_m \tag{6.12}$$

Where $\mathbf{m}$ is a final representation obtained based on the contextual similarity between the headline and subgraph.

6.6.1.3 Aggregation and Classification

Given the headline representation vector $\mathbf{h}$, body bigram network representation vector $\mathbf{b}$ and representation vector $\mathbf{m}$ obtained based on the contextual similarity between headline and subgraphs, we obtained a feature vector $\mathbf{f}$ by applying attention between $\mathbf{h}$, $\mathbf{b}$ and $\mathbf{m}$.

- **Scaled attention :** This approach forms a matrix $\mathbf{L}$ by concatenating $\mathbf{b}$ and $\mathbf{m}$. Then estimate a similarity between $\mathbf{h}$ and $\mathbf{L}$ as defined below.

$$\mathbf{v} = \Big(\frac{\mathbf{h}\,\mathbf{L}^\top}{\sqrt{\mathbf{z}}}\Big) \tag{6.13}$$

  Where $\mathbf{z}$ is the dimension of $\mathbf{h}$. Subsequently, we convert similarity vector $\mathbf{v}$ into probability as defined below.

$$\alpha = \Big(\frac{\exp(\mathbf{v_i})}{\sum_j \exp(\mathbf{v_j})}\Big) \tag{6.14}$$

  Next, we obtain a feature vector $\mathbf{r}$ by considering weighted summation between similarity probability vector α and $\mathbf{L}$ as defined below.

$$\mathbf{r} = \Big(\sum_j \alpha_i\,\mathbf{L}_i\Big) \tag{6.15}$$

  Subsequently, a final feature vector $\mathbf{f}$ is formed by concatenating $\mathbf{r}$ and $\mathbf{h}$.

$$\mathbf{f} = \mathbf{r} \oplus \mathbf{h} \tag{6.16}$$

- **Additive attention :** This approach is similar to the scaled attention-based approach. But this approach estimates similarity α_1 between $\mathbf{h}$ and $\mathbf{b}$ using a parametric form of similarity estimation, as defined below.

$$\alpha_1 = \mathbf{v}^\top\Big(\mathbf{W}_u\mathbf{h} + \mathbf{W}_v\mathbf{b}\Big) \tag{6.17}$$

Similarly, we also estimate similarity α_2 between $\mathbf{h}$ and $\mathbf{m}$ using a parametric form of similarity estimation, as defined below.

$$\alpha_2 = \mathbf{v}^\top \left(\mathbf{W}_u \mathbf{h} + \mathbf{W}_v \mathbf{m} \right) \tag{6.18}$$

Where, $\mathbf{v}$, $\mathbf{W}_u$ and $\mathbf{W}_v$ are learnable parameters. All other steps to form a feature vector $\mathbf{f}$ are similar to the scaled attention-based approach.

- **Similarity and difference features** This approach forms a feature vector $\mathbf{f}$ by estimating similarity and dissimilarity-based features. We first estimate the angle and difference between $\mathbf{h}$ and $\mathbf{b}$ using equations 6.19 as defined below.

$$\mathbf{a}^+ , \mathbf{a}^- = \mathbf{h} \odot \mathbf{b} , \mathbf{h} - \mathbf{b} \tag{6.19}$$

Next, we estimate the angle and difference vector between $\mathbf{h}$ and $\mathbf{b}$ using equations 6.20.

$$\mathbf{b}^+ , \mathbf{b}^- = \mathbf{h} \odot \mathbf{m} , \mathbf{h} - \mathbf{m} \tag{6.20}$$

Subsequently, we estimate the angle and difference vector between $\mathbf{b}$ and $\mathbf{m}$ using equations 6.21 as defined below.

$$\mathbf{c}^+ , \mathbf{c}^- = \mathbf{b} \odot \mathbf{m} , \mathbf{b} - \mathbf{m} \tag{6.21}$$

Finally, we form a feature vector $\mathbf{f}$ using equation 6.22 as defined below.

$$\mathbf{f} = \left(\mathbf{a}^+ \oplus \mathbf{a}^- \odot \mathbf{b}^+ \odot \mathbf{b}^- \odot \mathbf{c}^+ \odot \mathbf{c}^- \odot \mathbf{h} \odot \mathbf{m} \odot \mathbf{b} \right) \tag{6.22}$$

Once we obtain a feature vector $\mathbf{f}$ using any of the above-discussed methods, we pass the feature vector $\mathbf{f}$ to the two layers of a fully connected neural network, followed by softmax for incongruent news classification.

6.6.2 Graph-based Dual Context Matching GDCM

Figure 6.6 presents a working diagram of the proposed model, *Graph-based Dual Context Matching* **GDCM**. Given a news article with a pair of headline $\mathcal{H}$ and news body $\mathcal{B}$, the **GDCM** model first obtains a separate bigram network $\mathcal{H}_n$ and $\mathcal{B}_n$ by considering unigram as node and edge between nodes if two-word co-occur, respectively. Figure 6.4 presents (left) the bigram network of news body shown in Figure 6.3. The proposed model **GDCM** adopts two approaches for dual context extraction and matching: *Max-Min similarity* and

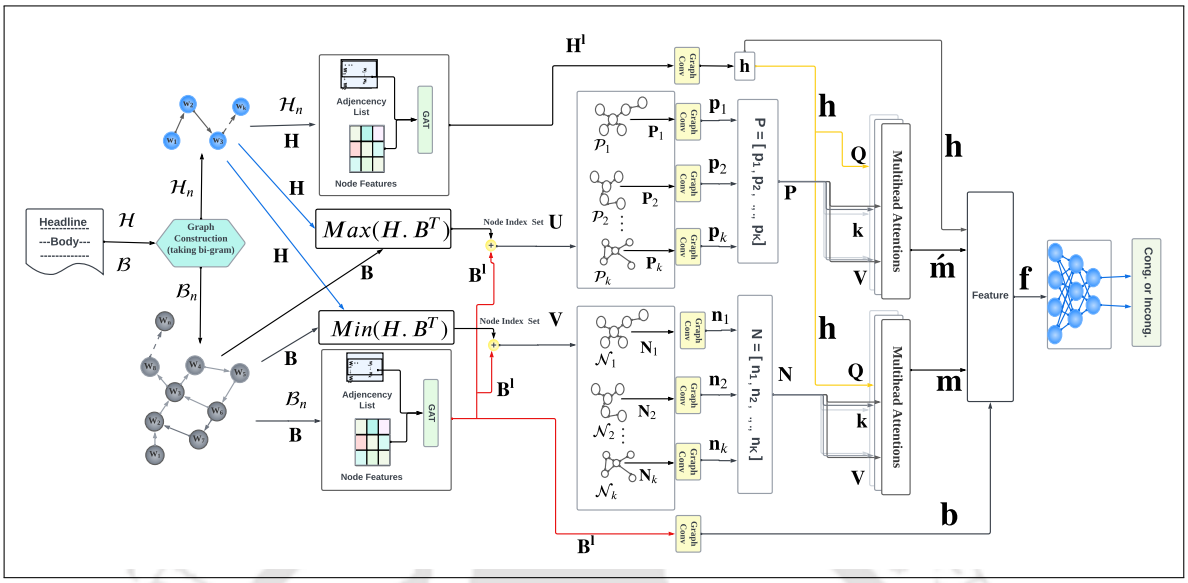


Fig. 6.6 The Working of proposed model **GDCM** models is presented in the diagram.

Graph complement. Both the approaches (*Max-Min similarity* and *Graph complement*) aim to extract positive context and negative context of headlines from news bodies, but they are different only in their approach. The primary motivation for extracting dual contexts (positive and negative) of headlines from various segments of the news body lies in their ability to differentiate between segments that align or deviate from the headline. Segments congruent with the headline are regarded as the positive context of the headline, whereas segments incongruent with the headline are considered the negative context. The *Max-Min similarity* extracts information (nodes in the body bigram network) from the news body which is highly similar to the headline as the positive context of the headline and information (nodes in the body bigram network) from the news body which is least similar to the headline as the negative context of the headline. Whereas *Graph complement* approach extracts information (nodes in the body bigram network) from the news body which is highly similar to the headline as the positive context, but information (nodes in the body bigram network) which is present in the body bigram network and not present in the positive context is considered negative context. Intuitively, if any information of the news body is not present in a positive context of the headline, then it is considered as a negative context of a headline in case of *Graph complement* approach. Whereas *Max-Min similarity* separately extracts the negative context of the headline from the news body along with the positive context.

6.6.2.1 Headline Centric Dual Context Extraction

Headline-centric dual context extraction aims to extract the positive and negative context for each word present in the headline from bigram network $\mathcal{B}_n$ of news body $\mathcal{B}$. The paragraphs where discussion regarding the headline is presented are called the positive context of the headline, and paragraphs where discussion regarding the headline is not present, are called the negative context of the headline. Positive context and negative context in this study are headline-centric because we extract positive and negative context from different segments of the news body based on the similarity between words in the headline and news body. The key institution behind extracting both positive and negative context is that in the case of congruent news, article headlines should be most similar to both positive and negative contexts extracted from the news body. Similarly, in the case of incongruent news, the headline will be least similar to both the positive and negative context extracted from the news body. Whereas in the case of partially incongruent news, headlines should be most similar to the positive context but is not similar to the negative context extracted from the news body. We obtain a matching matrix, **M**, to extract the headline-centric positive and negative context from the news body by following equation 6.23.

$$\mathbf{M} = \mathbf{H}^{\mathbf{T}} \cdot \mathbf{B} \tag{6.23}$$

Where **H** is the node embedding matrix of the headline bigram network $\mathcal{H}_n$ and **B** is the node embedding matrix of the body bigram network $\mathcal{B}_n$. As discussed above, **GDCM** adopts two approaches for dual context extraction and matching : *Max-Min similarity* and *Graph complement*.

Max-Min similarity approach : The *Max-Min similarity* based approach obtains a positive context node index set **U** by considering the column index of the row-wise maximum of the matching matrix **M** (as defined in equation 6.23). Similarly, also obtains a negative context node index set **V** by considering the column index of a row-wise minimum of the matching matrix **M** (as defined in equation 6.23). Intuitively, nodes from the body bigram network, which are highly similar to nodes in the headline bigram network, are considered as positive context nodes index and added to the positive context node index set **U**. Similarly, nodes from the body bigram network, which are least similar to nodes in the headline bigram network, are considered negative context nodes index and added to the negative context node index set **V**. The above *Max-Min similarity* based approach is also called as *dual similarity* because, for every node or word in the headline, the *Max-Min similarity* approach considers both the most and least similar node from the body bigram network as positive and negative

context node index respectively. Consequently, it will help to extract dual context (positive and negative context) for every word in the headline from different segments of the news body.

Graph complement approach : The *Graph complement* based approach obtains a positive context node index set **U** by considering the column index of the row-wise maximum of the matching matrix **M** (as defined in equation 6.23). The negative context node index set **V** is obtained by the equation defined below.

$$\mathbf{V} = \mathbf{G} \setminus \mathbf{U} \tag{6.24}$$

Where **G** is a set of nodes present in the body bigram network $\mathcal{B}_n$, **U** is a set of nodes in positive context node index and $\setminus$ denotes set difference operation. Intuitively, in the case of the *Graph complement* approach, all the nodes of the body bigram network $\mathcal{B}_n$ which are not present in the positive context node index set **U** are considered as negative context node index and accordingly added to the negative node index set **V**.

6.6.2.2 Global Representation of Headline and Body Bigram Network

Given node embedding matrix **H** and **B** of headline bigram network $\mathcal{H}_n$ and body bigram network $\mathcal{B}_n$, respectively, we apply **t** layer graph attention network **GAT** [93] to update the node's embedding of each node in $\mathcal{H}_n$ and $\mathcal{B}_n$ based on local neighbourhood structure of nodes and contexts of nodes in their neighbour. Our graph attention network **GAT** implementation setting is similar to study [93]. The attention coefficient $\alpha^l_{j,r}$ to nodes in neighborhood of $\mathbf{j}^{th}$ in body bigram network $\mathcal{B}_n$ is computed as follows:

$$\alpha^l_{j,r} = \left(\frac{\exp(\mathbf{e}^l_{j,r})}{\sum_r \exp(\mathbf{e}^l_{j,r})}\right) \tag{6.25}$$

Where $\mathbf{e}^l_{j,r}$ is estimated using the equation defined below.

$$\mathbf{e}^l_{j,r} = \mathbf{LeakyReLU}\left(\mathbf{a}^{\mathbf{T}} | [\mathbf{W}\mathbf{b}_j || \mathbf{W}\mathbf{b}_r]\right) \tag{6.26}$$

Where $\alpha^l_{j,r}$ denotes the importance of the node j to r. One convolution layer in **GAT** [93] involves the transformation from $\mathbf{B}^l$ to $\mathbf{B}^{l+1}$. The equation involved is as follows:

$$\mathbf{b}^{l+1}_j = \blacksquare_{r \in N(j)} \alpha_{j,r} \mathbf{W}^l \mathbf{b}^l_r \tag{6.27}$$

Similarly, we also obtain a transformed embedding matrix $\mathbf{H}^l$ by applying *GAT* over embedding matrix $\mathbf{H}$ of $\mathcal{H}_n$ by following equations 6.25, 6.26 and 6.27. Next, we obtain the headline bigram network $\mathcal{H}_n$ representation feature vector $\mathbf{h}$ and body bigram network $\mathcal{B}_n$ representation feature vector $\mathbf{b}$ by concatenating the min, max and average pulling over $\mathbf{H^l}$ and $\mathbf{B^l}$ as defined below.

$$\mathbf{h} = \Big(\mathbf{Max}\big(\mathbf{H}^l\big) \oplus \mathbf{Min}\big(\mathbf{H}^l\big) \oplus \mathbf{Avg}\big(\mathbf{H}^l\big)\Big) \tag{6.28}$$

$$\mathbf{b} = \Big(\mathbf{Max}\big(\mathbf{B}^l\big) \oplus \mathbf{Min}\big(\mathbf{B}^l\big) \oplus \mathbf{Avg}\big(\mathbf{B}^l\big)\Big) \tag{6.29}$$

Where Min, Max and Avg. indicate min pooling, max pooling and average pooling operation.

6.6.2.3 Dual Context Matching Between Headline and Body

Given a transformed word embedding matrix $\mathbf{H}^l$, $\mathbf{B}^l$ along with positive context node index set $\mathbf{U}$ and negative context node index set $\mathbf{V}$, we apply dual context matching between headline and body to verify the consistency between headline and body. We first construct a subgraph $\mathcal{P}_i$ for each node $\mathbf{u}_i$ in the positive context node index set $\mathbf{U}$ by sampling their $\mathbf{x}$ hop neighbour (*Radius*) in the body bigram network $\mathcal{B}_n$. Next, considering $\mathbf{N}\big(\mathbf{i}\big)$ as the list of node index of nodes in $\mathcal{P}_i$ except the root node $\mathbf{u}_i$, we obtain subgraph embedding matrix $\mathbf{P}_i$ by extracting embedding of nodes in $\mathbf{N}\big(\mathbf{i}\big)$ from bigram body embedding $\mathbf{B}^l$ as defined below.

$$\mathbf{P}_i = \Big(\mathbf{B}^l\big[\mathbf{u}(\mathbf{i})\big] \oplus \mathbf{B}^l\big[\mathbf{N}(\mathbf{i})\big]\Big) \tag{6.30}$$

Where $\mathbf{P}_i$ is the embedding matrix of subgraph $\mathcal{P}_i$. Similarly, we also construct a subgraph $\mathcal{N}_i$ for each node $\mathbf{v}_i$ in the negative context node index set $\mathbf{V}$ by sampling their $\mathbf{x}$ hop neighbour (*Radius*) in the body bigram network $\mathcal{B}_n$. Now, considering $\mathbf{N}\big(\mathbf{i}\big)$ as the list of the node index of the nodes in $\mathcal{N}_i$ except node the root node $\mathbf{v}_i$, we obtain subgraph embedding matrix $\mathbf{N}_i$ by extracting embedding of nodes in $\mathbf{N}\big(\mathbf{i}\big)$ from the embedding matrix of body bigram network $\mathbf{B}^l$ as defined below.

$$\mathbf{N}_i = \Big(\mathbf{B}^l\big[\mathbf{v}(\mathbf{i})\big] \oplus \mathbf{B}^l\big[\mathbf{N}(\mathbf{i})\big]\Big) \tag{6.31}$$

Where $\mathbf{N}_i$ is the embedding matrix of subgraph $\mathcal{N}_i$. Subsequently, we obtain subgraph representation vector $\mathbf{p}_i$ and $\mathbf{n}_i$ by following equation 6.32 and 6.33.

$$\mathbf{p}_i = \Big(\mathbf{Max}\big(\mathbf{P}_i\big) \oplus \mathbf{Min}\big(\mathbf{P}_i\big) \oplus \mathbf{Avg}\big(\mathbf{P}_i\big)\Big) \tag{6.32}$$

$$\mathbf{n}_i = \Big(\mathbf{Max}\big(\mathbf{N}_i\big) \oplus \mathbf{Min}\big(\mathbf{N}_i\big) \oplus \mathbf{Avg}\big(\mathbf{N}_i\big)\Big) \tag{6.33}$$

Next, we form a positive context embedding matrix $\mathbf{P}$ by concatenating the embedding of subgraph $\mathbf{P}_i \; \forall i \in [1,k]$. Similarly, we also form a negative context embedding matrix $\mathbf{N}$ by concatenating the embedding of subgraph $\mathbf{N}_i \; \forall i \in [1,k]$ respectively. Where $\mathbf{k}$ is the number of words in the headline, and there will be a k subgraph for both positive and negative contexts. Subsequently, we apply multi-head attention [58] between headline representation $\mathbf{h}$ and positive context representation embedding matrix $\mathbf{P}$ to obtain a representation vector $\mathbf{p}$. The main motivation behind applying multi-head attention between headline and positive context representation is that in the case of congruent news or partially incongruent news, the similarity between headline and representation of positive context should be high because positive context is extracted from paragraphs of the news body where discussion regarding the headline present, so in the case of congruent news or partially congruent news similarity between headline and representation positive context set should be high. Whereas in the case of incongruent news, the similarity between the headline and representation of a positive context set should be the least because the headline and news body do not align with each other. Accordingly, the context of the headline and news body should be different. Next, we also obtain a representation vector $\mathbf{n}$ by applying multi-head attention between headline representation $\mathbf{h}$ and negative context representation embedding matrix $\mathbf{N}$. The main motivation behind applying multi-head attention between headline and negative context representation is that in the case of incongruent news or partially incongruent news, the dissimilarity between headline and representation of negative context should be high because negative context is extracted from paragraphs of the news body where discussion regarding the headline is not present, so in the case of incongruent news or partially congruent news dissimilarity between headline and representation of negative context set should be high. Whereas in the case of congruent news, the similarity between the headline and representation of a negative context set should be high because the headline and news body are contextually similar and align with each other accordingly the context of the headline and news body should be the same. With the above intuitions, we pass headline representation vector $\mathbf{h}$ as query and positive context representation matrix $\mathbf{P}$ as key and value in multi-head attention.

The query $\mathbf{h}^q$, key $\mathbf{P}^k$ and value $\mathbf{P}^v$ matrices for $\mathbf{a}^{th}$ attention head are defined as follows.

$$\mathbf{h}_a^q, \mathbf{P}_a^k, \mathbf{P}_a^v = \mathbf{h} \cdot \mathbf{W}_a^q, \mathbf{P} \cdot \mathbf{W}_a^k, \mathbf{P} \cdot \mathbf{W}_a^v \tag{6.34}$$

Where $\mathbf{W}_a^q$, $\mathbf{W}_a^k$ and $\mathbf{W}_a^v$ are learnable parameter matrices of query, key and value projections respectively, for $\mathbf{a}^{th}$ attention head of multi-head attention and $\cdot$ is the dot product between matrix. Subsequently, attention weigh $\mathbf{d}_a$ for $\mathbf{a}^{th}$ attention head defined as follows:

$$\mathbf{d}_a = \sigma\left(\frac{\mathbf{h}_a^q \left(\mathbf{P}_a^k\right)^\top}{\sqrt{\mathbf{z}}}\right) \tag{6.35}$$

Where $\mathbf{d}_a$ is a weight vector of a^{th} attention head. $\mathbf{d_a}[\mathbf{i}]$ entry represents the similarity probability between $\mathbf{h}$ and subgraph representation of $\mathbf{i}^{th}$ subgraph in positive context representation matrix $\mathbf{P}$. $\mathbf{z}$ is the dimension of $\mathbf{P}_a^q$. Next, the weighted summation is applied over positive context subgraph representation $\mathbf{p}_i$ based on similarity with headline representation $\mathbf{h}$.

$$\mathbf{m}_a = \left(\sum_{j=1, i \neq j}^{k} \mathbf{d}_{a,i} \mathbf{P}_{a,j}^v\right) \tag{6.36}$$

Where $\mathbf{m}_a$ is the representation of positive context in $\mathbf{a}^{th}$ the attention head after weighted summation between $\mathbf{d}_a$ and positive context subgraph representation matrix $\mathbf{P}_a^v$. Next, we concatenate the representation $\mathbf{m}_a$ obtained for each attention and pass through a dense layer to obtain a final representation $\mathbf{m}$ as defined in the equation below.

$$\acute{\mathbf{m}} = \left(\mathbf{m}_1 \oplus \mathbf{m}_2 \oplus \ldots \oplus \mathbf{m}_i \oplus . \oplus \mathbf{m}_a\right) \mathbf{W}_m \tag{6.37}$$

Where $\acute{\mathbf{m}}$ is a final representation obtained based on the headline and positive context subgraph similarity. The negative context representation vector $\mathbf{m}$ is obtained by applying different methods for *Max-Min similarity* and *Graph complement*.

Max-Min similarity approach : Given negative context subgraph representation matrix $\mathbf{N}$ and headline subgraph representation $\mathbf{h}$, we obtain a negative context representation vector $\mathbf{m}$ based on the similarity between headline and negative context subgraph by the following equation 6.34, 6.35 and 6.37 but equation 6.36 is replaced by the equation 6.38 as defined below.

$$\mathbf{m}_a = \left(\sum_{j=1, i \neq j}^{k} \left(1 - \mathbf{d}_{a,i}\right) \mathbf{P}_{a,j}^v\right) \tag{6.38}$$

The key motivation behind considering $(1 - \mathbf{d}_{a,i})$ in equation 6.38 unlike in equation 6.36 is that we want to give higher weight to the representation vector $\mathbf{n}_i$ of subgraph $\mathcal{N}_i$ which is the least similar to the representation oh headline. Whereas in the equation 6.38 we are giving higher weight to the representation vector $\mathbf{p}_i$ of subgraph $\mathcal{P}_i$ which is the highly similar to the representation of headline. The primary motivation behind the contradiction between equation 6.36 and equation 6.38 is that if a positive context representation $\acute{\mathbf{m}}$ is obtained by giving high attention weight to a highly contextual similar positive context subgraph $\mathcal{P}_i$ (positive context subgraph representation $\mathbf{p}_i$ which is highly similar to headline representation $\mathbf{h}$) and a negative context representation $\mathbf{m}$ is obtained by giving high attention weight to the least contextual similar negative context subgraph $\mathcal{N}_i$ (negative context subgraph representation $\mathbf{n}_i$ which is least similar to headline representation $\mathbf{h}$) are similar to headline representation $\mathbf{h}$. Then the news article is congruent, otherwise incongruent.

Graph complement approach : In this approach, the negative context representation vector $\mathbf{m}$ is obtained by the following equation 6.39 and 6.40.

$$\mathbf{C} = \left(\mathbf{B}^{l}\left[\mathbf{V}\right]\right) \tag{6.39}$$

Where $\mathbf{C}$ is the embedding matrix of nodes $\mathbf{V}$. Here, negative context nodes index set $\mathbf{V}$ is obtained by equation 6.24.

$$\mathbf{m} = \left(\mathbf{Max}\left(\mathbf{C}\right) \oplus \mathbf{Min}\left(\mathbf{C}\right) \oplus \mathbf{Avg}\left(\mathbf{C}\right)\right) \tag{6.40}$$

Intuitively, in the case of *Graph complement* approach, we form an embedding matrix $\mathbf{C}$ (in equation 6.39) by concatenating node embeddings of nodes in $\mathbf{V}$ (obtained using equation 6.24) and then form a negative context representation feature vector $\mathbf{m}$ by applying maximum, minimum and average pooling over embedding matrix $\mathbf{C}$ (as defined in equation 6.40).

6.6.2.4 Aggregation and Classification

Once we obtain the headline bigram network representation vector $\mathbf{h}$, body bigram network representation vector $\mathbf{b}$, positive context representation vector $\acute{\mathbf{m}}$ and negative context representation vector $\mathbf{m}$, next we form a feature vector $\mathbf{f}$ by applying attention between $\mathbf{h}$, $\mathbf{b}$, $\mathbf{m}$ and $\acute{\mathbf{m}}$. The motivations behind estimating feature vector $\mathbf{f}$ by applying attention between $\mathbf{h}$, $\mathbf{b}$, $\acute{\mathbf{m}}$ and $\mathbf{m}$ is that: (i) In the case of congruent news, the context of the headline will be congruent to the context or discussion regarding the headline in different segments of news

body. Consequently, the headline representation **h** must be highly similar to representations of positive context subgraph (context obtained from different subgraph) **ḿ** and representations of negative context subgraph **m** and global representation of news body **b**. (ii) In the case of fully incongruent news cases, content in the headline and news body will not be contextually similar. Consequently, the headline representation **h** will be least similar to the representation of positive context **ḿ**, representations of negative context **m** and global representation of body bigram network **b**. (iii) In the case of partially incongruent news, a few paragraphs of the news body will be congruent with the headline, but a few paragraphs will not be congruent with the headline. Accordingly, in the case of **GDCM** the model, the positive context representation vector **ḿ** represents the positive context in different segments of the news body, which are congruent with the headline. Similarly, a negative context representation vector **m** represents the context in different segments of the news body which are incongruent with the headline. Consequently, the headline representation **h** may be similar to representations of positive context representation **ḿ** but the headline representation **h** will be least similar to representations of negative context **ḿ** and global representation of news body **b**. With the above-discussed intuitions, we adopt three approaches: *Scaled attention*, *Additive attention* and *Similarity feature-based* approach to estimate feature vector **f**.

- **Scaled attention approach :** This approach forms a matrix **E** by concatenating **b**, **ḿ** and **m**. Then estimate a similarity between **h** and **E** as defined below.

$$\mathbf{v} = \left(\frac{\mathbf{h}\,\mathbf{E}^{\top}}{\sqrt{\mathbf{z}}}\right) \tag{6.41}$$

  Where **z** is the dimension of **h**. Subsequently, we convert similarity vector **v** into probability, as defined below.

$$\alpha = \left(\frac{\exp(\mathbf{v_i})}{\sum_j \exp(\mathbf{v_j})}\right) \tag{6.42}$$

  Next, we obtain a feature vector **r** by considering weighted summation between similarity probability vector α and **E** as defined below.

$$\mathbf{r} = \left(\sum_j \alpha_i\,\mathbf{E}_i\right) \tag{6.43}$$

  Subsequently, a final feature vector **f** is formed by concatenating **r** and **h**.

$$\mathbf{f} = \mathbf{r} \oplus \mathbf{h} \tag{6.44}$$

  Where $\oplus$ denotes vector concatenation operations.

- **Additive attention approach :** This approach is similar to the scaled attention approach discussed above, but it estimates similarity α_1 between $\mathbf{h}$ and $\mathbf{b}$ using a parametric form of similarity estimation, as defined below.

$$\alpha_1 = \mathbf{v}^\top \left(\mathbf{W}_u \mathbf{h} + \mathbf{W}_v \mathbf{b} \right) \tag{6.45}$$

Similarly, we also estimate similarity α_2 between $\mathbf{h}$ and $\acute{\mathbf{m}}$ and α_3 between $\mathbf{h}$ and $\mathbf{m}$ using a parametric form of similarity estimation, as defined in equation 6.46 and 6.47.

$$\alpha_2 = \mathbf{v}^\top \left(\mathbf{W}_u \mathbf{h} + \mathbf{W}_v \acute{\mathbf{m}} \right) \tag{6.46}$$

$$\alpha_3 = \mathbf{v}^\top \left(\mathbf{W}_u \mathbf{h} + \mathbf{W}_v \mathbf{m} \right) \tag{6.47}$$

Where, $\mathbf{v}$, $\mathbf{W}_u$ and $\mathbf{W}_v$ are learnable parameters. All other steps to form a feature vector $\mathbf{f}$ are similar to the scaled attention-based approach. Next, we obtain a feature vector $\mathbf{r}$ by considering weighted summation as defined below.

$$\mathbf{r} = \left(\alpha_1 \mathbf{b} + \alpha_2 \acute{\mathbf{m}} + \alpha_3 \mathbf{m} \right) \tag{6.48}$$

Subsequently, a final feature vector $\mathbf{f}$ is formed by concatenating $\mathbf{r}$ and $\mathbf{h}$.

$$\mathbf{f} = \mathbf{r} \oplus \mathbf{h} \tag{6.49}$$

- **Similarity and difference features :** The *Scaled attention* and *Additive attention* approach only verify consistency and similarity between headline and news body by forming a feature vector $\mathbf{f}$ based on the similarity between headline and positive context, between headline and negative context and between headline and body bigram network representation, but does not verify the consistency or similarity within different segments of the news body. To verify the consistency or similarity within different segments of the news body, the *Similarity and difference features* approach also estimates similarity and dissimilarity between positive context representation $\acute{\mathbf{m}}$ and negative context representation $\mathbf{m}$. The prime intuition behind estimating similarity between positive context representation $\acute{\mathbf{m}}$ and negative context representation $\mathbf{m}$ is that in the case of congruent news, the different paragraphs of news body will be contextually similar to each other, accordingly similarity between positive context representation $\acute{\mathbf{m}}$ and negative context representation $\mathbf{m}$ should be high. Whereas, in the case of a partially incongruent news article, paragraphs which are congruent to the headline and paragraphs which are not congruent to the headline will not be contextually similar to each other accordingly positive context representation $\acute{\mathbf{m}}$ will not be similar to negative context representation $\mathbf{m}$. This approach forms a feature vector $\mathbf{f}$ by estimating similarity and dissimilarity-based features. We estimate the angle and difference between $\mathbf{h}$ and $\mathbf{b}$ using equations 6.50 to form feature vectors $\mathbf{a}^+$ and $\mathbf{a}^-$ based on similarity and dissimilarity between headline representation $\mathbf{h}$ and

global representation of news body **b**.

$$\mathbf{a}^{+}, \mathbf{a}^{-} = \mathbf{h} \odot \mathbf{b}, \mathbf{h} - \mathbf{b} \tag{6.50}$$

Next, we estimate the angle and difference vector between headline representation **h** and negative context representation **m** using equations 6.51 to form feature vectors $\mathbf{b}^{+}$ and $\mathbf{b}^{-}$ based on similarity and dissimilarity between headline representation **h** and negative context representation **m**.

$$\mathbf{b}^{+}, \mathbf{b}^{-} = \mathbf{h} \odot \mathbf{m}, \mathbf{h} - \mathbf{m} \tag{6.51}$$

Subsequently, we estimate the angle and difference vector between headline representation **h** and positive context representation $\acute{\mathbf{m}}$ using equation 6.52 to form feature vectors $\mathbf{c}^{+}$ and $\mathbf{c}^{-}$ based on similarity and dissimilarity between headline representation **h** and positive context representation $\acute{\mathbf{m}}$.

$$\mathbf{c}^{+}, \mathbf{c}^{-} = \mathbf{b} \odot \acute{\mathbf{m}}, \mathbf{b} - \acute{\mathbf{m}} \tag{6.52}$$

As discussed above, we also estimate the angle and difference vector between positive context representation $\acute{\mathbf{m}}$ and negative context representation **m** using equation 6.53 to form feature vectors $\mathbf{d}^{+}$ and $\mathbf{d}^{-}$ based on similarity and dissimilarity between negative context representation **m** and positive context representation $\acute{\mathbf{m}}$.

$$\mathbf{d}^{+}, \mathbf{d}^{-} = \mathbf{m} \odot \acute{\mathbf{m}}, \mathbf{m} - \acute{\mathbf{m}} \tag{6.53}$$

Finally, we form a feature vector **f** using equation 6.54 as defined below.

$$\mathbf{f} = \left(\mathbf{a}^{+} \oplus \mathbf{a}^{-} \oplus \mathbf{b}^{+} \oplus \mathbf{b}^{-} \oplus \mathbf{c}^{+} \oplus \mathbf{c}^{-} \oplus \mathbf{h} \oplus \mathbf{m} \oplus \mathbf{b} \oplus \mathbf{d}^{+} \oplus \mathbf{d}^{-} \oplus \acute{\mathbf{m}}\right) \tag{6.54}$$

Once we obtain a feature vector **f** using any of the above-discussed methods, we pass the feature vector **f** to the two layers of a fully connected neural network, followed by softmax for incongruent news classification.

6.7 Evaluation Methodology

In this section, we describe the experiment settings for accessing the performance of the candidate methods. We also provide reproducible information for the shown results and analyses.

6.8 Experimental Setup

6.8.1 Baselines.

We consider sixteen state-of-art methods from the literature as a baseline to compare and study the effectiveness of our proposed model **GDCM**. These sixteen baseline models can be further grouped into *features*, *encoding*, *hierarchical*, *deep hierarchical* and *summarization* based on whether a model uses bag-of-words based feature, sequential encoding of news headline and body or exploit summarization methods. This study considers *FNC* [22], *UCLMR* [35] and *stackLSTM* [23] from literature as bag-of-words based baseline. Similarly, this study also considers hierarchical encoding-based models *AHDE* [28], *HDSF* [36], *GHDE* [1] and deep hierarchical encoding-based models *HoBERT* [39], *HeLSTM* [39], $GraSHE^{(=)}$ [39] and *RaSHE* [39] from literature as baseline. We further consider summarization-based model *FEDS* [38, 31] and dual summarization-based model *MADS* [21] from literature as a baseline. In addition to the above state-of-the-art baseline models from the literature, we also consider bidirectional encoder representations from transformers (BERT) [80] encoding-based baseline model *BERT* and *Recurrence of BERT* (*RoBERT*) as a baseline to study the effectiveness of our proposed model.

- `BERT` [80]: Considering the encouraging observations of BERT (for various NLP tasks) in recent studies, we also built a classifier using BERT [80] encoding base baseline models. First, we obtain an encoded representation of the headline and news body. Next, estimate the angle and difference between the encoded representation of the headline and the news body. Subsequently, we concatenate the encoded representation of the headline and news body along with angle and difference and pass it to a two-layer fully connected neural network followed by softmax for classification.

- `Recurrence over BERT (RoBERT)`: As reported in the study [125], hierarchical transformer and recurrence over BERT models are more effective in large document classifications. Motivated by the above observations, we also build recurrence over BERT *RoBERT* models, which first split the news body into a set of sentences and then obtain an encoding of each sentence within the news body and headline. Next, apply BiLSTM over the encoded representation of sentences to obtain an encoded representation of the news body. Subsequently, we form a feature by concatenating the encoded representation of the headline and body and the angle and difference between them. Then, we pass it to a two-layer fully connected neural network followed by softmax for classifications. Further, the details of *RoBERT* can be studied in [125].

Table 6.1 Details of hyperparameters used to produce results.

Hyperparameters	**Values**
Epoch	100
Batch Size	50
Word Embedding Dimension	300
Learning Rate	0.01
Loss Function	Cross Entropy
Number of Layer in MLP	2
number of words in headline	12
# attention head	1,2,3,4,6,8
n-gram values	2,3,4
# hope neighbour in subgraph	1,2,3
# Number GAT convolution layer	3

6.8.2 Experimental Setups

We consider F-measure *(F.)*, Accuracy (*Acc.*) and class-wise *F-measure* as evaluation metrics for evaluating our proposed and baseline models. Table 6.1 presents the details of hyperparameters that have been used to produce the results presented in this study. Though we have experimented with different values of hyperparameters, Table 6.1 presents the value which gives the best accuracy over the development split of the dataset. We consider a maximum of twelve words in the headline. Accordingly, we get twelve subgraph in both the positive context set and the negative context set.

6.9 Results and Discussions

Table 6.3 presents the performance of proposed and baseline models over three benchmark datasets. As evident in Table 6.3, the baseline methods are grouped into four; *feature*, *Encoding*, *hierarchical* and *summarization*. First, we study the response of feature-based baseline models, which mainly rely on bag-of-words-based features. Table 6.3 shows that *stackLSTM* outperforms *FNC* and *UCLMR* over NELA and FNC. The possible reason behind the superior performance of *stackLSTM* is that *stackLSTM* uses both bag-of-word-based features and sequential encoding, whereas *FNC* and *UCLMR* solely relies on bag-of-word-based

Table 6.2 Comparison of the performances of different models over three benchmark datasets. Here, (Acc) and (F) indicate accuracy and F-measure, respectively. Similarly, (Cong.) and (Incong.) indicate F-measure of congruent and incongruent class, respectively. Color indicates the best performance across models over a dataset.

		Models	NELA-17				ISOT				FNC			
			Acc	F	Cong.	Incong.	Acc	F	Cong.	Incong.	Acc	F	Cong.	Incong.
Baseline	Feat.	**FNC** [22]	0.586	0.586	0.564	0.608	0.844	0.844	0.847	0.842	0.586	0.496	0.282	0.709
		UCLMR [35]	0.589	0.588	**0.608**	0.569	**0.997**	**0.997**	**0.997**	**0.997**	0.964	0.955	0.934	0.975
		StackLSTM [98]	**0.597**	**0.591**	0.541	**0.641**	0.992	0.992	0.992	0.992	**0.971**	**0.963**	**0.946**	**0.982**
	Enc	**BiLSTM**	0.555	0.55	0.563	0.547	0.99	0.99	0.99	0.99	0.616	0.504	0.269	0.74
		BERT	0.572	0.563	**0.624**	0.503	0.894	0.894	0.894	0.891	0.722	0.419	0.21	0.838
		RoBERT	**0.615**	**0.613**	0.54	**0.642**	**0.996**	**0.996**	**0.996**	**0.996**	0.828	0.755	0.622	0.888
	Hierarchical	**AHDE** [28]	0.606	0.606	0.614	0.598	0.913	0.913	0.909	0.909	0.691	0.454	0.094	0.814
		HDSF [36]	0.517	0.494	0.602	0.386	0.720	0.712	0.665	0.759	0.758	0.666	0.492	0.841
		GHDE [1]	0.55	0.331	0.331	0.332	-	-	-	-	-	-	-	-
		HoBERT [39]	0.635	0.634	0.65	0.618	0.991	0.991	0.991	0.991	0.861	0.823	0.741	0.905
		HeLSTM [39]	0.602	0.602	0.607	0.598	0.997	0.997	0.997	0.997	0.809	0.764	0.661	0.867
		GraSHE$^{((=)}$ [39]	**0.70**	**0.699**	**0.71**	**0.688**	**0.999**	**0.999**	**0.999**	**0.999**	0.842	0.804	0.718	0.89
		RaSHE [39]	0.677	0.677	0.678	0.679	**0.999**	**0.999**	**0.999**	**0.999**	**0.876**	**0.844**	**0.775**	**0.914**
	Summ.	**FEDS** [38, 31]	0.533	0.532	0.550	0.515	**0.998**	**0.998**	**0.998**	**0.998**	0.878	0.837	0.755	0.918
		MADS(BiLSTM) [21]	**0.641**	**0.640**	**0.652**	**0.629**	0.998	0.998	0.998	0.998	0.969	0.960	0.942	0.978
		MADS(S-BERT) [21]	0.63	0.628	0.603	0.654	0.984	0.984	0.984	0.984	**0.971**	**0.963**	**0.947**	**0.98**
		DuSum(BiLSTM,θ) Chapter 5	0.668	0.668	0.673	0.664	**0.998**	**0.998**	**0.998**	**0.998**	0.986	0.982	0.974	0.990
		DuSum(ST,θ) Chapter 5	**0.701**	**0.701**	**0.703**	**0.7**	0.99	0.99	0.99	0.99	0.99	0.987	0.981	0.993
		DuSum(BiLSTM,k) 5	0.627	0.625	0.62	0.63	0.997	0.997	0.997	0.997	0.985	0.981	0.973	0.990
		DuSum(ST,k) Chapter 5	0.647	0.643	0.66	0.626	**0.998**	**0.998**	**0.998**	**0.998**	**0.991**	**0.988**	**0.983**	**0.993**
		DuSum(BiLSTM,θ) Chapter 5	0.668	0.668	0.673	0.664	**0.998**	**0.998**	**0.998**	**0.998**	0.986	0.982	0.974	0.990
		DuSum(ST,θ) Chapter 5	**0.701**	**0.701**	**0.703**	**0.7**	0.99	0.99	0.99	0.99	0.99	0.987	0.981	0.993
		DuSum(BiLSTM,k) Chapter 5	0.627	0.625	0.62	0.63	0.997	0.997	0.997	0.997	0.985	0.981	0.973	0.990
		DuSum(ST,k) Chapter 5	0.647	0.643	0.66	0.626	**0.998**	**0.998**	**0.998**	**0.998**	**0.991**	**0.988**	**0.983**	**0.993**
Proposed	Scaled	**GCM(AH = 1)**	0.921	0.921	0.92	0.922	0.977	0.977	0.977	0.977	0.958	0.947	0.924	0.971
		GCM(AH = 2)	0.926	0.926	0.925	0.927	0.995	0.995	0.995	0.995	0.966	0.957	0.938	0.977
		GCM(AH = 6)	**0.929**	**0.929**	**0.928**	**0.93**	**0.998**	**0.998**	**0.998**	**0.998**	**0.972**	**0.965**	**0.95**	**0.981**
	Additive	**GCM(AH = 1)**	0.902	0.902	0.902	0.902	0.997	0.997	0.997	0.997	0.866	0.822	0.746	0.909
		GCM(AH = 2)	**0.911**	**0.910**	**0.909**	**0.912**	**0.998**	**0.998**	**0.998**	**0.998**	0.908	0.884	0.832	0.936
		GCM(AH = 6)	0.902	0.902	0.901	0.904	**0.998**	**0.998**	**0.998**	**0.998**	**0.937**	**0.918**	**0.88**	**0.957**
	sim	**GCM(AH = 1)**	**0.924**	**0.924**	**0.923**	**0.925**	0.997	0.997	0.997	0.997	0.965	0.954	0.934	0.975
		GCM(AH = 2)	0.901	0.9	0.897	0.904	0.998	0.998	0.998	0.998	0.963	0.953	0.933	0.974
		GCM(AH = 6)	0.911	0.911	0.91	0.912	0.998	0.998	0.998	0.998	**0.972**	**0.965**	**0.95**	**0.981**

features. From the above observation, we can conclude that bag-of-word-based features alone are insufficient for the incongruent news detection task. Next, we study the performance of encoding-based baseline models. From Table 6.3, it is apparent that the BERT [80] encoding-based model *RoBERT* outperforms other encoding-based baseline models *BERT* and *BiLSTM* over FNC and NELA dataset. Similarly, the *BiLSTM* model outperforms *BERT* and *RoBERT* over ISOT dataset. From Table 3.1, it is evident that the NELA dataset has thirteen paragraphs on average, and FNC has eleven paragraphs on average, but the ISOT dataset has only three paragraphs. This indicates that news samples in NELA and FNC are large in the number of paragraphs, whereas news samples in the ISOT dataset are small. As concluded in the study [125] that *RoBERT* is superior to BERT [80] and BiLSTM [123] for large document classification. So, the possible reason behind the superior performance of *RoBERT* the model over NELA and FNC is that *RoBERT* the model is suitable for encoding news articles with large (high number of paragraphs) news bodies. Whereas *BiLSTM* and *BERT* are more suitable for encoding small news documents, accordingly performance of the *BiLSTM* model is superior to *RoBERT* over the ISOT dataset. Subsequently, we study the performance of hierarchical encoding-based baseline models, which can be further grouped into *hierarchical encoding* and *deep hierarchical encoding*. Table 6.3 presents the performances of hierarchical encoding-based baseline models *AHDE* [28], *HDSF* [36], *GHDE* [1], *HeLSTM* [39] and deep hierarchical encoding-based baseline models *RaSHE* [39], $GraSHE^{(=)}$ [39], *HoBERT* [39]. From Table 6.3 it is apparent that the $GraSHE^{(=)}$ [39] outperforms all other hierarchical and deep hierarchical based baseline models over NELA and ISOT datasets. Similarly, the *RaSHE* [39] outperforms all other hierarchical encoding and deep hierarchical encoding-based baseline models over FNC datasets. As discussed in subsection 4.6.1 NELA dataset represents the characteristics of partially incongruent news, and FNC datasets represent the characteristics of fully incongruent news. Both $GraSHE^{(=)}$ [39] and *RaSHE* [39] are deep hierarchical encoding-based models, and their working is also similar. The only difference is that $GraSHE^{(=)}$ [39] assigns unique weights to paragraphs while encoding news body based on its inability to represent another paragraph in the news body, but *RaSHE* [39] gives equal importance to all paragraphs while encoding news body. Assigning unique weight helps highlight the paragraphs incongruent with the news headline. Therefore, the performance of $GraSHE^{(=)}$ [39] is superior in partially incongruent news detection (NELA dataset). Similar observations regarding the superior performance of $GraSHE^{(=)}$ [39] are made in study [39].

Further, we study the performance of summarizations-based baseline models. From Table 6.3, it is visible that the baseline model *MADS* [21] outperforms *FEDS* [38, 31] over FNC and NELA datasets. The possible reason behind the superiority of *MADS* [21] over

Table 6.3 Comparison of the performances of different models over three benchmark datasets. Here, (Acc) and (F) indicate accuracy and F-measure, respectively. Similarly, (Cong.) and (Incong.) indicate F-measure of congruent and incongruent class, respectively. The colour indicates the best performance of the model over the respective dataset.

		Models	NELA-17				ISOT				FNC			
			Acc	F	Cong.	Incong.	Acc	F	Cong.	Incong.	Acc	F	Cong.	Incong.
Baseline	Feat.	**FNC** [22]	0.586	0.586	0.564	0.608	0.844	0.844	0.847	0.842	0.586	0.496	0.282	0.709
		UCLMR [35]	0.589	0.588	**0.608**	0.569	**0.997**	**0.997**	**0.997**	**0.997**	0.964	0.955	0.934	0.975
		StackLSTM [98]	**0.597**	**0.591**	0.541	**0.641**	0.992	0.992	0.992	0.992	**0.971**	**0.963**	**0.946**	**0.982**
	Enc	**BiLSTM**	0.555	0.55	0.563	0.547	0.99	0.99	0.99	0.99	0.616	0.504	0.269	0.74
		BERT	0.572	0.563	**0.624**	0.503	0.894	0.894	0.894	0.891	0.722	0.419	0.21	0.838
		RoBERT	**0.615**	**0.613**	0.54	**0.642**	**0.996**	**0.996**	**0.996**	**0.996**	**0.828**	**0.755**	**0.622**	**0.888**
	Hierarchical	**AHDE** [28]	0.606	0.606	0.614	0.598	0.913	0.913	0.909	0.909	0.691	0.454	0.094	0.814
		HDSF [36]	0.517	0.494	0.602	0.386	0.720	0.712	0.665	0.759	0.758	0.666	0.492	0.841
		GHDE [1]	0.55	0.331	0.331	0.332	-	-	-	-	-	-	-	-
		HoBERT [39]	0.635	0.634	0.65	0.618	0.991	0.991	0.991	0.991	0.861	0.823	0.741	0.905
		HeLSTM [39]	0.602	0.602	0.607	0.598	0.997	0.997	0.997	0.997	0.809	0.764	0.661	0.867
		GraSHE$^{(=)}$ [39]	**0.70**	**0.699**	**0.71**	**0.688**	**0.999**	**0.999**	**0.999**	**0.999**	0.842	0.804	0.718	0.89
		RaSHE [39]	0.677	0.677	0.678	0.676	**0.999**	**0.999**	**0.999**	**0.999**	**0.876**	**0.844**	**0.775**	**0.914**
	Summ.	**FEDS** [38, 31]	0.533	0.532	0.550	0.515	**0.998**	**0.998**	**0.998**	**0.998**	0.878	0.837	0.755	0.918
		MADS(BiLSTM) [21]	**0.641**	**0.640**	**0.652**	**0.629**	0.998	0.998	0.998	0.998	0.969	0.960	0.942	0.978
		MADS(S-BERT) [21]	0.63	0.628	0.603	0.654	0.984	0.984	0.984	0.984	**0.971**	**0.963**	**0.947**	**0.98**
		DuSum(BiLSTM,θ) Chapter 5	0.668	0.668	0.673	0.664	**0.998**	**0.998**	**0.998**	**0.998**	0.986	0.982	0.974	0.990
		DuSum(ST,θ) Chapter 5	**0.701**	**0.701**	**0.703**	**0.7**	0.99	0.99	0.99	0.99	0.99	0.987	0.981	0.993
		DuSum(BiLSTM,k) 5	0.627	0.625	0.62	0.63	0.997	0.997	0.997	0.997	0.985	0.981	0.973	0.990
		DuSum(ST,k) Chapter 5	0.647	0.643	0.66	0.626	**0.998**	**0.998**	**0.998**	**0.998**	**0.991**	**0.988**	**0.983**	**0.993**
		DuSum(BiLSTM,θ) Chapter 5	0.668	0.668	0.673	0.664	**0.998**	**0.998**	**0.998**	**0.998**	0.986	0.982	0.974	0.990
		DuSum(ST,θ) Chapter 5	**0.701**	**0.701**	**0.703**	**0.7**	0.99	0.99	0.99	0.99	0.99	0.987	0.981	0.993
		DuSum(BiLSTM,k) Chapter 5	0.627	0.625	0.62	0.63	0.997	0.997	0.997	0.997	0.985	0.981	0.973	0.990
		DuSum(ST,k) Chapter 5	0.647	0.643	0.66	0.626	**0.998**	**0.998**	**0.998**	**0.998**	**0.991**	**0.988**	**0.983**	**0.993**
Proposed	Max-Min	**GDCM(AH = 1)**	**0.905**	**0.905**	**0.901**	**0.909**	**0.998**	**0.998**	**0.998**	**0.998**	0.817	0.709	0.529	0.89
		GDCM(AH = 6)	0.850	0.850	0.851	0.849	0.998	0.998	0.998	0.998	0.955	0.944	0.918	0.969
		GDCM(AH = 8)	0.904	0.904	0.9	0.907	0.998	0.998	0.998	0.998	**0.963**	**0.953**	**0.932**	**0.974**
		GDCM(AH = 1)	0.88	0.881	0.88	0.881	**0.997**	**0.997**	**0.997**	**0.997**	0.86	0.825	0.746	0.903
		GDCM(AH = 6)	0.88	0.881	0.88	0.881	0.995	0.995	0.995	0.995	**0.889**	**0.854**	**0.781**	**0.926**
		GDCM(AH = 8)	**0.9**	**0.900**	**0.9**	**0.899**	0.996	0.996	0.996	0.996	0.826	0.719	0.545	0.893
		GDCM(AH = 1)	0.859	0.887	0.859	0.914	0.998	0.998	0.998	0.998	0.968	**0.968**	**0.968**	0.968
		GDCM(AH = 6)	0.913	0.915	0.91	0.915	0.995	0.995	0.995	0.995	**0.974**	0.886	0.952	0.82
		GDCM(AH = 8)	**0.922**	**0.922**	**0.921**	**0.923**	**0.999**	**0.999**	**0.999**	**0.999**	0.973	0.966	0.951	**0.981**
	Graph Complement	**GDCM(AH = 1)**	**0.918**	0.883	**0.916**	0.849	0.998	0.998	0.998	0.998	**0.965**	**0.956**	**0.936**	**0.976**
		GDCM(AH = 4)	0.898	0.898	0.896	0.9	0.998	0.998	0.998	0.998	0.96	0.949	0.927	0.972
		GDCM(AH = 8)	0.9	**0.913**	0.896	**0.93**	**0.999**	**0.999**	**0.999**	**0.999**	0.957	0.946	0.922	0.97
		GDCM(AH = 1)	**0.903**	**0.903**	**0.901**	**0.905**	**0.998**	**0.998**	**0.998**	**0.998**	0.8	668	0.458	0.877
		GDCM(AH = 4)	0.889	0.888	0.889	0.888	0.997	0.997	0.997	0.997	**0.959**	**0.949**	**0.925**	**0.972**
		GDCM(AH = 8)	0.888	888	0.884	0.892	0.998	0.998	0.998	0.998	0.892	0.886	0.881	0.891
		GDCM(AH = 1)	**0.891**	**0.891**	**0.886**	**0.895**	0.997	0.997	0.997	0.997	0.97	0.963	0.946	0.979
		GDCM(AH = 4)	0.817	0.816	0.806	0.826	0.998	0.998	0.998	0.998	0.969	0.962	0.944	0.979
		GDCM(AH = 8)	0.879	0.879	0.876	0.881	**0.999**	**0.999**	**0.999**	**0.999**	**0.972**	**0.965**	**0.949**	**0.981**

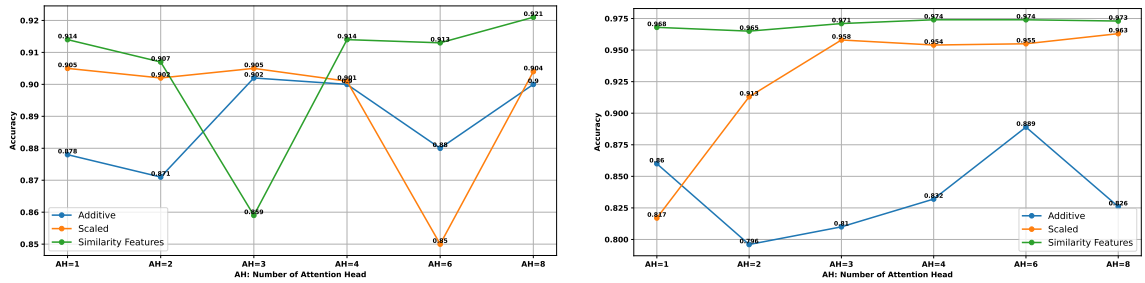
FEDS [38, 31] is that *MADS* [21] is dual summarization-based models, which is effective in partial incongruent news detection [21]. A model's response heavily depends on the underlying characteristics of samples in the dataset and the underlying distribution of datasets for different classes. Therefore, considering the characteristics of the FNC, NELA and ISOT datasets, the responses of the models also differ. As discussed in Subsection 4.6.1 *Fake* class sample NELA dataset represents the characteristics of the partially incongruent news article, and *Fake* class samples in the FNC dataset represent the characteristics of the fully incongruent news article. Similarly, the ISOT dataset represents the characteristics of real fake news. Relating the performance of baseline models from Table 6.3 and the characteristics of datasets, we can note the following important points.

- *The baselines and proposed model's performance improves when documents are smaller and have fewer paragraphs.* News articles (samples) in the ISOT dataset are small in terms of the number of paragraphs, and it is a balanced dataset (see Table 3.1). So, all the baselines and proposed model provided comparable performance with a small margin, and the performance of the proposed and baseline models is very high over ISOT dataset. Studies [167, 168, 21, 39] also reported similar performance of models and observations over the ISOT dataset.
- *The baseline models fail to detect partially incongruent articles efficiently.* NELA dataset represents the characteristics of partially incongruent and FNC represents the characteristics of fully incongruent news (refer Subsection 4.6.1). The performance of baseline models over the FNC dataset is excellent compared to the performance of baseline models over the FNC dataset (see Table 6.3). The performance of baseline models over the NELA dataset is average. From such observation about the performance of baseline models over the FNC and NELA dataset, we can conclude that the baseline models are adequate for fully incongruent news detection. However, the baseline models are ineffective for partially incongruent news article detection.

Motivated by the above contradicting responses of baseline models over NELA and FNC datasets and the inability of baseline models to detect partially incongruent news. We propose a *Graph-based Dual Context Matching* $(GDCM)$ model as shown in Figure 6.6. Table 6.3 presents the performance of two different setups of *GDCM*, namely (i) *Max-Min* and (ii) *Graph complement*. These two setups of proposed model *GDCM Max-Min* and *Graph complement* only differ in the way we form negative context node index set V (please refer to Section 6.6.2.1) and the way we obtain negative context representation $\mathbf{m}$ (please refer to Section 6.6.2.3). The proposed model *GDCM* with *Max-Min* and *GDCM* with *Graph complement* are further divided into three different setups, namely *Scaled*, *Additive* and *Sim* based on the attention methods used to form feature vector $\mathbf{f}$ (please refer Subsection 6.6.2.4). From Table 6.3, it is evident that the proposed model $GDCM(AH = 8)$ with *Max-Min* and *Sim* setup

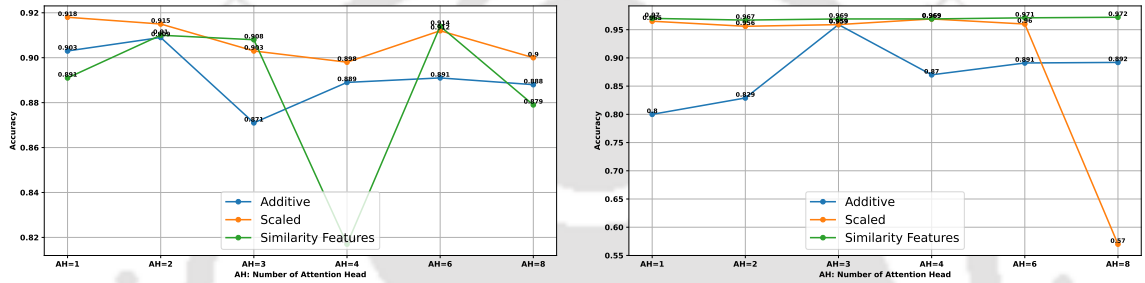
outperforms all the baseline methods (feature, encoding, hierarchical and summarization-based baseline) over NELA, FNC and ISOT dataset. Similarly, from the Table 6.3, it is also apparent that $GDCM(AH = 1)$ with *Graph Complement* and *Scaled* setup outperforms all the baseline methods (feature, encoding, hierarchical and summarization-based baseline) over NELA dataset and $GDCM(AH = 8)$ with *Graph Complement* and *Sim* setup outperforms all the baseline methods (feature, encoding, hierarchical and summarization-based baseline) over FNC dataset. From such observations over the performance of different setups of the proposed model, *GDCM*, we can draw the following conclusion: (i) The different setups of proposed model *GDCM* are very effective in partially incongruent news article detection compared to any baseline models. (ii) The different setups of the proposed model *GDCM* are also effective in fully incongruent news article detection compared to any baseline models. (iii) The performance of the proposed model *GDCM* is superior with *Max-Min* setup compared to the performance of the proposed model *GDCM* with *Graph Complement* setup. (iv) The performance of *GDCM* is also influenced by the number of attention heads used in the multi-head attention component of *GDCM* (see Subsection 6.6.2.3) and aggregation methods (*Scaled*, *Additive* and *Sim*) used to form a feature vector $\mathbf{f}$ in Subsection 6.6.2.4.

Similarly, Table 6.2 presents the performance of the proposed model *GCM*, which is further grouped into three groups based on how *GCM* the model aggregates the representation vector of- headline, headline-subgraph matching and body bigram network in the subsection 6.6.1.3, namely scaled, additive and sim. From Table 6.2, it is evident that $\mathbf{GCM\left(AH = 6\right)}$ with scaled attention outperforms all other baseline models over the NELA dataset with significantly high margins. Similarly, from table 6.2, it is also evident that $\mathbf{GCM}(AH = 6)$ a model with scaled and sim setup jointly outperforms other baseline models over FNC datasets. However, the performance of *GCM* the models is comparable to other baseline models in terms of accuracy for the ISOT dataset. Considering the performance of the *GCM* model and the characteristics of datasets, the following conclusion can be made: (i) As *GCM* models performance is superior to bag-of-word feature-based models (*FNC* [22], *UCLMR* [35] and *stackLSTM* [98]), hierarchical encoding-based models (*AHDE* [28]), *HDSF* [36], *HoBERT*, *HeLSTM*, $_{GraSHE^{((=)}}$ and *RaSHE* [21]) and summarization-based models (*FEDS* [38, 31] and *MADS* [21]) with a significantly high margin over NELA dataset and which represent the characteristics of partially incongruent news. Consequently, it can be concluded that the proposed model *GCM* is much more effective in detecting partially incongruent news compared to state-of-the-art methods in the literature. (ii) Considering the performance of the proposed model *GCM* on the FNC dataset, which represents characteristics of fully incongruent news, we can conclude that the proposed model *GCM* is also effective



(*a*) NELA Dataset (*b*) FNC Dataset

Fig. 6.7 Present the performance of proposed model *GDCM* with *Max-Min* setup, different attention heads, and different feature aggregation methods, namely, *Scaled*, *Additive* and *Sim*. Here, AH indicated attention head.



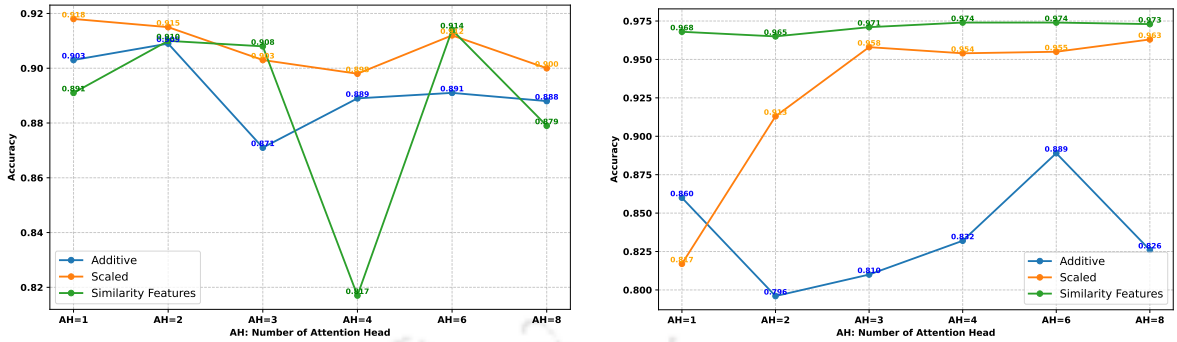
(*a*) NELA Dataset (*b*) FNC Dataset

Fig. 6.8 Present the performance of proposed model *GDCM* with *Graph Complement* setup, different attention heads and different feature aggregation methods, namely, *Scaled*, *Additive* and *Sim*. Here, AH indicates attention head.

in fully incongruent news detection.(iii) Performance of models over ISOT dataset is always high similar performance over ISOT dataset have been reported in studies [167, 168, 21, 39].

6.9.1 Optimal Attention Head and Effect of Aggregation Methods

Figure 6.7 and Figure 6.8 presents the performance of *GDCM* models with different attention heads and different aggregation methods (*Scaled*, *Additive* and *Sim*) used to form a feature vector **f** in Subsection 6.6.2.4 for *Max-Min* and *Graph Complement* setup. From Figure 6.7, we can observe that the performance of the proposed model *GDCM* with *Max-Min* setup is significantly better with eight attention heads and *Sim* setup over FNC, NELA and datasets. Similarly, from Figure 6.8, it is evident that the performance of the proposed model *GDCM* with *Graph Complement* setup is significantly better with one attention heads and *Scaled* setup over NELA dataset, whereas performance is superior with eight attention heads and



(*a*) NELA Dataset (*b*) FNC Dataset

Fig. 6.9 Present the performance of proposed model *GCM* with different attention heads and different feature aggregation methods, namely, *Scaled*, *Additive* and *Sim*. Here, AH indicated attention head.

Sim. setup over FNC datasets. From the above observations, the following conclusion can be made. (i) The number of attention heads also influences the performance *GDCM* model, and the performance of *GDCM* is optimal with eight attention heads in most of the setup. The aggregation methods (*Scaled*, *Additive* and *Sim.*) used to form a feature vector **f** in Subsection 6.6.2.4 also influences the performance *GDCM* model, and the performance of *GDCM* is maximum with *Sim* over FNC and NELA datasets in most cases. The possible reason behind the superiority in performance of the proposed model *GDCM* with *Sim* setup compared to the performance of *GDCM* with *Scaled* and *Additive* is that the similarity and difference (*Sim.*) approach forms a feature vector by estimating several combinations of similarity and difference between headline representation **h**, global representation of news body **b**, positive context representation **ḿ** and negative context representation **m**, which helps *Sim.* approach to verify the similarity between headline and news body along with consistency within news body. Whereas *the Additive* and *Scaled* approach forms a feature vector based on how similar headline representation **h** is with a global representation of news body **b**, positive context representation **ḿ** and negative context representation **m** and does not consider similarity within a global representation of news body **b**, positive context representation **ḿ** and negative context representation **m**. Accordingly, *Additive* and *Scaled* only verify how close the headline is with a different representation of the news body and does not form any feature vector which helps to verify consistency within the news body.

Similarly, We also conduct an empirical study to understand the impact of number of attention heads in the multi-head attention component of headline and subgraph context matching, as discussed in subsection 6.6.1.2. Figure 6.9 present the performance of the *GCM* model with different attention heads over NELA and FNC datasets, respectively. From

Table 6.4 The table presents the comparison between the performance of **GDCM**(**M**) and **GDCM**(**G**) for different values of radius and attention mechanism. **GDCM**(**M**) indicate the *Max-Min similarity* setup of the propose **GDCM** models, and **GDCM**(**G**) indicate the *Graph complement* setup of the proposed **GDCM** model. In the case of **GDCM**(**M**) setup, attention heads 1, 3 and 8 have been used for the NELA dataset, and attention heads 8, 4 and 6 have been used for the FNC dataset for scaled, additive and sim attention mechanism, respectively. Similarly, in the case of **GDCM**(**G**) setup, attention heads 1, 2 and 6 have been used for the NELA dataset, and attention heads 4, 3 and 6 have been used for the FNC dataset for scaled, additive and sim attention mechanism, respectively.

Model	Dataset	Radius	Scaled			Additive			Sim		
			Acc.	cong.	Incong.	Acc.	cong.	incong.	Acc.	cong.	incong.
GDCM(M)	NELA	1	0.905	0.901	0.909	**0.902**	**0.903**	**0.901**	**0.922**	**0.921**	**0.923**
		2	0.905	0.902	0.902	0.848	0.846	0.851	0.899	0.894	0.903
		3	**0.911**	**0.912**	**0.911**	0.890	0.886	0.894	0.906	0.903	0.910
	FNC	1	**0.974**	**0.974**	**0.974**	**0.889**	**0.781**	**0.926**	**0.974**	**0.952**	**0.982**
		2	0.956	0.920	0.970	0.886	0.775	0.924	0.972	0.949	0.980
		3	0.940	0.934	0.975	0.875	0.750	0.917	0.972	0.949	0.980
GDCM(G)	NELA	1	**0.918**	**0.916**	**0.920**	**0.909**	**0.910**	**0.909**	0.914	0.911	0.917
		2	0.916	0.914	0.918	0.838	0.837	0.840	**0.916**	**0.914**	**0.917**
		3	0.915	0.914	0.916	0.836	0.832	0.835	0.906	0.903	0.910
	FNC	1	**0.969**	**0.944**	**0.979**	**0.959**	**0.925**	**0.972**	**0.972**	**0.949**	**0.981**
		2	0.955	0.919	0.969	0.886	0.764	0.925	0.971	0.947	0.980
		3	0.94	0.934	0.975	0.875	0.750	0.917	0.954	0.917	0.968

Figure 6.9, it is evident that the performance of proposed model *GCM* is superior for attention head 6, 2 and 1 in scaled, additive and similarity feature (sim) setup over NELA dataset respectively. Similarly, from Figure 6.9, it is also evident that the performance of the proposed model *GCM* is superior for attention head 6 in scaled, additive and sim setup over the FNC dataset. From such observations, it is apparent that an optimal number of attention heads depends on the underlying characteristics of the dataset. Next, we compare the performance of the proposed model *GCM* in different setups used in subsection 6.6.1.3, namely scaled, additive and similarity features (sim). From Figure 6.9, it is apparent that the performance of proposed *GCM* models is superior in the case of *GCM* model with scaled setup in comparison to *GCM* model with additive and similarity feature (sim) over NELA datasets. Similarly, from Figure 6.9, it is evident that the performance of proposed *GCM* models is superior in the case of *GCM* model with similarity feature (sim) setup in comparison to *GCM* model with additive and similarity feature (sim) over FNC datasets. Consequently, we can conclude that the performance of the proposed model in different setups (discussed in Subsection6.6.1.3) relies on the underlying characteristics of datasets.

Table 6.5 The table presents the comparison between performance **GCM** models for 1, 2 and 3 hop neighbour need to be extracted body bigram network to construct subgraph for each word in the headline. Here, We fix the attention head of the proposed *GCM* model to attention heads 6, 2 and 1 have been used for the NELA dataset's scaled, additive and sim attention mechanism, respectively. Similarly, attention heads six have been used for scaled, additive and sim for the FNC dataset. We can observe that the performance *GCM* model is superior for the subgraph with the one-hop neighbour. Consequently, extracting bigram context (subgraph with one-hop neighbour) from different segments of the news body for each word in the headline as a subgraph and then matching the context of the subgraph and headline is sufficient for incongruent news detection for FNC and NELA datasets.

Dataset	Radius	Scaled			Additive			Sim		
		Acc.	cong.	Incong.	Acc.	cong.	incong.	Acc.	cong.	incong.
NELA	1	**0.929**	**0.928**	**0.930**	**0.911**	**0.909**	**0.912**	**0.924**	**0.923**	**0.925**
	2	0.905	0.902	0.907	0.883	0.883	0.883	0.923	0.922	0.923
	3	0.903	0.901	0.905	0.897	0.896	0.898	0.914	0.913	0.916
FNC	1	**0.972**	**0.950**	**0.981**	**0.937**	**0.880**	**0.957**	**0.972**	**0.950**	**0.981**
	2	0.955	0.919	0.969	0.894	0.808	0.927	0.965	0.963	0.965
	3	0.957	0.921	0.970	0.899	0.818	0.930	0.951	0.912	0.966

6.9.2 Impact of Radius in Subgraph

As discussed in Subsection 6.6.2.3 we construct a subgraph for $\mathcal{P}_i$ for every node $\mathbf{u}_i$ in the positive context node index set $\mathbf{U}$ by extracting $\mathbf{x}$ hop neighbour (*Radius*) of every node $\mathbf{u}_i$ from the body bigram network for both *Max-Min similarity* and *Graph complement* setup of our proposed model *GDCM*. Similarly, in the case of *Max-Min similarity* configuration of our proposed model, *GDCM*, we also construct a subgraph for $\mathcal{N}_i$ for every node $\mathbf{v}_i$ in the negative context node index set $\mathbf{V}$ by extracting $\mathbf{x}$ hop neighbour (*Radius*) of every node $\mathbf{v}_i$ from the body bigram network. So now the question is what is the optimal value of $\mathbf{x}$ or how many hop neighbour (*Radius*) nodes of $\mathbf{u}_i$ and $\mathbf{v}_i$ should be extracted from body bigram network to form subgraph $\mathcal{P}_i$ and $\mathcal{N}_i$ respectively. We conduct this empirical study to answer the above question and understand the impact of $\mathbf{x}$ hop neighbour (*Radius*) on the performance of the proposed model *GDCM*. For example, the value of $\mathbf{x}$ in Figure 6.4 is three for subgraph with root node "trump" and "whistleblower" because in Figure 6.3 we extract three hop neighbour of node "trump" and "whistleblower" from the body bigram network to form subgraph for keywords in the headline "trump" and "whistleblower". Intuitively, considering one-hop neighbour captures bigram context, two-hop neighbour captures the trigram context

of nodes in the body bigram network and so on. Table 6.4 presents the performance of the proposed model *GDCM* with *Max-Min similarity* setup $GDCM(M)$ and *Graph compliment* setup $GDCM(G)$ for different value of **x** (*Radius*). This study considers one, two and three hop neighbors (*Radius*) to study the impact of **x** hop neighbour (*Radius*) on the performance of the proposed model *GDCM*. From Table 6.4 it is evident that, except the case of *Scaled* attention for $GDCM(M)$ and *Sim* feature setup for $GDCM(G)$, the performance of both $GDCM(M)$ and $GDCM(G)$ is superior for one hop neighbour (*Radius*). Similarly, it is also apparent that as we increase the value of **x** (*Radius*) the performance of both $GDCM(M)$ and $GDCM(G)$ decrease except the case of *Scaled* attention for $GDCM(M)$ and *Sim* feature setup for $GDCM(G)$. The above observations indicate that extracting one-hop neighbour of $\mathbf{u}_i$ and $\mathbf{v}_i$ from the body bigram network to form subgraph $\mathcal{P}_i$ and $\mathcal{N}_i$ respectively is sufficient for the incongruent news detection task. In other words, extracting the *Bigram* dual context (positive and negative context) of words in the headline from different segments of news body and then estimating similarity between extracted *Bigram* dual context (positive and negative context) from news body and the headline is sufficient for incongruent news detection. All the results of *GDCM* presented in Table 6.3 and Figure 6.7, 6.8 are produced by considering one hop neighbour *Radius*. From Table 6.3 and Figure 6.7, 6.8 it is evident that the performance *GDCM* over ISOT dataset is maximum and consistent irrespective of hyperparameters. Therefore, this study does not consider the ISOT dataset in this empirical study.

Similarly, Table 6.5 also presents the performance of the *GCM* model for different values (one, two, three) of hop neighbour nodes (radius) considered while constructing the subgraph for headline keywords. Table 6.5 shows that the proposed *GCM* model performance is superior for a subgraph with one hop neighbour (radius). Consequently, extracting bigram context (subgraph with one-hop neighbour) from different segments of the news body for each word in the headline as a subgraph and then matching the context of the subgraph and headline is sufficient for incongruent news detection for FNC and NELA datasets.

6.9.3 Effect of ngram Network

As reported in the studies [169, 170], the performance of the proposed model is also influenced by ngram network in document classification and other NLP tasks. Intuitively, in the case of bigram network, there is an edge between two words if they appear together in the bigram context, for the trigram network there is an edge between two nodes if they appear together in the trigram context, and so on. As discussed in Section 6.6 given a

Table 6.6 Comparison of **GDCM**(**M**) and **GDCM**(**G**) performance for different N-gram networks. **GDCM**(**M**) refers to the *Max-Min similarity* setup of the proposed **GDCM** models, while **GDCM**(**G**) indicates the *Graph complement* setup of the proposed **GDCM** model. For the **GDCM**(**M**) setup, attention heads 1, 3, and 8 have been used for the NELA dataset, and attention heads 8, 4, and 6 have been used for the FNC dataset for scaled, additive, and sim attention mechanisms, respectively. Similarly, for the **GDCM**(**G**) setup, attention heads 1, 2, and 6 have been used for the NELA dataset, and attention heads 4, 3, and 6 have been used for the FNC dataset for scaled, additive, and sim attention mechanisms, respectively.

Model	N-gram	NELA						FNC					
		Scaled		Additive		Sim		Scaled		Additive		Sim	
		Acc.	F.	Acc.	F.	Acc.	F.	Acc.	F.	Acc.	F.	Acc.	F.
GDCM(M)	**Bigram**	0.905	0.900	**0.902**	**0.902**	0.922	0.922	**0.963**	**0.953**	**0.889**	**0.853**	**0.974**	**0.967**
	Trigram	**0.919**	**0.919**	0.866	0.865	0.917	0.917	0.958	0.947	0.846	0.775	**0.974**	**0.967**
	4gram	0.909	0.909	0.892	0.892	**0.923**	**0.923**	0.876	0.820	0.801	0.675	0.972	0.972
GDCM(G)	**Bigram**	**0.918**	**0.918**	**0.909**	**0.909**	**0.914**	**0.914**	**0.969**	**0.961**	**0.959**	**0.948**	0.972	0.933
	Trigram	0.913	0.913	0.871	0.870	0.913	0.912	0.960	0.949	0.921	0.900	0.972	0.964
	4gram	0.914	0.914	0.872	0.872	0.912	0.912	0.958	0.947	0.920	0.920	**0.974**	**0.967**

Table 6.7 The table compares the performance of **GCM** model for bigram, trigram and 4gram networks. Attention heads 6, 2 and 1 have been used for the NELA dataset's scaled, additive and sim attention mechanism. Similarly, attention heads six have been used for scaled, additive and sim for the FNC dataset. From this table, we can observe that the performance of the proposed model *GCM* is superior when the headline and body are represented using a bigram network. Similarly, we can also observe the reduction in performance of the proposed model *GCM* in the case of trigram or 4gram network. Consequently, we can conclude that representing headlines and bodies using the bigram network is sufficient for incongruent news detection tasks.

	NELA						ISOT						FNC					
	Scaled		Additive		Sim		Scaled		Additive		Sim		Scaled		Additive		Sim	
	Acc.	F.	Acc.	F.	Acc.	F.	Acc.	F.	Acc.	F.	Acc.	F.	Acc.	F.	Acc.	F.	Acc.	F.
Bigram	**0.929**	0.914	**0.911**	**0.911**	**0.924**	**0.924**	**0.998**	**0.998**	**0.998**	**0.998**	**0.998**	**0.998**	**0.972**	**0.965**	0.937	0.918	**0.972**	**0.965**
Trigram	0.918	0.917	0.897	0.897	0.908	0.908	0.997	0.997	0.996	0.996	0.998	0.998	0.97	0.962	0.787	0.786	0.97	0.962
4gram	0.919	**0.919**	0.638	0.601	0.918	0.918	**0.998**	**0.998**	0.997	0.997	0.997	0.997	0.957	0.957	**0.963**	**0.953**	0.962	0.953

news article with a pair of headlines $\mathcal{H}$ and news body $\mathcal{B}$, the *GDCM* model first obtains a separate bigram network $\mathcal{H}_n$ and $\mathcal{B}_n$ for headline and body respectively. So, to understand the influence ngram network on the performance of *GDCM* we conduct this empirical study by comparing the performance of *GDCM* model over bigram, trigram and 4gram networks. Table 6.6 presents the performance of the proposed model *GDCM* with *Max-Min similarity* setup *GDCM*(M) and *Graph compliment* setup *GDCM*(G) over bigram, trigram and 4gram network. From Table 6.6, it is evident that the performance of *GDCM*(M) and *GDCM*(G) is superior over bigram network in most of the cases for both NELA and FNC dataset. The above observation suggests that representing news headlines and bodies in the form of network-based on bigram context helps to capture better document structure and context, and representing news headlines and documents based on bigram context is sufficient for incongruent news detection tasks. All the results of *GDCM* presented in Table 6.3, 6.4 are produced by using bigram network. From Table 6.3 and Figure 6.7, and 6.8, it is evident that the performance of *GDCM* over ISOT dataset is maximum and consistent irrespective of hyperparameters. Therefore, this study does not consider the ISOT dataset in this empirical study.

Similarly, we also study the performance of the proposed model *GCM* when we represent the headline and body in the form of a trigram or 4gram network instead bigram network (by default bigram). We found the performance of the proposed model *GCM* is superior when we represent the headline and body using the bigram network instead of trigram and or 4gram network (refer Table 6.7).

6.10 Summary

This study proposes graph context matching based methods *Graph-based Dual Context Matching* **GDCM** and *Graph-based Context Matching* model *GCM* to detect incongruent news articles of different characteristics. **GDCM** first represents the headline and news body in the form of bigram network, and then extracts dual context (positive and negative) of the headline from the body bigram network. Then estimates the similarity between the headline and extracted dual context for incongruent news article detection. Similarly, *Graph-based Context Matching GCM* also constructs a bigram network news headline and news body, then also constructs a subgraph for each node in the headline by extracting the neighbour k-hop neighbour (radius) from body bigram network of the node which is contextually similar to the respective headline node. Then construct a feature vector based on the contextual similarity between the headline bigram network, body bigram network and subgraphs and

pass it to the neural network for incongruent news classifications. We conduct experiments over three publicly available benchmark datasets. Our experimental results suggest that the proposed methods **GCM** and **GDCM** model outperforms existing state-of-the-art methods in literature and efficiently detects partially incongruent news articles. The key observations from different empirical studies are as follows: i) extracting the *Bigram* dual context (positive and negative context) of words in the headline from different segments of new body and then estimating the similarity between extracted *Bigram* dual context (positive and negative context) from news body and the headline is sufficient for incongruent news detection and also helps in detecting partial incongruent news efficiently (refer to 6.9.2). (ii) Representing news headlines and bodies in the form of a bigram network based on bigram context helps to capture better document structure and context, and representing news headlines and documents based on bigram context is sufficient for incongruent news detection tasks 6.9.3. We identify two potential future research directions; (i) expansion of headlines with facts to overcome text length mismatch between headline and new body, (ii) Explainability of incongruent news detection.

Chapter 7

Conclusions and Future Work

This thesis aims to address the challenges of incongruent news article detection. First, this thesis proposes a `Gated Recursive And Sequential Deep Hierarchical Encoding` **GraSHE** method for detecting incongruent news articles by extending the hierarchy structure of news body from body to word level and incorporating incongruent weights. The proposed model, *(GraSHE)* captures the long-term dependencies and syntactic structure by incorporating sequential information at the paragraph and body level (using BiLSTM) and syntactic structure at the sentence level (child-sum Tree LSTM [41]). Further, unlike headline guided attention models [28][27], *(GraSHE)* also incorporates incongruity weight to capture non-dominant textual segments which are not congruent with other part of the news body. Second, this thesis proposes dual summarization-based methods *Multi-head Attention Dual Summarization* **MADS** and *dual-summarization based approach,* namely **DuSum** to detect partially incongruent news. Third, this thesis proposes two graph-based context matching methods, *Graph-based Context Matching* **GCM** and *Graph-based Dual Context Matching* **GDCM**. We also proposed datasets for incongruent and fake news detection in Hindi.

7.1 Gated Recursive And Sequential Deep Hierarchical Encoding

We proposed *Gated Recursive And Sequential Deep Hierarchical Encoding* model, namely *GraSHE*, to detect incongruent news articles. The proposed models capture long-term dependencies between words and syntactic structures of sentences at the sentence level and sequential structures at the body and paragraph level. From various experiments over three

datasets, it is observed that capturing structural properties at the sentence level improved the performance of incongruent news article detection tasks. The key observations from different empirical studies are as follows: (i) Incongruent news articles detection is a domain-dependent task in case of encoding-based models, i.e., the models perform inferior if news articles in training and testing datasets are from different domains or topics (refer to 4.7.3) ii) Encoding sentence structure instead of sequential encoding of sentence enhanced the performance (refer to 4.4), and iii) Performance of hierarchical structure-based models are superior to non-hierarchical structure-based models (refer to 4.5).

7.2 Dual Summarization

This thesis chapter proposed dual summarization-based methods, a *Multi-head Attention Dual Summarization* model, **MADS** and dual summary-based method **DuSum**, for detecting incongruent news articles of different characteristics. *MADS* extract two types of summary, viz. multi-head attention and convolution summary over positive and negative set separately. Similarly, **DuSum** also extract two different forms of summaries: *(i)* sequential weighted summation summaries and *(ii)* convolution summaries. Subsequently, estimate similarity features to check the similarity between headline and body and consistency within the news body for incongruent news article detection. From various experimental observations, it is evident that the proposed models outperform all of the baseline models for all three datasets of different characteristics. It is further observed that the proposed model is capable of capturing not only incongruent news articles but also partially incongruent news articles. We also conducted several ablation studies to study the strengths and weaknesses of the proposed models *MADS* and *DuSum*. The key observations from our ablation studies are as follows: *(i)* considering the summaries of both highly and poorly set is more effective than considering the summaries of only poorly congruent set or least k sentences; however, considering the summaries of only poorly congruent set or least k sentence is more effective than state-of-the-art similarity and summarization-based approach for incongruent news article detection (refer subsection 5.9.2). Similarly, in the case of *MADS*, we also observe that considering the summaries of both positive and negative sets is more effective than considering the summaries of only the negative set; however, considering the summaries of only the negative set is more effective than state-of-the-art similarity and summarization-based approach for incongruent news article detection (refer subsection 5.9.3). *(ii)* Considering sequence-weighted summation and convolution summaries together boosts the performance of both the proposed *MADS* and *DuSuM* (refer to subsection 5.9.4). *(iii)* performance of *DuSUM* models

is superior for θ and K value 0.5 and 5 respectively (refer to subsection 5.9.5). *(iv)* The performance of *DuSum* with contextualized LSTM is superior over *DuSum* with BiLSTM (refer subsection 5.9.6). This study can be extended in multiple directions; one significant extension of this study could be to study the effect of considering triplet loss function, contrastive learning, or considering attention between summaries instead of estimating similarity between summaries by estimating similar features in subsection 5.6.1.5.

7.3 Graph Context Matching

This study proposes graph context matching based methods *Graph-based Dual Context Matching* **GDCM** and *Graph-based Context Matching* model *GCM* to detect incongruent news articles of different characteristics. **GDCM** first represents the headline and news body in the form of bigram network, and then extracts dual context (positive and negative) of the headline from the body bigram network. Then estimates the similarity between the headline and extracted dual context for incongruent news article detection. Similarly, *Graph-based Context Matching GCM* also constructs a bigram network news headline and news body, then also constructs a subgraph for each node in the headline by extracting the neighbour k-hop neighbour (radius) from body bigram network of the node which is contextually similar to the respective headline node. Then construct a feature vector based on the contextual similarity between the headline bigram network, body bigram network and subgraphs and pass it to the neural network for incongruent news classifications. We conduct experiments over three publicly available benchmark datasets. Our experimental results suggest that the proposed methods **GCM** and **GDCM** model outperforms existing state-of-the-art methods in literature and efficiently detects partially incongruent news articles. The key observations from different empirical studies are as follows: i) extracting the *Bigram* dual context (positive and negative context) of words in the headline from different segments of new body and then estimating the similarity between extracted *Bigram* dual context (positive and negative context) from news body and the headline is sufficient for incongruent news detection and also helps in detecting partial incongruent news efficiently (refer to 6.9.2). (ii) Representing news headlines and bodies in the form of a bigram network based on bigram context helps to capture better document structure and context, and representing news headlines and documents based on bigram context is sufficient for incongruent news detection tasks 6.9.3.

Table 7.1 Present the mapping between incongruent news detection characteristics and literature studies addressing respective characteristics.

Models			C1	C2	C3	C4
Baseline	Feature	*FNC* [22]	✓			
		UCLMR [35]	✓			
		StackLSTM [98]	✓	✓		
	Encoding	*BiLSTM*	✓	✓		
		BERT	✓	✓		
		RoBERT	✓	✓		
	Hierarchical	*AHDE* [28]	✓	✓		
		HDSF [36]	✓	✓		
		GHDE [1]	✓	✓		
	Summ.	*FEDS* [38, 31]	✓	✓		
		Poshan [37]			✓	
Proposed	Hierarchical	*HoBERT* (Chapter 4) [39]	✓	✓		
		HeLSTM (Chapter 4) [39]	✓	✓		
		GraSHE$^{(=)}$ (Chapter 4) [39]	✓	✓		
		RaSHE (Chapter 4) [39]	✓	✓		
	Summ.	*MADS* (Chapter 5) [21]	✓	✓		✓
		DuSum (Chapter 5)	✓	✓		✓
	Graph	*GCM* (Chapter 6)	✓	✓	✓	✓
		GDCM (Chapter 6)	✓	✓	✓	✓

7.4 Mapping between Characteristics of Incongruent News and Literature Studies Addressing Respective Characteristics

As discussed in Section 1.3 and also reported in [3, 21], there are four key characteristics of incongruent news. **Characteristics (C1) :** The claim made in the headline is unrelated to or contradicts the claim made in the news body. **Characteristics (C2) :** News headlines and body are from the same topic and about the same event, but the content in the headline and body are not related to each other. **Characteristics (C3) :** The headline and body communicate or describe an authentic event or incident, but the dates or names of entities in the news headline and body are manipulated. **Characteristics (C4) :** one or more paragraphs of the news body are congruent to the headline, and one or more paragraphs of the news body

are not congruent to the headline (referred to as *partially incongruent*). Table 7.1 presents the relation between model strength and characteristics of incongruent news detection. From Table 7.1, the following conclusion can be made. *(i)* Except *StackLSTM* [98] other bag-of-word feature-based models could not detect incongruent news articles with *(C1)*, *(C2)* and *(C4)* characteristics. However, *StackLSTM* [98] also fails to detect incongruent news articles with *(C3)* and *(C4)* characteristics. *(ii) Encoding* and *hierarchical* encoding based effectively handle incongruent news articles with *(C1)* and *(C2)* characteristics but fails to detect incongruent news articles with it *(C3)* and *(C4)* characteristics. The study [37] proposed *POSHAN* models which efficiently handle *(C3)* characteristics of incongruent news articles. However, *Feature*, *Encoding* and *Hierarchical* encoding-based models fail in the case of text length mismatch between headline and news body [21, 1]. The study [31, 38] proposed a summarization-based model *FEDS* to overcome the text length mismatch between headline and body. However, *FEDS* [31, 38] model also fails to detect partial incongruent news (Characteristics *((C4))*. In Chapter 4, a recursive encoding-based model named **GraSHE** is proposed. This model acquires recursive encoding of headlines and deep hierarchical encodings of news bodies. Subsequently, it estimates the similarity between the encodings of headlines and news bodies for incongruent news article detection. The deep hierarchical encoding utilized in **GraSHE** assists in addressing (Characteristics *((C1))* and (Characteristics *((C2))*. However, this model encounters difficulties in handling numeric figures within headlines and dealing with partially incongruent news. Consequently, it falls short in addressing Characteristics (Characteristics *((C3))* and (Characteristics *((C4))*. In Chapter 5, dual summarization-based models, MADS and DuSum, are proposed for detecting partially incongruent news articles. Although the performance of MADS and DuSum surpasses that of **GraSHE**, achieving even better performance in detecting partially incongruent news articles is indeed crucial. Though the dual summarization-based models proposed in Chapter 5 to detect partially incongruent news articles (Characteristics *(C4)*) but dual summarization based models also fails to detect partial incongruent news articles efficiently. Therefore, Chapter 6 introduces the *Graph Context Matching* model (**GCM**) and the *Graph-based Dual Context Matching* model (**GDCM**), both of which efficiently detect partially incongruent news articles (Characteristics *(C4)*) while effectively addressing news articles with the characteristics *(C1)* and *(C2)* of incongruent news.

7.5 Future works

The thesis identifies several potential future research directions:

1. *Knowledge-based Summarization*: Given that many news articles pertain to real-world events or entities, enhancing summarization techniques by incorporating knowledge bases such as Wikipedia or news archives could be explored.

2. *Explainability of Incongruent News Detection*: There is a need to delve into the explanation of predictions made by models, shedding light on which parts of the news body are deemed congruent or incongruent, and offering insights into why the model classifies news articles as such.

3. *Large Language Models (LLMs) for Incongruent News Article Detection*: Investigating the effectiveness of Large Language Models (LLMs) for the detection of incongruent news articles presents a promising avenue for future research.

7.6 Publications

Asterisk (*) denotes equal contributions.

7.6.1 From Thesis

- **Conference Publication from thesis**

    1. **Sujit Kumar**, Gaurav Kumar, and Sanasam Ranbir Singh. "**Text_Minor at CheckThat!-2022: Fake News Article Detection Using RoBERT**." In CLEF (Working Notes), pp. 554-563. 2022.
    2. **Sujit Kumar,** Gaurav Kumar, and Sanasam Ranbir Singh. "**Detecting incongruent news articles using multi-head attention dual summarization**." In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 967-977. 2022.
    3. **Sujit Kumar**, Rohan Jaiswal, Mohit Ram Sharma and Sanasam Ranbir Singh. "**Multiset Dual Summarization for Incongruent News Article Detection**" In Proceedings of the 20th International Conference on Natural Language Processing (**ICON-2023**)

- **Journal publication from thesis**

    1. **Sujit Kumar** Durgesh Kumar, and Sanasam Ranbir Singh. **Gated Recursive and Sequential Deep Hierarchical Encoding for Detecting Incongruent News Articles** In *IEEE Transaction on Computational Social System,* March 2023

2. Sujit Kumar, Saurabh Kumar and Sanasam Ranbir Singh. **A Headline-Centric Graph-Based Dual Context Matching Approach for Incongruent News Detection.** In *IEEE Transaction on Computational Social System,* March 2024
3. **Sujit Kumar**, Anant Shankhdhar, Divyam Singal, Bhuvan Aggarwal, Ahaan Sameer Malhotra, and Sanasam Ranbir Singh. **Fake News Article Detection Datasets for Hindi Language** In *Language Resources and Evaluation* [**Revision submitted**]
4. **Sujit Kumar**, Gaurav Kumar, and Sanasam Ranbir Singh. **Incongruent News Article Detection Using Dual Summary** In *IEEE Transactions on Neural Networks and Learning Systems* [**Under review**]
5. **Sujit Kumar**, Saurabh Kumar and Sanasam Ranbir Singh. **A Headline-Centric Approach for Incongruent News Detection Using Graph-Based Context Matching** In *Information Processing and Management* **[Under Review]**.

7.6.2 Outside Thesis

- **Conference publication outside the scope of thesis**

1. **Sujit Kumar**, Mohan Kumar, and Sanasam Ranbir Singh. **Language Independent Fake News Detection Over Social Media Networks Using Centrality Aware Graph Convolution Network**. In *Proceedings of the 9th International Conference on Mathematics and Computing* ***(ICMC-2023)***, Lecture Notes in Networks and Systems (LNNS, volume 697)
2. **Sujit Kumar**, Aditya Sinha, Soumyadeep Jana, Rahul Mishra and Sanasam Ranbir Singh. **jack-flood at SemEval-2023 Task 5: Hierarchical Encoding and Reciprocal Rank Fusion-Based System for Spoiler Classification and Generation** . In *Proceedings of the 17th International Workshop on Semantic Evaluation* ***(SemEval-2023)***, Association for Computational Linguistics
3. Monika Singh, **Sujit Kumar**, Tanveen and Sanasam Ranbir Singh. **ClusterCore at SemEval-2024 Task 7: Few Shot Prompting With Large Language Models for Numeral-Aware Headline Generation**. In *Proceedings of the 18th International Workshop on Semantic Evaluation* ***(SemEval-2024)***, Association for Computational Linguistics.

- **Journal publication outside the scope of thesis**

1. **Sujit Kumar**, Monika Singh, Abhishek Ranjan, Tanveen, and Sanasam Ranbir Singh. **Prompt-based Masked Language Modeling for Numerical Reasoning** In *ACM Transactions on Intelligent Systems and Technology (**ACM-TIST**)* **[Under Review]**.

2. **Sujit Kumar**, Shifali Agrahari, Priyank Soni, Aayush Sachdeva, and Sanasam Ranbir Singh. **Fake News Detection Using Social Context and Post Expansion in Social Media Networks.** In *Pattern Recognition Letters* **[Under Review]**.
3. **Sujit Kumar**, Mohit Ram Sharma , and Sanasam Ranbir Singh. **Topic Aware Summarization for Incongruent News Detection. [Preparing for submission]**.
4. **Sujit Kumar**, Advaita Mallik, Naman Anand, and Sanasam Ranbir Singh. **Prompt-based Masked Language Modeling for Multiple Choice Question Answering and Mathematical Reasoning [Preparing for submission]**.
5. **Sujit Kumar**, Abhishek Ranjan , and Sanasam Ranbir Singh. **Detecting incongruent news utilizing Prompt Tuning with Large Language Models (LLMs) [Preparing for submission]**.

7.7 GitHub Repositories

7.7.1 From Thesis

Gated Recursive and Sequential Deep Hierarchical Encoding

Multi-head Attention Dual Summarization

TextMinor at CheckThat! 2022: fake news article detection using RoBERT.

7.7.2 Outside Thesis

jack-flood at SemEval-2023 Task 5:

7.8 Miscellaneous Research Activities

7.8.1 Service

- **Reviewer:** EMNLP 2023, Information processing and management, journal.

References

[1] S. Yoon, K. Park, M. Lee, T. Kim, M. Cha, and K. Jung, "Learning to detect incongruence in news headline and body text via a graph neural network," *IEEE Access*, vol. 9, pp. 36 195–36 206, 2021.

[2] U. K. Ecker, S. Lewandowsky, E. P. Chang, and R. Pillai, "The effects of subtle misinformation in news headlines." *Journal of experimental psychology: applied*, vol. 20, no. 4, p. 323, 2014.

[3] S. Chesney, M. Liakata, M. Poesio, and M. Purver, "Incongruent headlines: Yet another way to mislead your readers," in *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 56–61.

[4] W. Wei and X. Wan, "Learning to identify ambiguous and misleading news headlines," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 4172–4178.

[5] N. Higdon, *The anatomy of fake news: A critical news literacy education*. University of California Press, 2020.

[6] K. Park, T. Kim, S. Yoon, M. Cha, and K. Jung, "Baitwatcher: A lightweight web interface for the detection of incongruent news headlines," in *Disinformation, Misinformation, and Fake News in Social Media*. Springer, 2020, pp. 229–252.

[7] A. Wang, P. Guo, and E. Wong, "Fake news detection: Misleading headlines and satire," *International Journal of Computational and Biological Intelligent Systems*, vol. 3, no. 2, 2021.

[8] D. A. Effron and M. Raj, "Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share," *Psychological science*, vol. 31, no. 1, pp. 75–87, 2020.

[9] P. H. Tannenbaum, "The effect of headlines on the interpretation of news stories," *Journalism Quarterly*, vol. 30, no. 2, pp. 189–197, 1953.

[10] J. Rieis, F. de Souza, P. V. de Melo, R. Prates, H. Kwak, and J. An, "Breaking the news: First impressions matter on online news," in *Proceedings of the international AAAI conference on web and social media*, vol. 9, no. 1, 2015, pp. 357–366.

[11] M. Gabielkov, A. Ramachandran, A. Chaintreau, and A. Legout, "Social clicks: What and who gets read on twitter?" in *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, ser. SIGMETRICS '16. Antibes Juan-les-Pins, France: Association for Computing Machinery, 2016, p. 179–192.

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

[13] U. K. Ecker, S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga, and M. A. Amazeen, "The psychological drivers of misinformation belief and its resistance to correction," *Nature Reviews Psychology*, vol. 1, no. 1, pp. 13–29, 2022.

[14] Y. Tsfati, H. G. Boomgaarden, J. Strömbäck, R. Vliegenthart, A. Damstra, and E. Lindgren, "Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis," *Annals of the International Communication Association*, vol. 44, no. 2, pp. 157–173, 2020.

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

[16] R. K. Garrett, E. C. Nisbet, and E. K. Lynch, "Undermining the corrective effects of media-based political fact checking? the role of contextual cues and naïve theory," *Journal of Communication*, vol. 63, no. 4, pp. 617–637, 2013.

[17] U. K. Ecker, S. Lewandowsky, and M. Chadwick, "Can corrections spread misinformation to new audiences? testing for the elusive familiarity backfire effect," *Cognitive Research: Principles and Implications*, vol. 5, pp. 1–25, 2020.

[18] C. G. Horner, D. Galletta, J. Crawford, and A. Shirsat, "Emotions: The unexplored fuel of fake news on social media," *Journal of Management Information Systems*, vol. 38, no. 4, pp. 1039–1066, 2021.

[19] B. Bago, D. G. Rand, and G. Pennycook, "Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines." *Journal of experimental psychology: general*, vol. 149, no. 8, p. 1608, 2020.

[20] A. M. Guess, M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, and N. Sircar, "A digital media literacy intervention increases discernment between mainstream and false news in the united states and india," *Proceedings of the National Academy of Sciences*, vol. 117, no. 27, pp. 15 536–15 545, 2020.

[21] S. Kumar, G. Kumar, and S. R. Singh, "Detecting incongruent news articles using multi-head attention dual summarization," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, 2022, pp. 967–977.

[22] D. Pomerleau and D. Rao, "Exploring summarization to enhance headline stance detection," in *Natural Language Processing and Information Systems*, ser. NLDB 2021, E. Métais, F. Meziane, H. Horacek, and E. Kapetanios, Eds., vol. vol 12801. Cham: Springer International Publishing, 2021, pp. 243–254.

[23] A. Hanselowski, P. Avinesh, B. Schiller, and F. Caspelherr, "Description of the system developed by team athene in the fnc-1, 2017," *Online: https://github.com/hanselowski/athene_system/blob/master/system_description_athene.pdf. Accessed*, vol. 1, no. 1, pp. 03–13, 2018.

[24] A. Hanselowski, A. PVS, B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, and I. Gurevych, "A retrospective analysis of the fake news challenge stance-detection task," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug 2018, pp. 1859–1874. [Online]. Available: https://aclanthology.org/C18-1158

[25] L. Borges, B. Martins, and P. Calado, "Combining similarity features and deep representation learning for stance detection in the context of checking fake news," *Journal of Data and Information Quality (JDIQ)*, vol. 11, no. 3, pp. 1–26, 2019.

[26] H. Karimi and J. Tang, "Learning hierarchical discourse-level structure for fake news detection," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, jun 2019, pp. 3432–3442.

[27] C. Conforti, M. T. Pilehvar, and N. Collier, "Towards automatic fake news detection: cross-level stance detection in news articles," in *Proceedings of the First Workshop on Fact Extraction and VERification*, 2018, pp. 40–49.

[28] S. Yoon, K. Park, J. Shin, H. Lim, S. Won, M. Cha, and K. Jung, "Detecting incongruity between news headline and body text via a deep hierarchical encoder," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 791–800, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/3756

[29] R. Sepúlveda-Torres, M. Vicente, E. Saquete, E. Lloret, and M. Palomar, "Headlinestancechecker: Exploiting summarization to detect headline disinformation," *Journal of Web Semantics*, vol. 71, p. 100660, 2021.

[30] R. Mishra, P. Yadav, R. Calizzano, and M. Leippold, "Musem: Detecting incongruent news headlines using mutual attentive semantic matching," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Miami, FL, USA: IEEE, 2020, pp. 709–716.

[31] G. Kim and Y. Ko, "Effective fake news detection using graph and summarization techniques," *Pattern Recognition Letters*, vol. 151, pp. 135–139, 2021.

[32] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1906–1919.

[33] S. Chen, F. Zhang, K. Sone, and D. Roth, "Improving faithfulness in abstractive summarization with contrast candidate generation and selection," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Jun. 2021, pp. 5935–5941.

[34] Z. Cao, F. Wei, W. Li, and S. Li, "Faithful to the original: Fact aware neural abstractive summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[35] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel, "A simple but tough-to-beat baseline for the fake news challenge stance detection task," *arXiv preprint arXiv:1707.03264*, vol. 1, no. 1, pp. 1–6, 2017.

[36] H. Karimi and J. Tang, "Learning hierarchical discourse-level structure for fake news detection," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3432–3442. [Online]. Available: https://aclanthology.org/N19-1347

[37] R. Mishra and S. Zhang, "Poshan: Cardinal pos pattern guided attention for news headline incongruence," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 1294–1303.

[38] G. Kim and Y. Ko, "Graph-based fake news detection using a summarization technique," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 3276–3280. [Online]. Available: https://aclanthology.org/2021.eacl-main.287

[39] S. Kumar, D. Kumar, and S. R. Singh, "Gated recursive and sequential deep hierarchical encoding for detecting incongruent news articles," *IEEE Transactions on Computational Social Systems*, 2023.

[40] S. Kumar, G. Kumar, and S. R. Singh, "Textminor at checkthat! 2022: fake news article detection using robert," *Working Notes of CLEF*, 2022.

[41] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, Jul 2015, pp. 1556–1566.

[42] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 39.

[43] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International journal of machine learning and cybernetics*, vol. 1, pp. 43–52, 2010.

[44] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, vol. 1, no. 1, pp. 1–12, 2013.

[45] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[46] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From traditional to deep learning," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 2, pp. 1–41, 2022.

[47] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[48] A. J. Robinson and F. Fallside, *The utility driven dynamic error propagation network*. University of Cambridge Department of Engineering Cambridge, 1987, vol. 1.

[49] P. J. Werbos, "Generalization of backpropagation with application to a recurrent gas market model," *Neural networks*, vol. 1, no. 4, pp. 339–356, 1988.

[50] R. J. Williams, "Complexity of exact gradient computation algorithms for recurrent neural networks," Technical Report Technical Report NU-CCS-89-27, Boston: Northeastern . . . , Tech. Rep., 1989.

[51] R. J. Williams and D. Zipser, "Gradient-based learning algorithms for recurrent networks and their computational complexity," in *Backpropagation*. Psychology Press, 2013, pp. 433–486.

[52] M. C. Mozer, "A focused backpropagation algorithm for temporal pattern recognition," in *Backpropagation*. Psychology Press, 2013, pp. 137–169.

[53] B. Ghojogh and A. Ghodsi, "Recurrent neural networks and long short-term memory networks: Tutorial and survey," *arXiv preprint arXiv:2304.11461*, 2023.

[54] T. Lin, B. Horne, P. Tiňo, and C. Giles, "Learning long-term dependencies is not as difficult with narx networks," *Advances in Neural Information Processing Systems*, vol. 8, 1995.

[55] H. Jaeger, M. Lukoševičius, D. Popovici, and U. Siewert, "Optimization and applications of echo state networks with leaky-integrator neurons," *Neural networks*, vol. 20, no. 3, pp. 335–352, 2007.

[56] I. Sutskever and G. Hinton, "Temporal-kernel recurrent neural networks," *Neural Networks*, vol. 23, no. 2, pp. 239–243, 2010.

[57] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, p. 1724.

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[59] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[60] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[61] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[62] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[63] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International conference on machine learning*. PMLR, 2018, pp. 4055–4064.

[64] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, "Developing real-time streaming transformer transducer for speech recognition on large-scale dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5904–5908.

[65] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[66] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[67] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction," *ACS central science*, vol. 5, no. 9, pp. 1572–1583, 2019.

[68] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, 2021.

[69] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," *Advances in neural information processing systems*, vol. 28, 2015.

[70] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power, "Semi-supervised sequence tagging with bidirectional language models," *arXiv preprint arXiv:1705.00108*, 2017.

[71] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[72] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*. Association for Computational Linguistics (ACL), 2015, pp. 632–642.

[73] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of NAACL-HLT*, 2018, pp. 1112–1122.

[74] B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[75] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171—-4186.

[76] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[77] M. Mohtarami, R. Baly, J. Glass, P. Nakov, L. Màrquez, and A. Moschitti, "Automatic stance detection using end-to-end memory networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 767–776. [Online]. Available: https://aclanthology.org/N18-1070

[78] J. Li, M.-T. Luong, D. Jurafsky, and E. Hovy, "When are tree structures necessary for deep learning of representations?" *arXiv preprint arXiv:1503.00185*, vol. 1, no. 1, pp. 1–11, 2015.

[79] N. K. Tran and W. Cheng, "Multiplicative tree-structured long short-term memory networks for semantic representations," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. New Orleans, Louisiana: Association for Computational Linguistics, Jun 2018, pp. 276–286.

[80] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[81] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order boltzmann machine," *Advances in neural information processing systems*, vol. 23, 2010.

[82] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," *Advances in neural information processing systems*, vol. 27, 2014.

[83] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[84] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[85] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

[86] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606–615.

[87] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[88] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[89] Y. Ma, H. Peng, and E. Cambria, "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[90] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in neural information processing systems*, vol. 28, 2015.

[91] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[92] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[93] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[94] T. A. Neyazi, A. Kalogeropoulos, and R. K. Nielsen, "Misinformation concerns and online news participation among internet users in india," *Social Media+ Society*, vol. 7, no. 2, p. 20563051211009013, 2021.

[95] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, I. Traore, I. Woungang, and A. Awad, Eds. Cham: Springer International Publishing, 2017, pp. 127–138.

[96] ——, "Detecting opinion spams and fake news using text classification," *Security and Privacy*, vol. 1, no. 1, p. e9, 2018.

[97] B. Horne, S. Khedr, and S. Adali, "Sampling the news producers: A large news and feature data set for the study of the complex media landscape," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, pp. 518–527, Jun. 2018.

[98] A. Hanselowski, A. PVS, B. Schiller, F. Caspelherr, D. Chaudhuri, C. M. Meyer, and I. Gurevych, "A retrospective analysis of the fake news challenge stance-detection task," in *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2018, pp. 1859–1874.

[99] P. Patwa, M. Bhardwaj, V. Guptha, G. Kumari, S. Sharma, S. Pykl, A. Das, A. Ekbal, M. S. Akhtar, and T. Chakraborty, "Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts," in *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Springer, 2021, pp. 42–53.

[100] M. Bhardwaj, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Hostility detection dataset in hindi," *arXiv preprint arXiv:2011.03588*, 2020.

[101] J. Badam, A. Bonagiri, K. Raju, and D. Chakraborty, "Aletheia: A fake news detection system for hindi," in *5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, 2022, pp. 255–259.

[102] S. Kumar and T. D. Singh, "Fake news detection on hindi news dataset," *Global Transitions Proceedings*, 2022.

[103] T. Fredriksson, D. I. Mattos, J. Bosch, and H. H. Olsson, "Data labeling: An empirical investigation into industrial challenges and mitigation strategies," in *International Conference on Product-Focused Software Process Improvement*. Springer, 2020, pp. 202–216.

[104] D. Q. Sun, H. Kotek, C. Klein, M. Gupta, W. Li, and J. D. Williams, "Improving human-labeled data through dynamic automatic conflict resolution," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 3547–3557.

[105] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: a big data-ai integration perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328–1347, 2019.

[106] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4411–4421.

[107] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," 2020.

[108] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," *arXiv preprint arXiv:2003.07082*, 2020.

[109] O. Ngada and B. Haskins, "Fake news detection using content-based features and machine learning," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, 2020, pp. 1–6.

[110] S. Mishra, P. Shukla, and R. Agarwal, "Analyzing machine learning enabled fake news detection techniques for diversified datasets," *Wireless Communications and Mobile Computing*, vol. 2022, 2022.

[111] H. E. Wynne and Z. Z. Wint, "Content based fake news detection using n-gram models," in *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, 2019, pp. 669–673.

[112] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81–106, 1986.

[113] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[114] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, "Fake news detection using machine learning ensemble methods," *Complexity*, vol. 2020, 2020.

[115] A. Agarwal and A. Dixit, "Fake news detection: an ensemble learning approach," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2020, pp. 1178–1183.

[116] J. Thorne, M. Chen, G. Myrianthous, J. Pu, X. Wang, and A. Vlachos, "Fake news stance detection using stacked ensemble of classifiers," in *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 80–83.

[117] G. Bhatt, A. Sharma, S. Sharma, A. Nagpal, B. Raman, and A. Mittal, "Combining neural, statistical and external features for fake news stance identification," in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 1353–1357.

[118] R. E. Schapire, "A brief introduction to boosting," in *Ijcai*, vol. 99, 1999, pp. 1401–1406.

[119] ——, "Explaining adaboost," *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pp. 37–52, 2013.

[120] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[121] J. R. Quinlan *et al.*, "Bagging, boosting, and c4. 5," in *Aaai/Iaai, vol. 1*, 1996, pp. 725–730.

[122] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[123] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[124] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, and P. Kumar, "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages," in *Findings of EMNLP*, 2020.

[125] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak, "Hierarchical transformers for long document classification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Singapore: IEEE, 2019, pp. 838–844.

[126] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[127] Y. Fung, C. Thomas, R. G. Reddy, S. Polisetty, H. Ji, S.-F. Chang, K. McKeown, M. Bansal, and A. Sil, "Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 1683–1698.

[128] J. Rieis, F. de Souza, P. Vaz de Melo, R. Prates, H. Kwak, and J. An, "Breaking the news: First impressions matter on online news," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 9, no. 1, pp. 357–366, Aug. 2021.

[129] B. C. Andrew, "Media-generated shortcuts: Do newspaper headlines present another roadblock for low-information rationality?" *Harvard International Journal of Press/Politics*, vol. 12, no. 2, pp. 24–43, 2007.

[130] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[131] S. Kumar and N. Shah, "False information on web and social media: A survey," *arXiv preprint arXiv:1804.08559*, vol. 1, no. 1, pp. 1–35, 2018.

[132] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–36, 2018.

[133] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, "Combating fake news: A survey on identification and mitigation techniques," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 3, pp. 1–42, 2019.

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

[135] S. B. Parikh and P. K. Atrey, "Media-rich fake news detection: A survey," in *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. Miami, FL, USA: IEEE, 2018, pp. 436–441.

[136] A. D'Ulizia, M. C. Caschera, F. Ferri, and P. Grifoni, "Fake news detection: a survey of evaluation datasets," *PeerJ Computer Science*, vol. 7, p. e518, 2021.

[137] S. Colaco, S. Kumar, A. Tamang, and V. G. Biju, "A review on feature selection algorithms," *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2018, Volume 2*, pp. 133–153, 2019.

[138] G. Bhatt, A. Sharma, S. Sharma, A. Nagpal, B. Raman, and A. Mittal, "On the benefit of combining neural, statistical and external features for fake news identification," *arXiv preprint arXiv:1712.03935*, 2017.

[139] T. Saikh, A. De, A. Ekbal, and P. Bhattacharyya, "A deep learning approach for automatic detection of fake news," in *Proceedings of the 16th International Conference on Natural Language Processing*. International Institute of Information Technology, Hyderabad, India: NLP Association of India, Dec 2019, pp. 230–238.

[140] D. Paschalides, C. Christodoulou, R. Andreou, G. Pallis, M. D. Dikaiakos, A. Kornilakis, and E. Markatos, "Check-it: A plugin for detecting and reducing the spread of fake news and misinformation on the web," in *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 2019, pp. 298–302.

[141] J. Jang, Y.-S. Cho, M. Kim, and M. Kim, "Detecting incongruent news headlines with auxiliary textual information," *Expert Systems with Applications*, vol. 199, p. 116866, 2022.

[142] C. Scanlan, *Reporting and writing: Basics for the 21st century*. San Diego, California: Harcourt College Publishers, 2000.

[143] G. Kim and Y. Ko, "Graph-based fake news detection using a summarization technique," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr 2021, pp. 3276–3280.

[144] Y. Wang, H.-Y. Lee, and Y.-N. Chen, "Tree transformer: Integrating tree structures into self-attention," 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)*. Hong Kong, China: Association for Computational Linguistics, Nov 2019, pp. 1061–1070.

[145] Q. Zhou, N. Yang, F. Wei, and M. Zhou, "Selective encoding for abstractive sentence summarization," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)"*. Vancouver, Canada: Association for Computational Linguistics, Jul 2017, pp. 1095–1104.

[146] S. T. Dumais *et al.*, "Latent semantic analysis," *Annu. Rev. Inf. Sci. Technol.*, vol. 38, no. 1, pp. 188–230, 2004.

[147] O. Ajao, D. Bhowmik, and S. Zargari, "Sentiment aware fake news detection on online social networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2507–2511.

[148] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares, "Sentiment analysis for fake news detection," *Electronics*, vol. 10, no. 11, p. 1348, 2021.

[149] C. Iwendi, S. Mohan, E. Ibeke, A. Ahmadian, T. Ciano *et al.*, "Covid-19 fake news sentiment analysis," *Computers and electrical engineering*, vol. 101, p. 107967, 2022.

[150] C. Li, H. Peng, J. Li, L. Sun, L. Lyu, L. Wang, S. Y. Philip, and L. He, "Joint stance and rumor detection in hierarchical heterogeneous graph," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2530–2542, 2021.

[151] F. Xu, V. S. Sheng, and M. Wang, "A unified perspective for disinformation detection and truth discovery in social sensing: A survey," *ACM Computing Surveys*, vol. 55, no. 1, pp. 1–33, 2021.

[152] B. Kim, A. Xiong, D. Lee, and K. Han, "A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions," *Plos one*, vol. 16, no. 12, p. e0260080, 2021.

[153] F. Zhou, X. Qi, K. Zhang, G. Trajcevski, and T. Zhong, "Metageo: a general framework for social user geolocation identification with few-shot learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[154] D. Patra, B. Jana, S. Mandal, and A. A. Sekh, "Understanding fake news detection on social media: A survey on methodologies and datasets," in *Artificial Intelligence: First International Symposium, ISAI 2022, Haldia, India, February 17-22, 2022, Revised Selected Papers*. Springer, 2023, pp. 226–242.

[155] D. Patra and B. Jana, "Fake news identification through natural language processing and machine learning approach," in *Computational Intelligence in Communications and Business Analytics: 4th International Conference, CICBA 2022, Silchar, India, January 7–8, 2022, Revised Selected Papers*. Springer, 2022, pp. 269–279.

[156] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[157] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver,

Canada: Association for Computational Linguistics, Jul. 2017, pp. 1073–1083. [Online]. Available: https://aclanthology.org/P17-1099

[158] M. Vicente, C. Barros, and E. Lloret, "Statistical language modelling for automatic story generation," *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 5, pp. 3069–3079, 2018.

[159] H. Chi and B. Liao, "A quantitative argumentation-based automated explainable decision system for fake news detection on social media," *Knowledge-Based Systems*, vol. 242, p. 108378, 2022.

[160] S. Li, J. Yang, G. Liang, T. Li, and K. Zhao, "Sybilflyover: Heterogeneous graph-based fake account detection model on social networks," *Knowledge-Based Systems*, vol. 258, p. 110038, 2022.

[161] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 1412–1421. [Online]. Available: https://aclanthology.org/D15-1166

[162] Y. Tay, A. T. Luu, and S. C. Hui, "Hermitian co-attention networks for text matching in asymmetrical domains." in *IJCAI*, 2018, pp. 4425–4431.

[163] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3982–3992.

[164] A. Mousa and B. Schuller, "Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1*. Association for Computational Linguistics, Apr. 2017, pp. 1023–1032.

[165] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. [Online]. Available: https://aclanthology.org/D14-1181

[166] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations*, 2013.

[167] M. K. Elhadad, K. F. Li, and F. Gebali, "A novel approach for selecting hybrid features from online news textual metadata for fake news detection," in *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*. Springer, 2019, pp. 914–925.

[168] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid cnn-rnn based deep learning approach," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100007, 2021.

[169] G. Nikolentzos, A. Tixier, and M. Vazirgiannis, "Message passing attention networks for document understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8544–8551.

[170] Y. Zhang, X. Yu, Z. Cui, S. Wu, Z. Wen, and L. Wang, "Every document owns its structure: Inductive text classification via graph neural networks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 334–339.

[171] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.