



# **Identify, Inspect and Intervene Multimodal Fake News**

By

Shivangi Singhal

Under the supervision of

Dr Rajiv Ratn Shah, IIIT Delhi

Prof Ponnurangam Kumaraguru, IIIT Hyderabad

Indraprastha Institute of Information Technology Delhi

June, 2023





# **Identify, Inspect and Intervene Multimodal Fake News**

By

Shivangi Singhal

Submitted in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy

to

Indraprastha Institute of Information Technology Delhi

June, 2023

# Certificate

This is to certify that the thesis titled “**Identify, Inspect and Intervene Multimodal Fake News**” being submitted by **Shivangi Singhal** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by her under our supervision. In our opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

**June, 2023**

**Advisor**

Dr Rajiv Ratn Shah  
Assistant Professor  
Indraprastha Institute of Information Technology, Delhi  
New Delhi 110020

**Advisor**

Prof Ponnuram Kumaraguru  
Professor  
International Institute of Information Technology, Hyderabad  
Hyderabad, Telangana 500032



# Acknowledgements

First and foremost, I extend my utmost gratitude to my advisors, Dr Rajiv Ratn Shah and Prof Ponnurangam Kumaraguru (PK), for their guidance and support during the PhD programme. Rajiv Sir always had energy, which helped maintain a positive lab attitude. His vision not only helped me shape my initial years of research, but I understood the nitty-gritty of research under his supervision. On the other side, PK constantly encouraged me to strive to be a better version of myself. He is my agony aunt, and I could always count on him to be there for me. He has always embraced my research suggestions and assisted in formulating specific study objectives. I want to thank my advisors for believing in my abilities and work and never giving up on me.

I would also like to thank TCS Research for supporting my PhD. I would like to express my gratitude to Girish Palshikar, Principal Scientist at TCS and Dr Lipika Dey, Chief Scientist at TCS, for their continued support all through the time. Their rigour feedback and mentorship greatly aided me in developing the research.

Next, I would like to thank my monitoring committee members, Dr Arun Balaji Buduru and Prof AV Subrmayam, for closely monitoring my work. They gave me constructive feedback that helped me improve my work quality. I developed the ability to articulate my research because of the questions and clarifications each offered throughout the discussions of my yearly evaluations.

I also want to thank and extend my sincere gratitude to my mentors who have hosted me for research internships. It was an honour to work with Prof Shinichi Satoh during my visit to the National Institute of Informatics (NII) in Japan. I made it through six months in a foreign nation because of the hospitality offered by the professor and his lab. Not only this, my first research output came under his supervision. He helped me acquire a lot of knowledge and abilities. I would also like to thank Johannes Hoffart and his team for providing me with a fantastic experience during my stint at Goldman Sachs. He was a great manager who allowed me to explore ideas and lead discussions. My buddy, Manjunath Hedge, was readily available and helped me brainstorm ideas. He helped me acclimate to the group and supervised me the entire time. Last, I thank Raj Sharma for the smooth onboarding process. Our interactions during catch-ups helped in off-loading the work stress. He assisted us in providing a healthy and productive work environment.

I worked with several students throughout my PhD journey, and I wish to thank them all. Especially, I extend my gratitude to Anubha Kabra, Mudit Dhawan, Tanisha Pandey and Saksham Mrig for setting up different modules in our various detection works. I could pull off the works mentioned in the thesis because a team of ninjas worked relentlessly with me. I would also like to thank all the MIDAS and Precog lab members that provided a holistic work environment, special thanks to all the Pillars and Shepherds for reviewing my work and providing constructive feedback. I also want to mention a few folks I got to know at IIIT-Hyderabad. My Hyderabad stay was exquisite because of the beautiful souls I met there,

and I sincerely want to thank Mehul, Raj, Kshitija, Ankita, Akhila and Shivali for all the love and warmth.

My circle of friends provided unwavering support as I navigated the challenges of the PhD process. I sincerely thank Anusha, Anurag, Atish, Pakhi, Karmanya, Anwesha, Tarun, Garima and Himanshu for their constant presence and encouragement. A special thanks goes to Hitkul, whose key role in guiding me through this journey cannot be overstated. I feel fortunate to have a friend like him by my side. He stood by me through every challenge, bearing the brunt of our hardships together.

Finally, I want to express my gratitude to my family, who are the essence of my existence. Their unconditional love and unwavering support have been my constant companions throughout high and low.

# Abstract

Fake news refers to intentionally and verifiably false stories created to manipulate people's perceptions of reality. Fake news is destructive and has been used to influence voting decisions and spread hatred against religion, organizations or individuals, resulting in violence or even death. It has also become a method to stir up and intensify social conflict.

Fake news has existed for a long time, but what led to the change was the rise of Web 2.0 technologies and social media, which broadened communication horizons. Social media emerged as a multidisciplinary tool for exchanging information and ideas. However, there are always two sides to a coin; social media is no exception. On the positive side, social media aid users in generating content that is a backbone for the masses to interact. The negative impact, however, is significantly more profound. First, the availability of the Internet and smart phones at nominal prices, in tandem with lowering entry barriers on such platforms, has given fake news a vast audience and allowed it to spread rapidly and widely. Second, social media platforms suffer from a lack of centralized gatekeeping to regulate the volume of generated content. As a result, online users fall prey to misleading stories. Individuals tend to accept information supporting their ideologies, preventing them from making rational decisions. Third, one can gain monetary benefits from such platforms by engaging the audience. Users are always drawn to sensational and controversial content. As a result, manipulators tend to generate fake news that receives a lot of attention and engagement and is more likely to be spread on such platforms.

Therefore, it is essential to understand the nature of fake news spreading online, devise new technologies to combat it, analyze the current detection methods and improve intuitive understanding among online readers. Henceforth, this PhD thesis addresses three fundamental challenges. First, we focus on devising different methods to *Identify*, a.k.a., detect fake news online by extracting different feature sets from the given information. By designing foundational detection mechanisms, our work accelerates research innovations. Second, our research closely *Inspect* the fake stories from two perspectives. First, from the information point of view, one can inspect fabricated content to identify the patterns of false stories disseminating over the web, the modality used to create the fabricated content and the platform used for dissemination. To study such changing dynamics of fake news, we select India as the region and built an extensive dataset to aid researchers in investigating such issues. Next, from the model point of view, we inspect detection mechanisms used in prior work and their generalizability to other datasets. The thesis also suggests *Intervention* techniques to help internet users broaden their comprehension of fake news. We discuss potential practical implications for social media platform owners and policymakers.

We design different multimodal fake news detection baselines to answer the first part of the thesis. Typically, a news article consists of a headline, content, top image and other

corresponding images. We begin by designing *SpotFake- a multimodal framework for fake news detection*. Our proposed solution identifies fake news without taking into account any additional subtasks. It exploits both the textual and visual features of an article. Specifically, we used language models (like BERT) to learn contextual representations for the text, and image features are learned from VGG-19 pre-trained on the ImageNet dataset. Our proposed method outperforms the baselines by a margin of 6% accuracy on average. Next, we present *SpotFake+: A Multimodal Framework for Fake News Detection via Transfer Learning*. It is a multimodal approach that leverages transfer learning to capture semantic and contextual information from the news articles and its associated images and achieve better performance for fake news detection. SpotFake+ is one of the first attempts that performs a multimodal approach for fake news detection on a dataset that consists of full-length articles. Next, we observed that most of the research on fake news has focused on detecting fake news by leveraging information from both modalities, ignoring the other multiple visual signals present in a news sample. To this, we created *Inter-modality Discordance for Multimodal Fake News Detection*. The proposed method leverages information from multiple images in tandem with the text modality to perform multimodal fake news detection. The count of images varies per sample basis, and our designed method can incorporate such changes efficiently. We adopt a multimodal discordance rationale for multimodal fake news detection. Our proposed model effectively captures the intra and inter-modality relationship between the different modalities. Lastly, we observed that existing research capture high-level information from different modalities and jointly models them to decide. Given multiple input modalities, we hypothesize that not all modalities may be equally responsible for decision-making. Hence, we present *Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection*. Here, we design a novel architecture that effectively identifies and suppresses information from weaker modalities and extracts relevant information from the strong modality on a per-sample basis. We also capture the intra-modality relationship that first generates fragments of a modality and then learn fine-grained salient representations from the fragments.

In the first part of the thesis, we make numerous attempts to design methods that can effectively identify fake news. However, in the process, we observed that the results reported by state-of-the-art methods indicate achieving almost perfect performance. In contrast, such methods fail to cope with the changing dynamics of fake news. The reasons could be twofold. First, the issue can reside in the information itself; second, the designed method is incapable of extracting the informative signals. Hence, in the second part of the thesis, we inspect fake news from two perspectives. From an information viewpoint, we study the changing dynamics of fake news over time. We selected India as the region from which we could derive conclusions, as little effort was made to study the menace of fake news in India. To this end, we built an extensive dataset, *FactDrill: A Data Repository of Fact-Checked Social Media Content to Study Fake News Incidents in India*. Further, using the dataset, one

can investigate the changing dynamics of fake news in a multi-lingual setting in India. The resource would aid in examining the fake news at its core, i.e. investigating the different kinds of stories being disseminated, the modalities or combinations used to create the fabricated content and the platform used for dissemination. From a model viewpoint, we examine the apparent discrepancy between current research and real applications. We hypothesize that the performance claims of the current state-of-the-art have become significantly overestimated. The overestimation might be due to the systematic biases in the dataset and models leveraging such preferences and not taking actual cues for detection. We conduct experiments to investigate the prior literature from the input data perspective, where we study statistical bias in the dataset. Our finding state that though reported performances are impressive, leveraging multiple modalities to detect fake news is far from solved.

The final section of the thesis focuses on developing intervention strategies that enable readers to identify fake news. We design *SachBoloPls*- a system that validates news on Twitter in real-time. It is an effort to curb the proliferation of debunked fake news online, make audiences aware of fact-checking organizations, and educate them about false viral claims. There are three components of *SachBoloPls* that are independent and can be extended to other social media and instant messaging platforms like Instagram, WhatsApp, Facebook, and Telegram. The proposed prototype can also incorporate regional languages making it a viable tool to fight against fake news across India. Designing effective interventions can encourage social media users to exercise caution while reading or disseminating news online. Lastly, we discuss potential practical implications for social media platform owners and policymakers.

# Contents

<b>1</b>	<b>Introduction</b>	<b>22</b>
1.1	Thesis Questions . . . . .	23
1.2	Thesis Contribution . . . . .	24
1.2.1	Identify Fake News . . . . .	24
1.2.2	Inspect Fake News . . . . .	25
1.2.3	Intervene Fake News . . . . .	26
1.3	Thesis Roadmap . . . . .	26
1.4	Thesis Publications . . . . .	27
<b>2</b>	<b>Background</b>	<b>29</b>
2.1	What is Fake News? . . . . .	29
2.2	Information Disorder . . . . .	36
2.2.1	Disinformation, Misinformation and Malinformation . . . . .	36
2.3	How Big the Problem of Fake News is? . . . . .	38
2.4	Multimodality and its Importance . . . . .	40
<b>3</b>	<b>Literature Review</b>	<b>42</b>
3.1	Content-based Fake News Detection . . . . .	42
3.1.1	Text-Based Methods . . . . .	43
3.1.2	Image-based Methods . . . . .	43
3.1.3	Multimodal Methods . . . . .	44
3.2	Context-based fake News Detection . . . . .	45
3.3	Validation and Robustness of Fake News Detection Methods . . . . .	47
3.4	Intervention Methods . . . . .	48
3.5	Technologies to Curb Fake News . . . . .	49
3.6	Fake News Resources . . . . .	51
<b>4</b>	<b>Designing Simple Baselines for Multimodal Fake News Detection</b>	<b>55</b>
4.1	Introduction . . . . .	55
4.2	Research Objective . . . . .	57
4.3	Does Using Multiple Modalities Help Identify Fake News? . . . . .	58
4.3.1	Observations from Survey . . . . .	59

4.3.2	Inferences . . . . .	62
4.3.3	Discussion . . . . .	62
4.4	SpotFake: A Multi-modal Framework for Fake News Detection . . . . .	63
4.4.1	Data . . . . .	63
4.4.2	Methodology . . . . .	64
4.4.3	Experimental Setup . . . . .	66
4.4.4	Results . . . . .	67
4.4.5	Ethical Considerations . . . . .	69
4.4.6	Limitations . . . . .	69
4.4.7	Discussion . . . . .	70
4.5	SpotFake+: A Multimodal Framework for Fake News Detection via Transfer Learning . . . . .	70
4.5.1	Data . . . . .	70
4.5.2	Methodology . . . . .	71
4.5.3	Experimental Setup . . . . .	72
4.5.4	Results . . . . .	72
4.5.5	Discussion . . . . .	76
4.6	Conclusion and Future Works . . . . .	76
<b>5</b>	<b>Exploring the Role of Multiple Images for Multimodal Fake News Detection</b>	<b>78</b>
5.1	Introduction . . . . .	78
5.2	Research Objective . . . . .	79
5.3	Data . . . . .	80
5.4	Methodology . . . . .	81
5.4.1	Inter-modality Discordance Score . . . . .	83
5.4.2	Unimodal Visual Feature Extractor . . . . .	84
5.4.3	Unimodal Text Feature Extractor . . . . .	85
5.4.4	Multimodal Fake News Detector . . . . .	85
5.4.5	Loss Functions . . . . .	86
5.5	Experimental Setup . . . . .	87
5.6	Results . . . . .	88
5.7	Limitations . . . . .	92
5.8	Conclusion and Future Works . . . . .	93
<b>6</b>	<b>Extracting Intra and Inter Modality Relationship for Multimodal Fake News Detection</b>	<b>94</b>
6.1	Introduction . . . . .	94
6.2	Research Objective . . . . .	96
6.3	Dataset . . . . .	96
6.4	Methodology . . . . .	96
6.4.1	Self Attention . . . . .	97

6.4.2	Text Embeddings . . . . .	99
6.4.3	Image Embeddings . . . . .	100
6.4.4	Multimodal Fusion . . . . .	100
6.5	Results . . . . .	102
6.6	Conclusion . . . . .	105
6.7	Limitations and Future Works . . . . .	106
<b>7</b>	<b>Inspecting Fake News</b>	<b>108</b>
7.1	Overview . . . . .	108
7.2	Research Objective . . . . .	109
7.3	FactDrill: A Data Repository of Fact-Checked Social Media Content to Study Fake News Incidents in India . . . . .	110
7.3.1	Data Curation . . . . .	112
7.3.2	Basic Dataset Characterization . . . . .	118
7.3.3	Use Cases . . . . .	120
7.3.4	FAIR Principles and Limitations . . . . .	121
7.3.5	Conclusion . . . . .	121
7.4	Validating Fake News Detection Methods . . . . .	121
7.4.1	Research Objective . . . . .	122
7.4.2	Methodology . . . . .	122
7.4.3	Results . . . . .	124
7.4.4	Conclusion . . . . .	127
7.4.5	Future Works . . . . .	127
<b>8</b>	<b>Intervention Strategies: Enhancing Data Accessibility via System Design</b>	<b>130</b>
8.1	Overview . . . . .	130
8.2	SachBoloPIs: A Realtime Identification of Fake News Online . . . . .	130
8.3	Prototype Design . . . . .	131
8.4	Implications . . . . .	132
8.5	Limitations . . . . .	133
8.6	Conclusion . . . . .	134
<b>9</b>	<b>Conclusion, Limitations and Future Works</b>	<b>135</b>
9.1	Summary . . . . .	136
9.1.1	Designing Multimodal Fake News Detection Baselines . . . . .	136
9.1.2	Identifying the role of Multiple Images . . . . .	136
9.1.3	Extracting Intra and Inter Modality Relationship . . . . .	137
9.1.4	Resource Creation for Indic languages . . . . .	138
9.1.5	Data Accessibility: A System Design . . . . .	139
9.2	Limitations . . . . .	139
9.3	Future Works . . . . .	140





# List of Figures

1.1	An illustration of the contribution towards multimodal fake news detection. We first design multimodal fake news detection baselines, <i>SpotFake</i> [235] and <i>SpotFake+</i> [237]. Next, to identify the role of multiple images, we propose <i>Inter-modality Discordance for Multimodal Fake News Detection</i> [236]. Lastly, to extract intra and inter modality relationship, we design <i>Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection</i> [239]. . . . .	24
2.1	An example of satire news. The news reported by the New York Times reads that an NFL playoff led to a win due to the turndown possession advantage possessed by the winning team. The win resulted in an outcry among fans of the opponent that led the NFL to change the Overtime Rules. The same information was re-reported by The Onion, a satirical website, humorously.	30
2.2	An example of the parody news. A fictional content of a fake Washington post created a buzz among the D.C commuters. Jacques Servin, a member of the self-described "trickster art duo" the Yes Men, created the post with two other people. The artist took ludicrousness of the issue, ' <i>President resigning</i> ', that felt much saner than reality. The news went viral on social media and thus has resulted in different interpretations by the audience. The creators mentioned that the presented issue is unreal, but many Americans, possibly most, would like to see it in future. This is a classic example of parody where the audience's wish is picked up by tricksters and presented outlandishly. .	31
2.3	An example of fabricated news story. It is a made-up story and has no basis in reality. One pro-tip to identifying such news is checking whether other websites report the story. If none, it might be fake. . . . .	32
2.4	An example of the manipulated content. The Figure (a) is the original picture that has been morphed to demonstrate the support of Mark Zuckerberg for the Brazil Protest, as shown in Figure (b). . . . .	33
2.5	An example of False Connection. The most prominent form of such stories is Clickbait, where the creator uses catchy headlines to lure audiences to click.	34

2.6	An example of False Context. Figure (a) displays a video capturing the collapse of the Kambaniru Bridge in Indonesia during 2021. Manipulators deliberately associated this video with a news article published in 2022, as depicted in Figure (b). . . . .	35
2.7	An example of Imposter News. The snapshot pretends to be ABC News, but it is not. One pro-tip to identify imposter content is to check the website source on Google. If the domain URL does not match, the website might be an imposter. . . . .	36
2.8	An overview of how fake news is defined in the modern era. All different forms of the content: propaganda, lies, conspiracies, rumours, hoaxes, hyper-partisan content, falsehoods, or manipulated media, are grouped under Disinformation, Misinformation and Malinformation. Collectively, called it as Information Disorder. . . . .	37
2.9	An example of false climate change story circulating on the web. The hoax was posted by Natural News claiming NASA declaring that sun responsible for global warming and not the human activities. . . . .	39
3.1	A flowchart outlining the Literature Review. We provide an exhaustive literature that addresses sub-fields such as multimodal fake news detection (Section 3.1 and 3.2), validation and robustness (Section 3.3), various intervention strategies (Section 3.4), tools and technologies to curb fake news dissemination (Section 3.5) and list of fake news datasets (Section 3.6). . .	42
4.1	An example of fake news that claims that the actor Sylvester Stallone died due to prostate cancer. . . . .	55
4.2	The reply from the actor Sylvester Stallone after fake news of his death spread.	56
4.3	(i) real photo of two Vietnamese siblings but being presented as it was captured during the Nepal 2015 earthquakes; (ii) photos of spliced sharks taken during Hurricane Sandy in 2012; (iii) a beautiful artwork portrayed as a picture of Solar Eclipse of March 20, 2015. . . . .	57
4.4	What were user responses towards (i) what are the primary sources of users to consume news and (ii) what medium does a user find most trustable when consuming news. We find that about 64% of the users consume news via traditional sources and 33.7% via social media. About 91% of the respondents trust traditional news sources for news consumption. . . . .	59
4.5	How frequently a user discusses the worlds happening in their immediate social circle. We find that only a small % of users, about 23.6%, talk about news daily. . . . .	60

4.6	We asked what users believe to be the driving force behind the spread of false information online. We found that 53.9% of users believe that manipulators who share false information online intend to spread propaganda against a person or organization. . . . .	61
4.7	We asked what users believe to be the driving force behind spreading false information on traditional platforms. We found that participants think the goal is to change the reader’s perspective on a particular subject or point of view. . . . .	61
4.8	A histogram depicting the average length of sentences in the Twitter MediaEval (a) and Weibo (b) datasets, respectively. The final length value is decided to be the one where 95% of the sentences are below it. This is 23 tokens for the Twitter dataset and 200 characters for Weibo dataset. . . . .	64
4.9	A schematic diagram of the proposed SpotFake model. Value in () indicates the number of neurons in a layer. SpotFake consists of two sub-module, text and visual feature extractor. The intermediate representations are fused to form the news vector that is then fed to the classification layer to determine the veracity of the sample. . . . .	66
4.10	Image samples from the FakeNewsNet repository that are discarded after the cleaning process. The removed images are either the logos or GIFs present on the news website. . . . .	71
4.11	Our proposed SpotFake+ method for multimodal fake news detection. SpotFake+ consists of two sub-module, text and visual feature extractors. The intermediate representations are passed through numerous dense layers to form the unimodal representations, which are then fused to form the multimodal vector fed to the classification layer to determine the veracity of the sample. The values in () indicates the number of neurons in a layer. . . . .	73
4.12	Visualisations of the loss functions produced by SpotFake+ using the GossipCop and Politifact datasets. From the graph, we can conclude that the performance of SpotFake+ is not a result of over or under-fitting. For both datasets, training loss and validation loss are close, with validation loss slightly greater than the training loss. . . . .	76
5.1	A sample of news article present on online media websites. The text written in <i>pink</i> and <i>black</i> color depicts the headline and content of the news, respectively. The image highlighted in <i>blue</i> and <i>red</i> represents the top image (first-image) and other-image present within the given news sample, respectively. . . . .	78

5.2	Illustration of our proposed model with the primary task being multimodal fake news detection. We introduce three auxiliary learning tasks, i.e. measuring inter-modality discordance score via contrastive loss, multiple visual feature extractor and textual feature extractor. The first number in [ ] depicts the count of the components in news sample $N$ . For instance, the news sample has one headline ( $H_i$ ), content ( $C_i$ ), and two image components ( $M_1$ , $M_2$ ). The second number in [ ] depicts the feature vector size obtained after passing through each layer of the proposed method. For instance, it is 512, 768, 512 and 128 for the text feature extraction module. . . . .	82
5.3	Measuring modality discordance score on train and test set of Gossip-Cop (clean) and Politifact respectively. . . . .	91
6.1	A sample of tweet from the MediaEval dataset [23]. The corresponding text reads, ‘Husband Gave His Unfaithful Ex-Wife Half Of Everything He Owned – Literally.’ The intra-modality feature extractor, one of the sub-components of our proposed technique, curates the fine-grained salient representations for the text, capturing words like <i>Half</i> from the caption and the image segments highlighted by the blue colored boxes. . . . .	94
6.2	The framework of our proposed model. It comprises of two sub-modules, one extracting intra-modality relationship across modalities and other capturing the inter-modality relationship with an emphasises on the modality that shows least resistance towards fake news classification. . . . .	98
6.3	Different variants of fake news detected by our proposed model. Our proposed sub-module, <i>inter-modality feature extractor</i> (discussed in Section 6.4.4) utilizes a multiplicative multimodal method to identify strong modality on a per-sample basis to determine veracity of the news sample. . . . .	106
7.1	An excerpt from our proposed FactDrill dataset depicting the <i>investigation_reasoning</i> attribute. The attribute is exclusive of the FactDrill dataset and is not present in any existing fact-checking datasets. The attribute provides minute details of the fact-checking process, i.e. social media account or website that posted the fake content (highlighted in yellow), the platform that first encountered the fake content (highlighted in orange), links to the archive version of the post if the original content is deleted (highlighted in green), tools used by fact-checkers to investigate the claim (highlighted in pink), and links to the supporting or refuting reports related to the claim (highlighted in blue). . . . .	111
7.2	Our proposed dataset curation pipeline. Step 1 describes the data collection process. This is followed by Step 2, describing the data extraction methodology, and Step 3, discussing the data annotation and evaluation process. . . .	112

7.3	We present a screenshot from a fact-checking website (Vishvas News) to depict different attributes present in the proposed FactDrill dataset. . . . .	114
7.4	A excerpt from the dataset displaying different attributes present in a sample of the proposed <i>FactDrill</i> dataset. The feature list is paced under different headers namely, meta features, text features, social features, media features, and event information. The attributes are discussed in Section 7.3.1. . . . .	116
7.5	(a) shows the distribution of different languages in our proposed FactDrill dataset, (b) shows the spread of data across different fact-checking websites. It also depicts the language supported by each website. . . . .	118
7.6	Topic Distribution in English, Hindi and Regional languages (left to right). The figures clearly show that most fake news dissemination across the country is centred on the political domain. . . . .	119
7.7	Circulation of fake news over the years in India. Figure (a) demonstrates sharp peaks and drops in the graph that will be an exciting study in the future. Figure (b) shows the year-wise distribution of samples in the dataset. During the data collection stage, the last sample collected was in June 2020. Hence, the sample count is shown till June. . . . .	119
8.1	The second sub-module of SachBoloPls, i.e. User Interface Agent, which gets activated when a user calls it. The module is responsible for reading data and extracting features of the query on which SachBoloPls is invoked.	131
8.2	The third sub-module of SachBoloPls, i.e. Multimodal Search and Ranking System, which performs pattern matching over the queried data and the database to find the relevant news article, if it has been debunked before by the any of the agencies. In failure, the system returns a ‘not found’ message.	132
9.1	A summary of the research done for the PhD. The thesis studies the domain of multimodal fake news from the viewpoint of multiple modalities. We address three fundamental challenges, i.e. Identify, Inspect and Intervene. .	135

# List of Tables

3.1	List of datasets for detecting fake news in news articles. The table highlights datasets published between the years 2014 to 2021. Empty cells indicate that the information is not available. . . . .	52
3.2	List of datasets for detecting fake news on social media. The table highlights datasets published between the years 2015 to 2020. Empty cells indicate that the information is not available. . . . .	53
3.3	List of datasets for detecting fake news on social media. The table highlights datasets published in the years 2020 and 2021. Empty cells indicate that the information is not available. . . . .	54
4.1	Demographics of the participants in the survey. Values in the table are in percentage. . . . .	58
4.2	The confusion matrix shows user responses towards understanding the term <i>fake news</i> and their ability to distinguish between the same. We received a total of 89 responses in the survey. We found that 75.3% $[(31 + 8 + 29)/89]$ of participants understood the term fake news; among them, only 34.8% $[31/89]$ were confident in their ability to tell the difference between true and false news. . . . .	60
4.3	An overview of hyper parameter setting used in SpotFake. . . . .	67
4.4	Classification results on Twitter and Weibo datasets. SpotFake is our proposed model, and we compare it with numerous unimodal and multiple modality baselines for a fair comparison. Our proposed method outperforms the state-of-the-art on Twitter and Weibo datasets by 3.27% and 6.83%, respectively. . . . .	68
4.5	The number of samples in the FakeNewsNet repository. The values in the brackets indicate samples fit to use after data pre-processing. . . . .	71
4.6	An overview of hyper parameter setting used in the two sub-modules of the SpotFake+. . . . .	74
4.7	Classification results on the FakeNewsNet repository. SpotFake+ is our proposed model, and we compare it with numerous unimodal and multiple modality baselines for a fair comparison. Our proposed method outperforms the state-of-the-art by a relatively large margin. . . . .	75

5.1	The dataset statistics used during the experiments. Values in () signify the final count of samples used during experimentation. Politifact (raw) and Gossipcop (raw) are the data samples present in the FakeNewsNet repository [227]. Gossipcop (clean) is a clean version of the Gossipcop (raw) dataset presented by Giachanou et al. [82]. . . . .	81
5.2	The train-test split statistics of the datasets used during the experiments. Values in (.) signifies the count of fake samples. . . . .	88
5.3	Comparison of our proposed model with the text <sup>†</sup> , image <sup>‡</sup> and single-image multimodal <sup>‡</sup> fake news baselines. Our proposed method outperforms single and multiple modality baselines on the given datasets. However, we observe inconsistent performance on the GossipCop (raw) dataset. . . . .	90
5.4	Comparison of our proposed model with the multi-image multimodal fake news detection baselines. . . . .	90
5.5	Comparison of the proposed model with its different variants. L2 and L3 signify the variant comprising only the text and visual features. Whereas L2+L3+L4 represents the multimodal framework in the absence of the inter-modality discordance score. We observe an improvement in the L2+L3+L4 variant with the addition of L1 (inter-modality discordance component). . .	92
6.1	Comparison of our proposed model with the unimodal text <sup>†</sup> , image <sup>‡</sup> and multimodal <sup>‡</sup> fake news detection baselines. Our proposed model beats the strongest baseline, SpotFake by an average of 3.05% and 4.525% on accuracy and F1-score, respectively. . . . .	104
6.2	Comparison of our proposed model with its different variants. We can see an improvement of 1.8% and 2.7% in the accuracy measures on Twitter and Weibo datasets, respectively, on adding intra modality relationship module ( <i>Proposed</i> ) to the base detection method ( <i>w/o Multiplicative</i> ). Similarly, to examine the effectiveness of the extracted fragments for the text and image modality, we evaluate the performance of the <i>Proposed w/o Text</i> and <i>Proposed w/o Image</i> , respectively. . . . .	105
7.1	An overview of the fact-checking sources considered during the data collection. Empty cells indicate that the information is not available on the website. . . . .	113
7.2	Inter-annotator agreement for the two tasks. The values in bold indicates that Gwet’s AC(1) and AC(2) scores are calculated for the samples. We observe a mix of moderate (0.41-0.60) and substantial (0.61-0.80) agreement for the majority of the samples. . . . .	117



7.3	The table presents the performances of the N-gram Convolutional Neural Network (CNN) on the FakeNewsNet repository, highlighting the effectiveness of the model in detecting fake news. The CNN achieves an accuracy of 90.38% on the Politifact dataset and 87.73% on the Gossipcop dataset. . . .	124
7.4	The top 10 cues by strength for each class across datasets. The strongest cues for both the real and fake classes appear to be fairly random, suggesting that the model considers a wide range of cues, possibly including irrelevant ones, when making decisions. . . . .	125
7.5	The ablation study results highlight the model's reliance on cues and linguistic information within the dataset. The significant drop in performance observed with each ablation indicates a strong dependence of the model on these cues. This suggests that the model heavily relies on specific patterns and features present in the data to make accurate predictions. . . . .	126

# Chapter 1

## Introduction

Is Fake the New Real? We must investigate the epics for fake news on an epic scale.<sup>1</sup> In Mahabharata, Dronacharya was deceived by the fake news of his son's death, ultimately leading to the demise of Dronacharya.<sup>1</sup> Even the *Samudra Manthan* narrative described in the Vedas, Puranas, and Mahabharata might be superficial.<sup>2</sup> Considering the beginning of human history, fake news initially travelled via word of mouth. The form of communication was highly unreliable and faded with the emergence of textual communication via print media in AD 1439 [245]. For instance, in AD 1566, handwritten news sheets started circulating in Venice, whereas the onset of the 16th century witnessed a weekly circulation of printed newspapers in Germany.<sup>3</sup> By then, history had documented numerous instances of fake news that could blow one's mind. The ones that gained traction were the Lisbon earthquake in AD 1755 [1], Nazi propaganda machines in AD 1800 [1], and New York's sun-great moon hoax in AD 1835 [245]. However, in AD 1890, publishers began to publish sensationalizing and rumoured news in an effort to get readers to their publications. Such form of newspaper reporting is coined as *Yellow Journalism*. It should be highlighted that the motivations for the newspapers published in AD 1890 engaging in yellow journalism are same as for the fake news creators today.

Eventually, print media started fading in 2014<sup>4</sup> and what led to the change was the rise of Web 2.0 technologies and social media, which broadened communication horizons. Social media emerged as a multidisciplinary tool for exchanging information and ideas that led to a paradigm shift in news consumption via online platforms, resulting in the rise of digital journalism. There are always two sides to a coin; social media is no exception. On the positive side, social media aid users in generating content that is a backbone for the masses to interact. The negative impact, however, is significantly more profound. First, print media

---

<sup>1</sup><https://www.rediff.com/news/column/fake-news-has-been-around-from-mahabharat-times/20180902.htm>

<sup>2</sup><https://www.speakingtree.in/blog/samudra-manthan-is-it-real>

<sup>3</sup>[https://en.wikipedia.org/wiki/History\\_of\\_newspaper\\_publishing](https://en.wikipedia.org/wiki/History_of_newspaper_publishing)

<sup>4</sup><https://assets.pewresearch.org/wp-content/uploads/sites/13/2016/06/30143308/state-of-the-news-media-report-2016-final.pdf>

conduct a one-way communication that disables users from manipulating the news. On the other side, social media platforms suffer from a lack of centralized gatekeeping to regulate the volume of generated content. Anything and everything become news in the digital world, and the consumer is left clueless as to what exactly makes news. As a result, most online users fall prey to misleading stories. Second, during print media times, we had limited sources publishing the content. Hence, monitoring the authenticity of the news was much more manageable. On the other hand, because of the amount of data produced, verifying news online takes significantly more time and effort. The rate at which news is generated is significantly faster than the rate at which it is verified. Last but not least, compared to the period of traditional media, the online world has a considerably greater audience reach. Individuals tend to accept information from social supporting their ideologies, preventing them from making rational decisions.

We have examined countless instances thus far demonstrating that fake news is a permanent reality with social media serving as a primary conduit for its creation and dissemination. Despite the difficulty in identifying, tracking, and controlling unreliable content, there must be an effort to halt its expansion. We continue to see various attempts made by researchers worldwide to solve the problem [229,308]. Our research endeavors also contribute to tackling various aspects of fake news, encompassing identification, inspection, and intervention. The premise of our thesis is firmly placed at the point where we analyze multiple facets of user-generated content produced online (Social Computing) in the form of text and visuals (Multimodal Computing) to investigate the field of fake news. Next, we discuss in detail the core contributions made towards this thesis.

## **1.1 Thesis Questions**

The User-Generated Content (UGC) on social media is a blend of text and visual or audio signals. Hence, we specifically investigate the field of fake news from the perspectives of different modalities, namely text and visuals. Our research addresses three fundamental challenges.

### **Research Challenge I**

How to IDENTIFY Fake News?

- How can we identify informative features and harness them for fake news detection?
- How can we leverage the information included in the multiple images of news?
- How can the relationships between and among various news modalities be established?

### **Research Challenge II**

How to INSPECT Fake News?

- How can we study the changing dynamics of fake news?

- How can we interpret the core reason for vulnerabilities in the detection methods?

### Research Challenge III

How to design INTERVENTION methods to educate the masses?

- How can we enable readers to identify fake news?

## 1.2 Thesis Contribution

Our research aims to identify, inspect and intervene in multimodal fake news. The thesis work is uniquely placed at the intersection of Multimodal and Social Computing. The thesis findings can be applied to design automated solutions to reduce the spread of false information online. In addition, the conclusions drawn from the in-depth inspection of the fake news and existing systems will aid researchers in designing better mechanisms in future. Next, we discuss the core contributions of the thesis in detail.

### 1.2.1 Identify Fake News

Typically, a news article consists of a headline, content, top image and other corresponding images. It becomes challenging to examine the credibility of news amid so many cues. The first part of the thesis focuses on extracting multimodal signals in the form of text and images and devising methods to detect fake news. The contributions are shown in Figure 1.1. Given the theme of multimodal fake news, we first designed simple yet powerful detection baselines leveraging text and images. Next, a news would also comprise multiple images. Hence, we considered all images present in the news in tandem with the text signals. Lastly, we modelled the intra and inter relationship between the modalities to detect fake news.

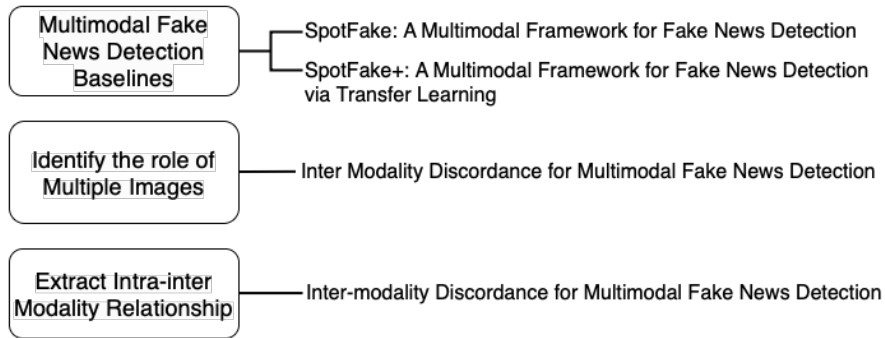


Figure 1.1: An illustration of the contribution towards multimodal fake news detection. We first design multimodal fake news detection baselines, *SpotFake* [235] and *SpotFake+* [237]. Next, to identify the role of multiple images, we propose *Inter-modality Discordance for Multimodal Fake News Detection* [236]. Lastly, to extract intra and inter modality relationship, we design *Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection* [239].

We begin by designing *SpotFake- a multimodal framework for fake news detection* [235]. Our proposed solution detects fake news without taking into account any additional subtasks. It exploits both the textual and visual features of an article. Specifically, we used language models (like BERT) to learn contextual representations for the text, and image features are learned from VGG- 19 pre-trained on the ImageNet dataset. Additionally, we develop *Spot-Fake+: A Multimodal Framework for Fake News Detection via Transfer Learning* [237]. It is a multimodal approach that leverages transfer learning to capture semantic and contextual information from the news articles and its associated images and achieve better performance for fake news detection. It is one of the first work that performs a multimodal approach for fake news detection on a dataset that consists of full-length articles. Next, we observed that most of the research on fake news has focused on detecting fake news by leveraging information from both modalities, ignoring the other multiple visual signals present in a news sample. To this, we present *Inter-modality Discordance for Multimodal Fake News Detection* [236]. The proposed method leverages information from multiple images in tandem with the text modality to perform multimodal fake news detection. The count of images varies per sample basis, and our designed method can incorporate such changes efficiently. We adopt a multimodal discordance rationale for multimodal fake news detection. Our proposed model effectively captures the intra and inter-modality relationship between the different modalities. Lastly, we observed that existing research capture high-level information from different modalities and jointly models them to decide. Given multiple input modalities, we hypothesize that not all modalities may be equally responsible for decision-making. Hence, we present *Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection* [239]. Here, we design a novel architecture that effectively identifies and suppresses information from weaker modalities and extracts relevant information from the strong modality on a per-sample basis. We also capture the intra-modality relationship. The idea is to generate fragments of a modality and then learn fine-grained salient representations from the fragments.

### 1.2.2 Inspect Fake News

In the first part of the thesis, we make numerous attempts to design methods that can effectively detect fake news. However, in the process, we observed that the results reported by state-of-the-art methods indicate achieving almost perfect performance. In contrast, such methods fail to cope with the changing dynamics of fake news. The reasons could be twofold. First, the issue can reside in the information itself; second, the designed method is incapable of extracting the informative signals. So in this part of the thesis, we inspect fake news from two perspectives.

From an information viewpoint, we study the changing dynamics of fake news over time. We select India as the region from which we could derive conclusions, as little effort was made to

study the menace of fake news in India. To this end, we built an extensive dataset, *FactDrill: A Data Repository of Fact-Checked Social Media Content to Study Fake News Incidents in India* [242]. Further, using the dataset, one can investigate the changing dynamics of fake news in a multi-lingual setting in India. The resource would aid in examining the fake news at its core, i.e. investigating the different kinds of stories being disseminated, the modalities or combinations used to create the fabricated content and the platform used for dissemination.

From a model viewpoint, we examine the apparent discrepancy between current research and real applications. We hypothesize that the performance claims of the current state-of-the-art have become significantly overestimated. The overestimation might be due to the systematic biases in the dataset and models leveraging such preferences and not taking actual cues for detection. We conduct experiments to investigate the prior literature from the input data perspective, where we study statistical bias in the dataset. Our finding state that though reported performances are impressive, leveraging multiple modalities to detect fake news is far from solved.

### **1.2.3 Intervene Fake News**

The final section of the thesis focuses on developing intervention strategies that enable readers to identify fake news. We design SachBoloPls- a system that validates news on Twitter in real-time. It is an effort to curb the proliferation of debunked fake news online, make audiences aware of fact-checking organizations, and educate them about false viral claims. There are three components of SachBoloPls that are independent and can be extended to other social media and instant messaging platforms like Instagram, WhatsApp, Facebook, and Telegram. The proposed prototype can also incorporate regional languages making it a viable tool to fight against fake news across India. Designing effective interventions can encourage social media users to exercise caution while reading or disseminating news online.

## **1.3 Thesis Roadmap**

The structure of the thesis document is as follows, aiming to cover the vast domain of fake news comprehensively. Chapter 2 serves as an introduction to the multimodal fake news domain, providing background information and defining relevant terminologies. This chapter helps readers grasp the broad scope of the field. In Chapter 3, an overview of existing literature is presented. This chapter focuses on previous research conducted in the fake news domain, specifically highlighting identification, inspection, and intervention methods. The first core area of the thesis, which is identifying fake news, is discussed in Chapters 4-6. Each chapter presents different proposed solutions for multimodal fake news detection, offering a comprehensive analysis of various approaches. Chapter 4 delves into designing simple yet effective multimodal fake news detection baselines. Chapter 5 addresses the challenge of handling multiple images found within a news sample. Lastly, Chapter 6

effectively identifies fake news by extracting intra and inter-modality relationships between the modalities. Chapter 7 delves into the second core area of the thesis, which is inspecting fake news. This chapter explores the examination of fake news, shedding light on its patterns, characteristics, and dynamics. The third main focus of the thesis, intervening in fake news, is covered in Chapter 8. We discuss our solution that encourages consumers to be cautious while reading or sharing news online. Finally, Chapter 9 concludes the thesis by discussing the overall impact of the work, potential future extensions, and any limitations encountered during the research process.

## 1.4 Thesis Publications

Here, we list the publications that contribute to the thesis.

- Chapter 4
  1. **Shivangi Singhal**, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. "Spotfake: A multi-modal framework for fake news detection." In 2019 IEEE fifth international conference on multimedia big data (BigMM), pp. 39-47. IEEE, 2019.
  2. **Shivangi Singhal**, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)." In Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 10, pp. 13915-13916. 2020.
- Chapter 5
  1. **Shivangi Singhal**, Mudit Dhawan, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. "Inter-modality Discordance for Multimodal Fake News Detection." In ACM Multimedia Asia, pp. 1-7. 2021. 13916. 2020.
- Chapter 6
  1. **Shivangi Singhal**, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. "Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection." In Companion Proceedings of the Web Conference 2022, pp. 726-734. 2022.
- Chapter 7
  1. **Shivangi Singhal**, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. "FactDrill: A data repository of fact-checked social media content to study fake news incidents in India." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 1322-1331. 2022.

2. **Shivangi Singhal**, Rishabh Kaushal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. “Fake news in India: scale, diversity, solution, and opportunities.” *Commun. ACM* 65, 11 (November 2022), 80–81. <https://doi.org/10.1145/3550493>

Addition publications while at IIIT-Delhi:

1. Hitkul Jangid, **Shivangi Singhal**, Rajiv Ratn Shah, and Roger Zimmermann. 2018. “Aspect-Based Financial Sentiment Analysis using Deep Learning.” In *Companion Proceedings of the The Web Conference 2018 (WWW ’18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1961–1966.
2. Hitkul Jangid, Simra Shahid, **Shivangi Singhal**, Debanjan Mahata, Ponnurangam Kumaraguru, and Rajiv Ratn Shah. “Aspect-based sentiment analysis of financial headlines and microblogs.” *Deep Learning-Based Approaches for Sentiment Analysis* (2020): 111-137.



## Chapter 2

# Background

Fake news has long existed and has instilled mistrust, fear, and confusion in the minds of its intended viewers. This, in turn, has led to widespread destruction causing harm to humankind. The term was also awarded as the word of the year by Collins in 2017.<sup>1</sup> However, earlier references to the word seem to misfit in the current scenario. Past studies have used the term to define related but distinct types of content, including satires [36, 210, 290], news parody [256], and news propaganda [269, 296]. However, current literature identifies fake news as false stories propagating on social media, particularly with an intent to discredit news organizations' critical reporting, further muddying discourse around fake news [229, 308]. This section reviews different terminologies used interchangeably with the term fake news. We provide definitions and examples to familiarize the readers with the distinct characteristics of each type.

### 2.1 What is Fake News?

Fake News combines two well-defined words. The term *fake* refers to something which is not genuine. It is also interchangeably used with other words like, forgery, fraud, non-credible and hoax. The term *news* refers to information about something that has happened recently. But weaving these two words together seems to introduce complexities and designing a universally accepted definition for it is still missing. If we go by the literature, there has been an evolution in the way researchers defined the terminology. Tandoc et al. [256] reviewed 34 academic research papers that used the term 'fake news' between the years 2003 and 2017 to see how literature has identified the term. Such studies have applied the term to define related but distinct types of content such as news satires, parody, fabricated content, manipulated stories to name a few. However, current studies [10, 235, 276] uses the term to define fake stories disseminating online with an intention to deliberately misinform or deceive readers. Next, we briefly recap the terms used interchangeably with fake news.

---

<sup>1</sup><https://www.theguardian.com/books/2017/nov/02/fake-news-is-very-real-word-of-the-year-for-2017>

## News Satires and News Parody

News Satire is a form of literary genre that takes the form of newscasts and uses humour, irony, and exaggeration to critique political, social or economic affairs. Whereas, News Parody is a literary work that draws the attention of the audience by relying on the comic effect introduced in the news. The non-factual information is used to inject humour into the news.

### The New York Times

#### N.F.L. to Change Postseason Overtime Rule After Bills' Playoff Loss

Each team will now get at least one overtime possession in playoff games.

Tom Pelissero @TomPelissero

NFL owners passed two other rules and resolutions today: one making permanent a health and safety-related change to free kick formations, and another allowing clubs to block personnel from taking assistant GM jobs elsewhere until after the draft.

**Approved 2022 Playing Rules Summary**

1. By Indianapolis and Philadelphia, amends Rule 16, to allow both teams an opportunity to possess the ball in overtime in the postseason.

**Approved 2022 Resolutions Summary**

G-1. By Baltimore, Buffalo, Philadelphia, and Tampa Bay; amends the Anti-Tampering Policy, in regard to Secondary Football Executive positions, to allow the employer club the choice to retain its player personnel staff through the Annual Selection Meeting. After the selection meeting through June 30, the employer club is required to grant permission for another club to interview and hire a non-high-level executive or non-secondary football executive for a secondary football executive position.

10:19 PM · Mar 29, 2022

240 Reply Share


Read 10 replies

PALM BEACH, Fla. — The N.F.L.'s 32 clubs passed a rule change on Tuesday to ensure that both teams would possess the ball at least once in overtime of postseason games. The measure comes months after Kansas City won a divisional round playoff game against the Buffalo Bills, who were not given a chance to score in overtime.

The change in the league's overtime rules was their first since 2010, when clubs voted to allow teams that scored a touchdown on the opening possession of overtime in a playoff game to win. (Before that, the team that scored first in any way in overtime won.) The rule, which by its nature gave an advantage to the team that won the overtime coin toss, was extended to the regular season in 2012.

the ONION® America's Finest News Source.

### NFL Satisfies Outraged Fans With New Overtime Rule That Both Teams Win



NEW YORK—Responding to outcry over the ending of a 2021 playoff victory by the Kansas City Chiefs over the Buffalo Bills, the National Football League reportedly satisfied fans Friday with a new overtime rule that both teams win.

"This rule change, which will be implemented next season and will apply to all playoff games, including the Super Bowl, means that no one will have to watch their team lose just because the other team played better—that simply wouldn't be fair," said NFL commissioner Roger Goodell, vowing that a situation where the Chiefs defeated the Bills just because the Bills defense couldn't stop them from scoring points was an "injustice" and that it would never happen again.

"We understand that fans were upset with the way our overtime rules functioned, so from now on, both teams will emerge victorious no matter what. The overtime period will last 15 minutes, or however long it takes each team to score as many points as they want to before they get sleepy and decide to go home. Fans watching a close, hard-fought game that ends tied in regulation will now be able to clap their hands in delight from the very beginning of the overtime period knowing that their team is going to win the game, along with the other team. This is the only fair way." Goodell added that if the new rule proved successful, the league was open to eliminating losses altogether and letting every team finish the season 16-0 with 1,000 touchdowns.

Figure 2.1: An example of satire news. The news reported by the New York Times reads that an NFL playoff led to a win due to the turndown possession advantage possessed by the winning team. The win resulted in an outcry among fans of the opponent that led the NFL to change the Overtime Rules. The same information was re-reported by The Onion, a satirical website, humorously.

Satires and parodies are an integral voice of Journalism that intends to communicate with the readers via injecting humour into the information. However, the difference lies in the injecting part. The core content of satires is based on actual news stories. It aims to present the news's direct commentary but in a hilarious manner. In contrast, parody picks up the ludicrousness of the issue and accentuates them by creating entirely fictitious stories. That

said, the content in parody is fabricated but not in satires. Moreover, while reading satires and parodies, it is presumed that both the author and reader share the joke. However, the issue arises when the presumption gets lost, i.e., the intention of the author and the gullibility of the reader goes out of sync. This creates a situation where the reader misunderstands the content and gets deceived by the information. It also results in sharing it with others without understanding the actual premise. Hence, this results in categorising news satires and parodies as fake news.



Figure 2.2: An example of the parody news. A fictional content of a fake Washington post created a buzz among the D.C commuters. Jacques Servin, a member of the self-described "trickster art duo" the Yes Men, created the post with two other people. The artist took ludicrousness of the issue, '*President resigning*', that felt much saner than reality. The news went viral on social media and thus has resulted in different interpretations by the audience. The creators mentioned that the presented issue is unreal, but many Americans, possibly most, would like to see it in future. This is a classic example of parody where the audience's wish is picked up by tricksters and presented outlandishly.

### News Fabrication

Fabrication represents entirely false articles, i.e., those with no factual basis but are presented in the style of actual news to mark legitimacy. The author of the fabricated news intends to misinform the readers and thus draws a parallel from the existing news stories to demonstrate authenticity. The success of fabricated stories is conditioned on the perseverance of the audience. If the readers demonstrate trust in a particular organization or a person, they are less likely to be vulnerable to fabricated stories and vice versa. Moreover, identifying fabricated news is challenging as non-news organizations or individuals could publish such stories under the veneer of legitimacy by adhering to the presentation styles. Further, when shared on social media, such stories earn authenticity since the source of acquiring the information

might be someone that readers generally trust.

By: Daniel Newton | [@NeonNettle](#) on 8th June 2018 @ 2.46pm



Twitter users discovered that National Human Trafficking Hotline was run by Hillary Clinton

Daughter of the former secretary of state Hillary Clinton has seemingly admitted that t 'Pizzagate' - the conspiracy theory born out of the Wikileaks emails allegedly exposing politicians running pedophile rings - is real on her Twitter on Thursday.

Figure 2.3: An example of fabricated news story. It is a made-up story and has no basis in reality. One pro-tip to identifying such news is checking whether other websites report the story. If none, it might be fake.

### Manipulated Content

Manipulated content is defined as news stories that perform modifications in the original text, images or videos to present a false narrative. Photo manipulation is a prevalent sort of manipulation in the age of the social media explosion [20, 138]. Cognitive Psychology demonstrates the efficacy of images in strengthening communication. Over the years, photographs have helped people better understand world events, including wars, scientific development, natural disasters and other countless noticeable occurrences. Photos verify that event did take place. With the advent of the Internet and the availability of smartphones at an affordable price, we all have become photographers, snapping events around us. However, the digital age has made it easy to alter pictures using manipulation software. Consumers need to develop a healthy scepticism when they encounter images. For instance, the Figure 2.4 (a) represents the image of Mark Zuckerberg holding a Thank You play card. The picture was taken when Facebook reached a milestone of 500 million users worldwide. To celebrate the success,

Facebook staffers took pictures of themselves with a thank you note. Mark Zuckerberg was one of the many staffers who acted. In contrast, the Figure 2.4 (b) shows the manipulated version of the original image. The picture is modified to demonstrate that Mark Zuckerberg supported the Brazilian protest at that time. This is a classic example of manipulated content where fabrication is introduced in the modality (text, image or video) to deceive the audiences.



Figure 2.4: An example of the manipulated content. The Figure (a) is the original picture that has been morphed to demonstrate the support of Mark Zuckerberg for the Brazil Protest, as shown in Figure (b).

### False Connection

As the name implies, it is a form of fake news where headlines, visuals or text do not support each other. That said, the parts of the news are legitimate, but the connection between them results in fabrication, and people can be taken in by the ruse. The most common example of this type of content is clickbait headlines. For instance, the Figure 2.5 illustrates a snapshot of a New York daily newspaper. The headline reads, '*Sugar as addictive as cocaine, heroin, studies suggest.*' It is a terrifying headline and sure to catch readers' eyes. And that is the point. Stories like this are created to lure audiences into clicking and reading the whole story. On reading, we might discover that theory claimed in the headline has only been proven to exist in rats, not humans. However, at that point, the damage has already been done. The sole aim of the creator was to make people click the content even if when people read the article, they feel that they have been deceived.

### Misleading Content

This type of content is when there is a misleading use of information to frame issues or individuals in specific ways by cropping photos or selecting quotes or statistics. Misleading content is the most challenging kind of fake news to uncover. Misleading content can find its way into various genuine stories. It is so hard to discover because it requires expertise or



## HEALTH

# Sugar as addictive as cocaine, heroin, studies suggest

By ROSEMARY BLACK  
DAILY NEWS STAFF WRITER • Dec 12, 2008 at 7:13 pm



Listen to this article



It's one addiction that won't land you in court or an inpatient rehab. But sugar - as anyone who mainlines sweets can attest - can be just as habit-forming as cocaine.

Researchers at [Princeton University](#) studying bingeing and dependency in rats have found that when the animals ingest large amounts of sugar, their brains undergo changes similar to the changes in the brains of people who abuse illegal drugs like cocaine and heroin.

"Our evidence from an animal model suggests that bingeing on sugar can act in the brain in ways very similar to drugs of abuse," says lead researcher and [Princeton](#) psychology professor [Bart Hoebel](#).

In the studies, he explains, animals that drank large amounts of sugar water when hungry experienced behavioral changes, too, along with signs of withdrawal and even long-lasting effects that resemble cravings.

Figure 2.5: An example of False Connection. The most prominent form of such stories is Clickbait, where the creator uses catchy headlines to lure audiences to click.

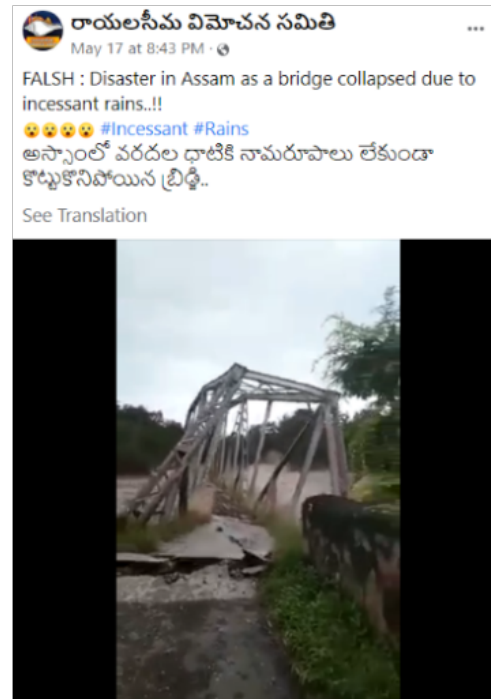
knowledge about a given subject to determine whether any news article's facts and details are misrepresented. Fact-checking resources can help you make sense of these details.

## False Context

One of the most common forms of fake news witnessed over the Internet. False context refers to news stories sharing genuine information with false contextual content. It is often used interchangeably with the words like Misrepresented, Misinterpreted or Misappropriated. One of the most significant issues with such kind of fake stories is that it is often seen being re-circulated out of their original context. For instance, The Figure 2.6 (a) shows a video of the Kambaniru Bridge in Indonesia collapsing in 2021. However, in Figure 2.6 (b), a separate video depicting a bridge collapse in Assam due to heavy rains started circulating on Facebook on 17 May 2022. It is important to note that this video had no connection to the actual floods in Assam. Manipulators falsely associated the old video from Indonesia to deceive and mislead readers, highlighting the use of false context in spreading fake news.



(a)



(b)

Figure 2.6: An example of False Context. Figure (a) displays a video capturing the collapse of the Kambaniru Bridge in Indonesia during 2021. Manipulators deliberately associated this video with a news article published in 2022, as depicted in Figure (b).

### Imposter Content

Imposter content refer to stories that are published by fake news websites trying to imitate legitimate news agencies. If the reader is not familiar with the source being authentic, identifying imposter websites might prove to be challenging. However, a careful inspection into the URL can almost always sniff-out them out.

To summarize, we discussed how literature had operationalized fake news: satires and parody, fabrication, manipulated content, misleading stories, false context and false connection. Current works investigating this vast domain of information pollution have decided to refrain from calling it fake news due to the following reasons:

- With the advent of the digital age, people have started consuming content online. The information can be a rumour or presented in the form of tweets, memes, manipulated videos, hyper-targeted dark ads and old photos re-shared as new. Thus, defining every piece of information available online as the news seems somewhat inappropriate.
- Individuals, websites, organizations or politicians have started using the terminology to undermine stories or clamp down upon disagreeable events.

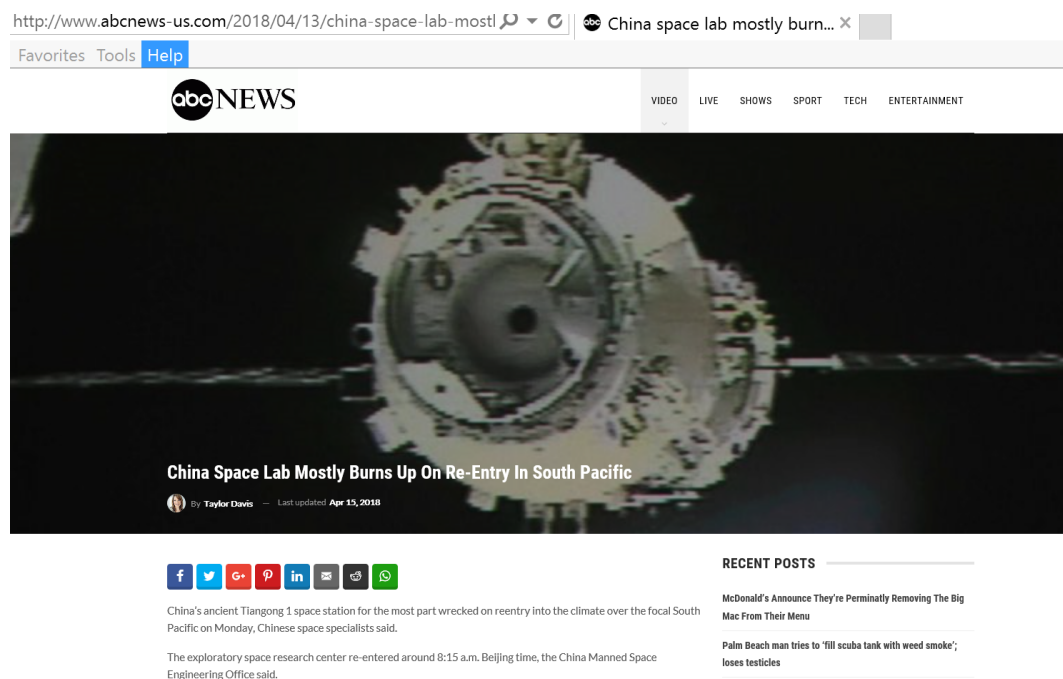


Figure 2.7: An example of Imposter News. The snapshot pretends to be ABC News, but it is not. One pro-tip to identify imposter content is to check the website source on Google. If the domain URL does not match, the website might be an imposter.

Next, we highlight the correct phrase in the literature for information pollution and further describe a conceptual framework for examining its different forms.

## 2.2 Information Disorder

In the previous section, we provide numerous examples to illustrate the typology of types of fake news. However, the failure of the term to capture the new reality is the reason not to use it. Claire Wardle and Hossein Derakhshan [280] advocate using the terms that best describe the content- propaganda, lies, conspiracies, rumours, hoaxes, hyper-partisan content, falsehoods or manipulated media. They present a conceptual framework to examine the information pollution by identifying them as mis-information, mal-information and dis-information. Collectively, called it as the information disorder.

### 2.2.1 Disinformation, Misinformation and Malinformation

Claire Wardle and Hossein Derakhshan [280] presented a conceptual framework for examining information disorder, identifying three different types. The categorization is performed based on dimensions of harm and falseness. Let us begin by defining each of the terms.



1. Dis-information: When false information is shared with an intent to harm. The creation of such stories is motivated by three factors: to make money, to have political influence, either foreign or domestic, or to cause trouble for the sake of it.
2. Mis-information: When false information is shared with no intent to harm. When disinformation is shared it often turns into misinformation. This happens when a reader encounters a false story and shares it without realizing it is false. Socio-psychological factors drive the sharing of misinformation.
3. Mal-information: When the information shared is genuine, but the intent is to cause harm. For instance, when Russian agents hacked into emails from the Democratic National Committee and the Hillary Clinton campaign and leaked specific details to the public to damage their reputations.<sup>2</sup>

To summarize, the literature refrains from using the term fake news to describe the vast spectrum of information pollution. Instead, Information Disorder is considered an apt word. Further, based on the two conditions: intent to harm and falseness, the online content is divided into three categories: misinformation, dis-information and mal-information. Figure 2.8 depicts the granular level categorization of the fake content.

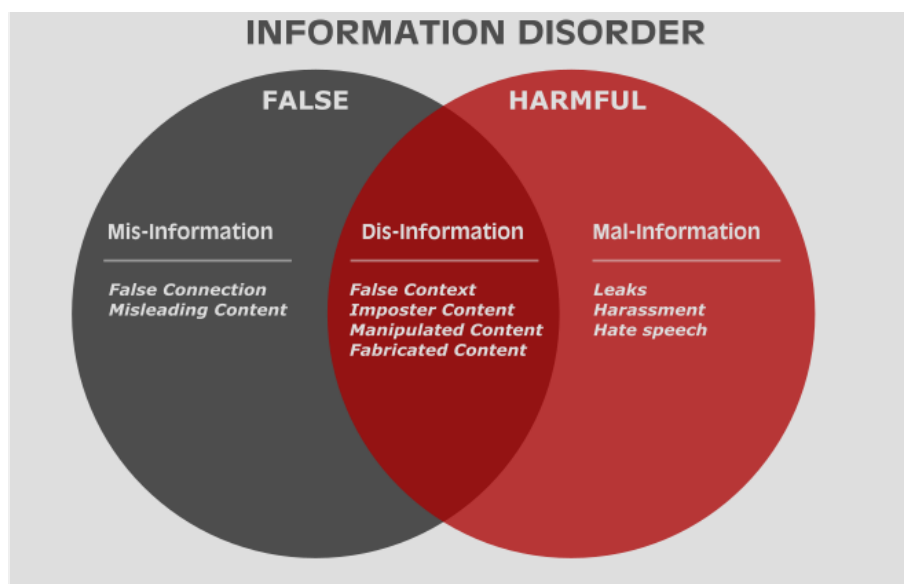


Figure 2.8: An overview of how fake news is defined in the modern era. All different forms of the content: propaganda, lies, conspiracies, rumours, hoaxes, hyper-partisan content, falsehoods, or manipulated media, are grouped under Disinformation, Misinformation and Malinformation. Collectively, called it as Information Disorder.

<sup>2</sup><https://apnews.com/article/technology-europe-russia-hacking-only-on-ap-dea73efc01594839957c3c9a6c962b8a>

## 2.3 How Big the Problem of Fake News is?

We all have witnessed fake news during our time. The section provides examples of fake stories that gained traction among audiences, created a buzz in the online world and have faced repercussions in the offline world.

1. Politics: Fake news is a growing threat to democratic events. We have witnessed the influence of fake news on election outcomes [257]. More specifically, the 2016 US Presidential Election [27, 87], Brexit Referendum and the 2019 Indian General Election [57] are the prime events that raised concerns. After all, voters may base the choice of their vote on incorrect information. In addition, the dissemination of false information online introduced changes in how political campaigns are run, ultimately forcing us to think about the legitimacy of elections.

Next, apart from elections, numerous examples in the history of fake news demonstrate how political parties used fake news to mock their opponents. Instances have also been seen where parties created fabricated stories to frame a positive opinion about them.

2. Health: The spread of health-related misconceptions on social media poses a severe threat to public health [248, 251, 261, 279]. Social media is abused to spread harmful health content, including unverified information about vaccines [35], disseminating unproven and erroneous information about cancer treatments [80] and spreading incorrect advice via rumours about curing HIV and AIDS [243]. Health misinformation reached a milestone during the COVID-19. The massive outbreak of false stories resulted in the declaration of an Infodemic.<sup>3</sup> An infodemic is a piece of information that is false or misleading in the digital and physical world during a disease outbreak. It confused and increased the indulgence of risk-taking behaviour by the masses. It also leads to mistrust in health authorities and undermines the public health response. In the past, such effects have also been observed during the outburst of Ebola [184] and Zika virus [265] epidemics.

3. Climate Change: Fake news has emerged as a quintessential climate problem that media literacy, policymakers, and gyaan pundits are tasked to solve [7, 50, 65, 153, 263]. More specifically, false information can be destructive for areas such as anthropocentric climate change, where understanding the facts and the scientific truth is essential. Climate change misinformation is closely linked to climate change skepticism, denial, and contrarianism [260].

For instance, numerous pieces of scientific evidence prove that the human-caused CO<sub>2</sub> emissions are increasing the temperature of our planet [183]. However, manipulators of fake stories have managed to drift audiences' perceptions with unverified claims, ignoring scientific evidence. Another piece of information shown in the Figure 2.9

---

<sup>3</sup>[https://www.who.int/health-topics/infodemic#tab=tab\\_1](https://www.who.int/health-topics/infodemic#tab=tab_1)



Figure 2.9: An example of false climate change story circulating on the web. The hoax was posted by Natural News claiming NASA declaring that sun responsible for global warming and not the human activities.

claims on social media that NASA has declared that it is the sun and not the human activities responsible for increasing global warming. The news is a hoax and was clarified by NASA later.<sup>4</sup>

4. Entertainment: The intrusion of hyperbolised fake articles into political campaigns or health and climate studies is havoc. However, the recent trend witnessed its presence in the cinematic realm with the release of Gore Verbinski's macabre asylum thriller *A Cure for Wellness*.<sup>5</sup> The marketing team of the movie teamed up with fake news websites to publish a series of false stories that included oblique references to the film and its fictitious realm. The stunt was performed to generate audience interest before the film was released. However, regular news outlets swiftly picked up the fabricated stories, re-purposing them, which generated significant engagement on social media despite being entirely false. In another event, soon after the death of a Bollywood star, Sushant Singh Rajput, a series of fabricated stories led to a creation of a wide range of fabricated stories<sup>6</sup> that started doing the rounds over the Internet [6].

<sup>4</sup><https://www.reuters.com/article/uk-factcheck-nasa-climate-change-idUSKBN2AI2KX>

<sup>5</sup><https://lwlies.com/articles/fake-news-viral-marketing-campaigns-a-cure-for-wellness/>

<sup>6</sup><https://zeenews.india.com/india/zee-news-busts-fake-news-in-sushant-singh-rajput-death-case-2301942.html>

5. Protest (Mass Gathering): The dissemination of fabricated stories has played a crucial role in inflaming or suppressing a social event. In India, some noticeable events saw the rage in the offline world due to increased deceptive activities in the online ecosystem. Propaganda specialists played with the emotion of the masses and created stories that aligned with the beliefs and practices of the audience. Some noticeable events include Phulwama Attack, CAA Act 2019 and the 2020-2021 Indian Farmers' Protest.

To summarize, we pointed out numerous examples from different spheres highlighting the voluminous intrusion of fake stories into human life. Fake news is destructive and can lead to hatred against religion, politics, celebrities or organizations resulting in riots/protests or even death. The destruction due to fake spread occurs at two levels. First, distract the masses from the original issue with a motive to keep it unresolved. Second, to intensify the social conflict to undermine people's trust in organizations and the democratic process.

## 2.4 Multimodality and its Importance

Communication is an act of imparting, transmitting or receiving information. There are various forms and modes by which two entities interact. Concerning written communication, people use different modalities (text, images or videos) to communicate their thoughts. Now the question arises, Can text trump visuals ? or do visuals rule the world? Communication is a complex phenomenon, and there are no two ways about it- thoughtful content and beautiful visuals can help make a piece of information look engaging and grab the audience's attention. Recently, there has been a massive shift in the way people present a message, explicitly shifting towards a combination of text and visuals. It has become easier to comb through an article with images between texts-visuals in the form of gifs, animations and eye-catching infographics. The reasons are three-fold.

1. Research by W. Howard Levie and Richard Lentz [142] proves that people following directions with text and illustrations do 323 percent better than those without illustrations.
2. The human brain works incredibly well in remembering images. Research shows that an individual can retain only 10 percent of the information if asked three days later. In contrast, the number goes to 65 percent if visuals accompany the textual information.<sup>7</sup>
3. From a technical perspective, there have been notable works in deep multimodal learning. Numerous instances have shown that models fusing data from different modalities outperform their uni-modal counterparts. Recently, Hang Zhao and Longbo Huang came up with a study [111] to provide theoretical justifications of how much more accurate an estimate of the latent space representation is when data gets combined from different modalities compared to the singular modality.

---

<sup>7</sup><http://brainrules.net/vision/>

To summarize, different modalities are characterized by different statistical properties. Choosing one over the other does not seem a plausible choice. In order to make progress in understanding the world around us, there is an alarming need to devise technologies that can interpret such multimodal signals together. Hence, in this PhD thesis, we plan to study the domain of fake news, a.k.a Information Disorder, from the viewpoint of multi-modality.

## Chapter 3

# Literature Review

Fake News is a vast domain, and multiple initiatives have been made in various directions to halt its expansion. The objective of the chapter is to provide exhaustive literature on multimodal fake news. Figure 3.1 provides a roadmap of the different sub-domains reviewed in the thesis.

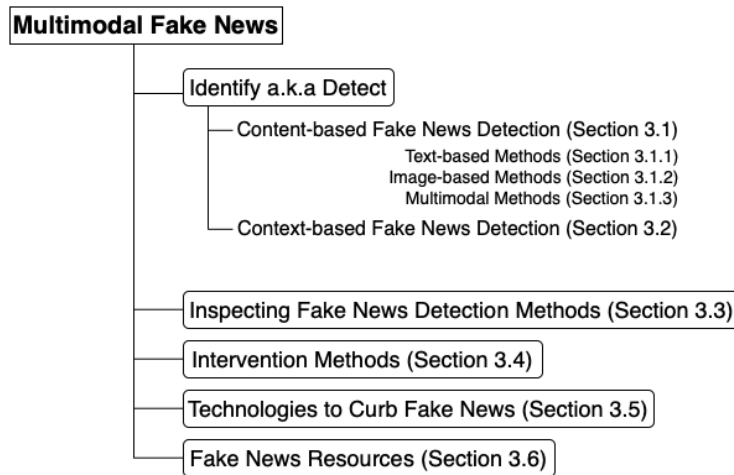


Figure 3.1: A flowchart outlining the Literature Review. We provide an exhaustive literature that addresses sub-fields such as multimodal fake news detection (Section 3.1 and 3.2), validation and robustness (Section 3.3), various intervention strategies (Section 3.4), tools and technologies to curb fake news dissemination (Section 3.5) and list of fake news datasets (Section 3.6).

### 3.1 Content-based Fake News Detection

A news article often has a headline, body of text, leading image, and further supplemental images. This section discusses studies that have looked at one modality or its combinations to identify fake news online.

### 3.1.1 Text-Based Methods

According to psychological theories like the Undeutsch hypothesis [12], one can differentiate between real and fake news by analyzing the linguistic style of the content. This is so because it is hypothesized that fake news consists of opinionated and inflammatory language that derives users' attention. Linguistic-based features can be retrieved from the text content by examining either the words (lexically and semantically), sentences (syntactic level) or documents (discourse level). With the advancement in machine learning, existing research extracted the linguistic features at the word level via counting the frequencies of each word [195, 304], identifying part of speech at the syntactic level [199, 200, 304] and capturing rhetorical relationships at the discourse level [211]. In addition, the textual features can be categorized under general and latent features [49]. General features describe the content style from four language levels: lexicon [304], syntax [74, 195, 304], discourse [124, 211, 304], and semantic [195, 210]. Based on prior study [308], such attributes and their corresponding computational features can be grouped along ten dimensions: quantity [162], complexity, uncertainty, subjectivity [262], non-immediacy, sentiment [313], diversity [262], informality [262], specificity [120], and readability [215]. Such features aid in identifying falsity in computer-mediated communications [77, 301] and testimonies [3] and have recently been used in fake news detection [25, 165, 195, 199, 304]. Latent textual features denote news text embedding. Such an embedding is obtained either at word [166, 191], sentence [16, 139], or document levels [16]. Such vector embeddings are then fed to traditional machine learning [304] or deep learning framework [44, 62, 103, 105, 109, 134, 140, 218, 233, 253] to capture the syntactic meaning of the text. Other works extracted meta features [123], linguistic features [86, 99, 128, 200, 206, 271], applied semi-supervised [88] and unsupervised [107] approach via tensor embeddings to detect fake news via textual cues. Recent research also explored ensemble method [247], reinforcement techniques [277], adversarial attacks [158], performed manipulation detection on web data including but not limited to Wikipedia [136] and devise methods to curb misinformation in low-resource languages [51, 148, 223].

### 3.1.2 Image-based Methods

With the advent of the Internet and the availability of smartphones at nominal prices, user activity on social media platforms has emerged in large scale. The increased user activity has resulted in the proliferation of fake stories online. Fake news attempts to utilize multimedia content with images or videos [144, 145, 293] to attract and mislead readers for monetary or other gains. Different kinds of manipulations can distort images. For instance, images can be deliberately manipulated via tampering, doctoring or photoshopping. It can also be generated automatically by deep generative networks. Other strategies include misusing images to depict the emerging event or portraying a real image with false context—such fabrication in the visual content results in misleading content [24, 299]. Gupta et al. [97] identified faking images based on numerous user-level and tweet-level hand-crafted attributes

using classification framework. In addition, visual features can be categorized into forensics, semantic, statistical, and context [32, 249, 298, 310]. The forensic features capture the distortion within the images and extract the camera-related specifications [75, 83, 159] or detect forgery operations performed on the images [112]. Research has also captured various signal and pixel-level features to identify the forgery performed via deep generative networks [84, 176] and has also devised numerous strategies to identify compression in images. To increase the viewership of the fake story, manipulators often rely on establishing a sensational backing for it. Such fabricated stories play with the emotion/sentiment [117, 137, 229, 250] of the reader. Hence, the literature [37] demonstrates few works that use semantic-level features to determine the post’s authenticity. Peng Qi et al. [203] hypothesize that fake news images might have different properties from real-news images at both physical and semantic levels. Such properties can be studied via the frequency and pixel domain, respectively. Therefore, the team proposed a novel framework Multi-domain Visual Neural Network (MVNN), to fuse the visual information of frequency and pixel domains for detecting fake news. The statistical features explore the distributional difference between real and fake news. Some basic statistical features that aided in detecting fake news were count, popularity and type [119, 284, 289]. Other advanced features proposed by [119] include visual clarity score, visual coherence score, visual similarity distribution headline, visual diversity score and visual clustering score.

### 3.1.3 Multimodal Methods

Jin et al. [116] made the first attempt towards multimodal fake news detection. Their paper proposed a recurrent neural network with an attention mechanism for fake news detection. It comprises of three sub-modules: first, sub-network uses RNN to combine text and social context features. The social context features are hashtags, mentions, retweets, and emotion polarity; Second, sub-network uses VGG19 pre-trained on the Imagenet database to generate representations for images present in tweets; Third, sub-network is a neural-level attention module that uses the output of RNN to align visual features. Yang et al. [294] made another attempt by designing a text and image information based Convolutional Neural Network (TI-CNN). The method extracts latent text and image features, represents them in a unified feature space, and then use learned features to identify fake news.

Another study by Wang et al. [276] proposed an event adversarial neural network for fake news detection. Core idea of the paper is to design a method that learns event-invariant features and preserve the shared features among all the events for fake news detection for newly emerged unseen events. The textual and visual features are extracted via Text-CNN and VGG19, respectively. The final representations are combined to form a multimodal feature vector utilized for fake news detection. In addition, the method uses an event discriminator to measure the dissimilarities among different events; it is a neural network that consists of two fully connected layers with corresponding activation functions. Khattar et al. [126] also came up with multimodal variational autoencoder for fake news detection. Model comprises



of three components: (i) encoder, responsible for generating the shared representation of features learnt from both the modalities, (ii) decoder, responsible for reconstructing data from the sampled multimodal representation and, (iii) fake news detector, that takes multimodal representation as input and classify the post as fake or not.

Another study attempted to detect fake news by leveraging spatial and frequency domain features from the image and textual features from the text present in a news [287]. Method uses multiple co-attention layers to learn the relationship between text and images. Visual features are first fused, followed by textual features; obtained fused representation from the last co-attention layer is used for fake news detection. Another attempt by Liu et al. [151] forgoes the standard vanilla fusion method and designs a method that combines the information from the images and text by introducing an image caption-based method. The proposed method can integrate the image description information into the text to bridge the semantic gap between text and images.

All the works mentioned above have focused on multimodal fake news detection ignoring the relationship between textual and visual cues present in news articles. Zhou et al. [307] proposed a similarity-aware fake news detection method to investigate relationship between the extracted features across modalities. Text features are extracted via Text-CNN, and image feature generation is a two-step process. First, images are passed through the image2sentence model to generate a caption for the image. Generated text is then passed through Text-CNN to get the desired representations. A modified version of cosine similarity is used to establish a cross-modal relationship between the modalities. Recently, [202] designed EM-FEND that captured the cross-modal correlations: entity inconsistency, mutual enhancement, and text complementation.

In the year 2020, a study by Giachanou et al. [82] introduced a new direction by exploiting the information from multiple images in accordance with the headline and the complementary *first image*. Giachanou et al. [82] proposed a multimodal multi-image module that encapsulates information from multiple images in the form of tags and semantic features via a pre-trained VGG-16 network. Next, to establish similarity between the different components of the two modalities, cosine similarity score is calculated between the text and image tags. Finally, textual and visual feature vectors are combined with the similarity score, in an additive manner to perform fake news detection. Recent developments in the area includes the use of external knowledge in the form of text-metadata [78], detect cross-modal inconsistencies [270], and inspect the connection between text and image [182].

### **3.2 Context-based fake News Detection**

The dissemination of fake news online is influenced by various reasons, including but not limited to user-based, time-based, and network-based variables. The field of study that encompasses additional factors in tandem with the headline, content and images forms context-based fake news detection. This section reviews prior work on multimodal fake news

identification while encapsulating such additional features.

User engagement is an influential factor that drives the dissemination of fake news online. It widens the research as various factors could be studied to identify the manipulation. Past literature explored methods to identify the source/publisher [19, 34, 45, 93, 230, 300], post [38, 171, 297] and user authenticity [41, 45, 59, 66, 67, 171, 179, 201, 212, 232, 273, 286, 297, 305] and curated the social features, either as a strong [94, 154, 179, 219, 271, 281] or weak signal [231]. Researchers have also explored other auxiliary features like, user comments [54, 225], leveraging information from multiple news [121], exploring the propagation networks [205] or incorporating relational knowledge [272, 285]. Some methods have also studied the changes or trends of the properties along the life-cycle [156]. Research has also explored the methods that leverages crowd [118], model the problem via graph-based methods [100, 114, 154, 164, 246, 306], perform agent-based modelling [81] or used of external knowledge [53, 108, 224, 278] to enhance the feature-set. In addition, the user and network properties has introduced new directions of propagation-based [152, 228] and network-based methods. Such domain aims to explore the news dissemination pattern in the online world and also study the network properties of the platform to draw inferences about the spread of fake news [60, 221]. Recent advancements toward fake news detection explore the unsupervised pathway [90, 292] due to the non-availability of the data.

Further, content-driven approaches can benefit significantly from context-driven approaches by enhancing their overall effectiveness in identifying and combating fake news. Here are a few ways in which content-driven approaches can leverage contextual information:

- **Improved Feature Selection:** Contextual features can be integrated as additional input into content-driven algorithms. This not only adds more dimensions for analysis but can also help in feature selection by identifying which contextual factors are most informative in distinguishing fake news. Additionally, incorporating contextual information like the author's reputation and publisher's credibility can improve the accuracy of fake news detection. Content analysis alone may not provide a comprehensive understanding of the news item, but when combined with context, the system can make more informed decisions.
- **User Trust:** Content-driven approaches can be improved by including context-driven features, which can enhance user trust in the system. Users are more likely to trust a system that can provide the reasons or sources behind its fake news classification.
- **Robustness:** Context-driven approaches can help content-driven methods become more robust. Fake news often relies on manipulating content in a way that might not be immediately obvious. By considering the context in which the news is published, the system can identify discrepancies and inconsistencies

### 3.3 Validation and Robustness of Fake News Detection Methods

Fake News Detection is not as easy as it seems to be. It has been the subject of extensive investigation. However, the lack of transparency in decision-making is the prime issue with fake news detection methods. It would be intriguing to inspect the patterns that the algorithm uses to ascertain how reliable or generalizable these patterns are. The most current developments in establishing robustness, generalizability, and fair fake news detection algorithms [187] are covered in this section.

Yang et al. [291] designed a fake news detection system that not only predicts the veracity of the information but also provides relevant explanations as prediction evidence. Another attempt by [252] proposed a novel explainability mechanism in the BERT-based fake news detectors. The methods used Local Interpretable Model-Agnostic Explanations (LIME) [169] and Anchors to perform the desired task. Brien et al. [181] attempted to investigate the generalizability capability of the fake news detectors. The author performs various analyses concluding that the model can identify subtle but consistent differences between the languages of the two types. Another attempt by [309] proposed ENDEF (entity debiasing framework) to mitigate the entity bias present in the fake news datasets that influences the generalizability ability of future data. Wu et al. [283] also attempted to investigate the generalizability of the fake news detectors by proposing a novel framework for debiasing evidence via causal intervention methods. Murayama et al. [173] on the other hand, exploited the fake news datasets, identified diachronic bias, and proposed strategies to mitigate it. The paper proposed masking methods using Wikidata that perform better for the out-of-domain datasets. The strategy is claimed to make the model more resistant to these biases. Another attempt by [302] also took a dig into numerous fake news datasets. The paper closely examines the collection and data split mechanism and points out flaws in the current approaches. It also provides suggestions for creating high-quality news datasets. Park et al. [186] attempts to study the fairness of the misinformation classification algorithms. The paper used a post-processing approach, Reject Option Classifier, to mitigate the bias while maintaining accuracy. On the other hand, Bozarth et al. [29] examined a subset of fake news detectors via performance evaluation and error analysis steps. The paper points out several factors that determine the performance of the models. Furthermore, simple evaluation metrics like accuracy or F1 scores are insufficient to evaluate and compare different methods. Hence, the paper highlights the need for systematic benchmarking to evaluate the performances of fake news detectors. Another attempt by Bozarth et al. [31] provides a comprehensive overview of fake and real news websites. The authors noticed websites generally have different labelling processes and cater to news from specific domains. They also highlight how selecting the preferred ground truth significantly affects fake news detection.

### 3.4 Intervention Methods

The proliferation of fake news on social media is one of the most prominent issues concerning society today. As a result, researchers and practitioners are interested in investigating the behaviour of online readers towards fake news. The branch of science that measures human behaviour is known as behavioral research. It uses qualitative and quantitative methods to examine and comprehend individual and societal behaviour. This section discusses the advancements done towards behavioural research for fake news to design effective interventions for social media platform owners.

Pennycook et al. [193, 194] conducted a study to understand why people share fake news and suggested solutions to reduce the sharing behaviour. The study infers that the veracity of news has minimal effect on the sharing intentions shown by the users. Moreover, nudging users to think about accuracy reduces the sharing behaviour. To this, Jahanbakhsh et al. [113] designed lightweight interventions to nudge users to assess the accuracy of the news before sharing it. Another study by Lin et al. [147] uses drift-diffusion modelling to investigate the role of accuracy prompts in decreasing fake news sharing. Epstein et al. [73] investigate how the perceived accuracy of a user towards a news changes when making a sharing decision. To test the applicability of accuracy prompts in a real-world scenario, an ongoing work by Guay et al. [89] suggests that Republicans are more prone to share ideologically concurrent fake news than Democrats. Another work by Epstein et al. [70] examines different approaches for accuracy prompts to find the most effective one.

The volume of fake news generated online is massive. An effective way to halt its dissemination is to use crowdsourcing efforts to identify fake news—one such attempt by Epstein et al. [72] aims to use the judgements by laypeople to combat misinformation. The authors conducted a survey where users judge the veracity of the news domains. Based on the rating by the users, the newsfeed algorithm would down-weight the content from such websites, and such content would then be less likely to be displayed on the platform. Another effort by Allen et al. [11] harnesses the efforts of laypeople to scale up the fact-checking process. Another intervention tested in the research is to attach label warnings to news stories. However, a potential consequence of such social corrections has been observed in the subsequent sharing behaviour of the users [172, 192]. In addition, Bashier et al. [33] find out that timing matters when correcting misinformation. Another study [71] examined whether providing how the labelling mechanism works boosts the effectiveness of the warnings. Whereas Martel et al. [160] conducted studies to examine the content of corrective messages.

### 3.5 Technologies to Curb Fake News

1. TweetCred [96]: It is a real-time, web-based system that evaluates the credibility of information on Twitter. The system utilizes a semi-supervised ranking model using SVM-rank to determine the authenticity of the post on users' timelines. The training data constitutes six high-impact crisis events of 2013. An extensive set of 45 features is used to determine each tweet's credibility score. The system provides a credibility rating between 1 (low) to 7 (high) for each tweet on the Twitter timeline. All features can be computed for a single tweet, including the tweet's content, characteristics of its author, and information about external URLs.
2. CredEye [197]: It is a system for automatic credibility assessment. Input in the form of a claim is analyzed for its credibility by considering relevant articles from the Web. The system is composed of three units. The first unit is responsible for retrieving articles from web sources by searching claim text as a query to a search engine. The second component performs the stance detection task to understand an article's stance. The Credibility aggregation model then merges per-article assessments to compute the overall scoring of the claim being true or false. Finally, the evidence extraction module extracts the supporting evidence from informative snippets from the relevant web articles. The training data for the task is curated from Snopes.com. The system utilizes 5,000 claims from Snopes, each labelled true or false and retrieved 30 relevant Web articles for each of them.
3. Real-Time Certification System on Sina Weibo: Zhou et al. [303] presents a real-time news certification system that can detect an event's credibility by providing keywords about it. The authors built a distributed data acquisition system to enable real-time data flow to gather event-related information through Sina Weibo. The average response time for each query is 35 seconds—this paper model the rumour detection problem from three aspects: content, propagation and information source. The content-based model leverages the event, sub-events and message information. In contrast, the propagation-based model captures propagation network influence. Finally, an ensemble method is utilized to capture the three aspects. In addition, to determine the credibility of an event, the system also provides information such as key users, key microblogs and the timeline of an event.
4. ClaimBuster [102]: It is a fact-checking platform that uses natural language processing and supervised learning techniques to determine factual claims in political discourses. The model is trained on a human-labelled dataset of check-worthy factual claims from the U.S General Election debate transcripts. The system performs the claim spotting task, giving each sentence a score indicating how likely it is to contain an essential factual claim that should be fact-checked. ClaimBuster helps fact-checkers to focus on the top-ranked sentences without searching through a large number of sentences.

5. XFake [291]: It is a fake news detection system that predicts the veracity of the information and provides relevant explanations as prediction evidence. The system comprises of three components that utilize the speaker and statement attributes. Specifically, MIMIC, ATTN and PERT frameworks are designed, where MIMIC is built for attribute analysis, ATTN is for statement semantic analysis, and PERT is for linguistic statement analysis. Explanations, supporting examples and visualization are provided to facilitate interpretation of the output.
6. Jennifer [146]: It is a chatbot maintained by a global group of volunteers. Building such a system aims to provide public information from trusted sources in an organized and efficient manner. Such information can be utilized during a crisis event or in general by the masses to understand public issues. The group also released a dataset, the COVID-19 question bank, consisting of 3,924 COVID-19-related questions.
7. PRTA [55]: PPropaganda persuasion Technique Analyzer, is a system that detects propaganda in text fragments. It also provides readers with information on what type of propaganda technique is used. The system attempts to promote media literacy among online audiences. With Prta, users can explore the contents of articles about several topics, crawled from various sources and updated regularly, and compare them based on their use of propaganda techniques. Prta is designed as a supervised multi-granularity gated BERT-based model, trained on a corpus of news articles annotated at the fragment level with 18 propaganda techniques, a total of 350K word tokens.
8. BRENDA [26]: BRENDA is a browser extension that aids in fact-checking. BRENDA performs two tasks- it identifies the fact-worthy claims and verifies the veracity of such claims from the online evidence. BRENDA uses SADHAN [168] to classify fake news.
9. BirdWatch [48]: Birdwatch is a community-based platform designed by Twitter that enables individuals to flag information in tweets they believe is misleading and write notes that provide informative context. The project is currently in the pilot stage. Notes are only visible on a separate Birdwatch site where pilot participants provide feedback on tweets and rate the helpfulness of feedback (notes) added by other contributors.
10. WhatsFarzi [135]: An app that can identify fake news on various instant messaging platforms like WhatsApp, Telegram and Hike. The backend system in the app utilizes text and images to identify the veracity of the claim. From the text, entities are extracted, which are then stored as a knowledge graph. For the images, state-of-the-art image tampering algorithms are used to detect parts of the image that have been doctored.

### **3.6 Fake News Resources**

Researchers and practitioners have proposed numerous resources to facilitate research on fake news. This section reviews the fake news and fact-checking datasets from the two viewpoints.

1. News articles: Fabricated content takes various forms. It can be published as news on online news websites consisting of a headline, content (body of the news) and images/videos associated with it. Table 3.1 review the datasets utilized to detect fake news mainly from the body of the news article. The style of each news article is an essential feature for detection.
2. Social Media Posts: Another pattern of fabrication is in the form of memes or tweets/posts on different social media platforms. Table 3.2 and 3.3 review datasets that are utilized to detect fake news, mainly from social media posts. User and network information in social media and text in social media posts are essential features.

Dataset	Instances	Labels	Topic Domain	Raters	Language	Year
PolitiFact14 [268]	221 headlines	5	Politics, Society	Fact-checking sites	English	2014
Buzzfeed_political [106]	71 articles	2	2016 US election	Buzzfeed page	English	2017
Random_political [106]	225 articles	3	Politics	List of Zimdars	English	2017
Ahmed2017 [4]	25,200 articles	2	News in 2016	Fact-checking site (PolitiFact)	English	2017
LIAR [9, 275]	12,836 claims	6		Fact-checking site (PolitiFact)	English	2017
TSHP-17_politiFact [204]	10,483 statements	6		Fact-checking site (PolitiFact)	English	2017
NELA-GT [180]	713K articles	2		NewsGuard, Pew Research Center, Wikipedia, OpenSources, MBFC, AllSides, BuzzFeed, PolitiFact	English	2018
FakeNewsAMT [195]	480 articles	2	Sports, Business, Entertainment, Politics, Technology, Education	Generated fake news by Crowdsourcing	English	2018
Celebrity [195]	500 articles	2	Celebrity	Fact-checking site (GossipCop)	English	2018
Kaggle_UTK	25,104 articles	2			English	2018
MisinfoText_Buzzfeed [259]	1413 articles	4		Fact-checking site (Buzzfeed)	English	2019
MisinfoText_Snopes [259]	312 articles	5		Fact-checking site (Snopes)	English	2019
FA-KES [213]	804 articles	2	Syrian War	Expert annotators	English	2019
Spanish-v1 [198]	971 articles	2	Science, Sport, Politics, Society, Environment, International	Fact-checking sites (VerificadoMX, Maldito Bulo, Caza Hoax)	Spanish	2019
Fauxtography [310]	1,233 articles	2		fact-checking site (Snopes)	English	2019
Breaking! [188]	679 articles	3	2016 US election	BS Detector	English	2019
TDS2020	46,700 articles	2		News sites (BreiBart, The Onion, InfoWars)	English	2020
FakeCovid [220]	12,805 articles	2-18			English	2020
TrueFact_FND	6,236 articles	2			English	2020
Spanish-v2 [198]	572 articles	2	Science, Sport, Politics, Society, Environment, International	Fact-checking sites (VerificadoMX, Maldito Bulo, Caza Hoax)	English	2021

Table 3.1: List of datasets for detecting fake news in news articles. The table highlights datasets published between the years 2014 to 2021. Empty cells indicate that the information is not available.



Dataset	Instances	Labels	Topic Domain	Raters	Platform	Language	Year
MediaEval_Dataset [23]	15,629 posts	2			Twitter, Facebook, Blog Post	English	2015
PHEME [312]	330 threads	3	Society, Politics	Crowdsourcing	Twitter	English	2016
Twitter-ma [155]	992 threads	2		Fact-checking site (Snopes)	Twitter	English	2016
RUMDECT [155]	4,664 threads	2		Sina community management	Weibo	Chinese	2016
RumorEval2017 [61]	297 threads	3		PHEME [312] [311]	Twitter	English	2016
Twitter15 [157]	1,478 threads	4		Fact-checking sites (Snopes, Emergent)	Twitter	English	2017
Twitter16 [157]	818 threads	4		Fact-checking sites (Snopes, Emergent)	Twitter	English	2017
BuzzFace [214]	2,263 threads	4	Politics	Buzzfeed	Facebook	English	2017
Some-like-it-hoax [254]	15,500 posts	2	Science	[38]	Facebook	English	2017
Media_Weibo [116]	9,528 posts	2		Sina community management	Weibo	Chinese	2017
PHEME_update [132]	6,425 threads	3	Society, Politics	PHEME [311]	Twitter	English	2018
FakeNewsNet [227]	23,921 posts	2	Politics, Celebrity	Fact-checking sites (Politifact, GossipCop)	Twitter	English	2018
Jiang2018 [115]	5,303 posts	5		Fact-checking sites (Politifact, Snopes)	Twitter, Youtube, Facebook	English	2018
RumorEval2019 [85]	446 threads	3	Natural disasters	Fact-checking sites (Politifact, Snopes)	Twitter, reddit	English	2018
Rumor-anomaly [255]	1,022 threads	6	Politics, Fraud and Scam, Crime, Science, etc.	Fact-checking site (Snopes)	Twitter	English	2019
WeChat_Dataset [277]	4,180 news	2		WeChat	WeChat	English	2020

Table 3.2: List of datasets for detecting fake news on social media. The table highlights datasets published between the years 2015 to 2020. Empty cells indicate that the information is not available.

Dataset	Instances	Labels	Topic Domain	Raters	Platform	Language	Year
Fang [179]	1,054 threads	2		PHEME [132], Twitter-ma [23], FakeNewsNet [227]	Twitter	English	2020
WhatsApp [207]	3,083 images	2	Brazilian Elections, Indian Elections	Fact-checking sites (aosfatos.org, boomlive.in, e-farsas, etc.)	WhatsApp		2020
Fakeddit [174]	1,063,106 posts	2,3,6		Expert annotators	Reddit	English	2020
Reddit_comments	12,597 claims	2		Fact-checking sites (PolitiFact, FactCheck.org, etc.)	Twitter	English	2020
HealthStory [56]	1,690 threads	2	Health	HealthNewsreview	Twitter	English	2020
HealthRelease [56]	606 threads	2	Health	HealthNewsreview	Twitter	English	2020
CoAID [52]	4,251 threads	2	COVID-19	Fact-checking sites (PolitiFact, FactCheck.org, etc.)	Twitter	English	2020
COVID-HeRA [63]	61,286 posts	5	COVID-19	CoAID, Expert Annotators	Twitter	English	2020
ArCOVID-19-Rumors [101]	162 threads	2	COVID-19	Fact-checking sites (Fatabyyano, Misbar)	Twitter	Arabic	2020
MM-COVID [143]	11,173 threads	2	COVID-19	Fact-checking sites (Snopes, Poynter)	Twitter	English, Spanish, Portuguese, Hindi, French, Italian	2020
Constraint [189]	10,700 posts	2	COVID-19	Fact-checking sites (PolitiFact, Snopes)	Twitter	English	2020
Indic-covid [122]	1,438 posts	2	COVID-19	Expert annotators	Twitter	Bengali, Hindi	2020
COVID-19-FAKES [69]	3,047,255	2	COVID-19	WHO, UN, UNICEF	Twitter	Arabic, English	2020
CHECKED [288]	2,104 threads	2	COVID-19	Sina community management	Weibo	Chinese	2021
COVID-Alam [8]	722 tweets	5	COVID-19	Expert annotators	Twitter	English, Arabic	2021
COVID-RUMOR [42]	2,705 posts	2	COVID-19	fact-checking sites (Snopes, Politifact, Boomlive)	Twitter, Websites	English	2021

Table 3.3: List of datasets for detecting fake news on social media. The table highlights datasets published in the years 2020 and 2021. Empty cells indicate that the information is not available.

## Chapter 4

# Designing Simple Baselines for Multimodal Fake News Detection

This chapter is partly a reproduction of papers published at the IEEE Multimedia Big Data (BigMM) 2019 [235] and the Association for the Advancement of Artificial Intelligence (AAAI) 2020 [237].

### 4.1 Introduction

The phenomenon of fake news has a long-standing history, and its detrimental impact on society has elevated it to a significant concern that the research community is actively striving to tackle. The term has become jargon, but how it is defined differs from the earlier studies. Earlier, any distinct content such as satires, hoaxes, news propaganda and clickbaits [40, 58] was termed as fake news. However, Allcott et al. [10] defined fake news as “news articles that are intentionally and verifiably false, and could mislead readers.” Moreover, such content is written to deceive someone. An example of such a false story is shown in Figure 4.1.



Figure 4.1: An example of fake news that claims that the actor Sylvester Stallone died due to prostate cancer.

Interestingly, the image shown in the news is photo-shopped to make it look similar to the news that is generally featured on popular news channels like CNN, BBC, and many others. The image made people believe that the news is real, but the news was later quashed by the victim himself (see Figure 4.2).



Figure 4.2: The reply from the actor Sylvester Stallone after fake news of his death spread.

There can be various reasons for the spread of fake news. The first one could result from a general lack of knowledge. The readers should acquire knowledge about the credibility of the sources [39, 95] and abilities for judging the truthfulness of the news. The second factor could be the failure of fact-checking initiatives to make it to the general public. Websites such as Politifact<sup>1</sup>, Full Fact<sup>2</sup> and AltNews<sup>3</sup> attempt to detect fake news, but their efforts go in vain. Additionally, false content that does not make it up to the eyes of fact-checking is considered authentic by the online readers [172, 192]. Third, there might be a case that manipulations are challenging for general AI models to spot. For instance, the text in Figure 4.3 (ii) says, “the presence of sharks during Hurricane Sandy 2012,”, whereas deep analysis of the image concludes that it was spliced to show fake sharks in the image. Similarly, the text in Figure 4.3 (iii) says, “picture of Solar Eclipse captured on March 20, 2015.” However, the image is an artwork done so beautifully that it is hard to distinguish from reality.

Examples mentioned above and numerous others found online share the trait of fusing multiple modalities. Moreover, the narrative style of news on social platforms differs from those on online websites. The word count bounds the content on social media; hence users express their thoughts via memes, images or videos in tandem with the text description. This necessitates the need to process multimodal data online to verify its veracity. Studying information constituting multiple modalities would be beneficial because (i) different modalities exhibit different aspects of news, (ii) information derived from different modalities complements each other in detecting the authenticity of the news, (iii) different sources

<sup>1</sup><http://www.politifact.com/>

<sup>2</sup><http://www.fullfact.org/>

<sup>3</sup><https://www.altnews.in/>

manipulate different modalities based on their expertise (e.g., some people have experience in creating fake news by manipulating images and others may have experience in manipulating modalities such as text, audio and videos), and (iv) since real-world texts, photos, and videos are complex, contextual information is also essential in addition to content information.

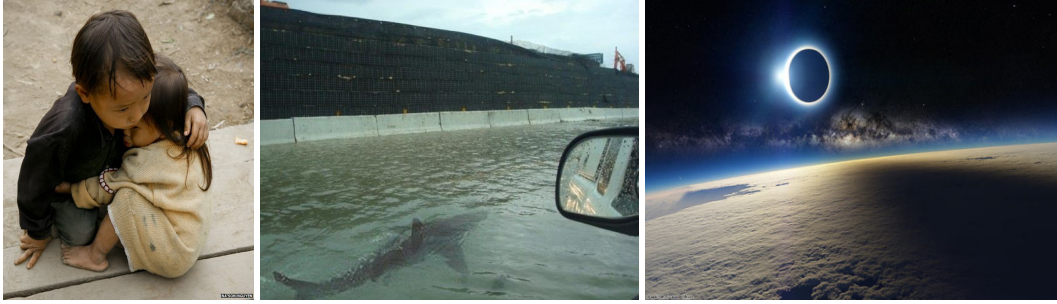


Figure 4.3: (i) real photo of two Vietnamese siblings but being presented as it was captured during the Nepal 2015 earthquakes; (ii) photos of spliced sharks taken during Hurricane Sandy in 2012; (iii) a beautiful artwork portrayed as a picture of Solar Eclipse of March 20, 2015.

As a result, in this Chapter, we study the field of fake news to devise effective baselines for fake news identification. We discuss our three primary contributions, each of which addresses three issues. First, we conduct survey research on a sample population to confirm the necessity of multimodal signals. Next, we adopt advanced deep learning and vision-based algorithms to design simple baselines for multimodal fake news detection. Lastly, we talk about another approach that can accurately spot false information in lengthy publications.

## 4.2 Research Objective

1. **Efficacy of Multiple Modalities:** Our first core contribution is to carry out a survey to understand how individuals perceive fake news. We also examine whether having multiple modalities makes it easier for people to spot fake news (Section 4.3).
2. **Designing Simple Multimodal Baseline:** In 2017, Jin et al. [116] made the first attempt toward multimodal fake news detection. The paper proposed a content-based multimodal fake news detection method that uses a recurrent neural network with an attention mechanism to combine text and social context features. It uses VGG-19 [234] pre-trained on the Imagenet database to generate representations for images present in tweets. Several other works discussed in Section 3.1.3 performed similar tasks. However, we did find contradictions in the existing literature. First, none of the approaches extracts contextual information from the text. Each method captures the syntactic and semantic features of the text. Second, there are multimodal fake news detection systems in the literature, but they solve the fake news problem by considering an additional sub-task like an event discriminator [276] and finding correlations across the modalities [126]. The results of fake news detection are heavily dependent on the

subtask, and in the absence of subtask training, the performance of fake news detection degrades by 10% on an average. Hence, our second core contribution is to propose a solution that detects fake news without taking into account any other subtasks. We present *SpotFake- a multimodal framework for fake news detection*<sup>4</sup> that exploits both the textual and visual features of an article. Additionally, SpotFake excels at managing short-length news stories (Section 4.4).

3. **Handling Long-length Content:** In recent years, there has been a substantial rise in news consumption via online platforms. The ease of publication and lack of editorial rigour in such platforms have further led to the proliferation of fake news. Hence, we introduce *SpotFake+, a multimodal approach that leverages transfer learning*<sup>5</sup> to capture semantic and contextual information from the long-length news articles and its associated images and achieve better accuracy for fake news detection (Section 4.5).

### 4.3 Does Using Multiple Modalities Help Identify Fake News?

The survey aims to learn how individuals view false news and to look for any potential sources that may be contributing to its spread. We also examine whether having multiple modalities makes it easier for people to spot fake news. The survey is divided into three components. The first series of questions aims to comprehend users’ general news-consuming habits. The second batch of questions probe users’ awareness of fake news and their vigilance towards it. The final set of questions assesses how well users can identify fake news and what aspects of the content drive their decisions.

The survey was conducted in June, 2018 and consists of 20 questions (See Appendix A for actual questions). In total, we collected responses from 89 participants. We display the demographics of the survey respondents in Table 4.1. We include demographic questions such as the age and gender of the participants. We carried out data collection using Google Forms.

Gender	Female	37.1
	Male	62.9
Age	<21	31.5
	21-29	65.2
	30-39	1.1
	>41	2.2

Table 4.1: Demographics of the participants in the survey. Values in the table are in percentage.

<sup>4</sup>At the time of writing the thesis, SpotFake had 230+ citations.

<sup>5</sup>At the time of writing the thesis, SpotFake+ had 80+ citations.

### 4.3.1 Observations from Survey

In this section, we present the main findings of the survey.

#### What is the general news consumption habit of the users?

We query the participants on the following to analyse consumption patterns: (i) what are the primary sources of users to consume news? (ii) what medium does a user find most trustable when consuming news? Furthermore, (iii) how frequently does a user discuss the worlds happening in their immediate social circle?

**Observation 1:** As shown in Figure 4.4, about 64% of the users consume news via traditional sources like TV, radio, newspaper and news app on smartphones. About 33.7% of the users consume news via social media, and only a small portion of consumers use WhatsApp to read the news. Further, when asked which news sources they trusted, traditional media received 91 percent of the respondents' votes. This demonstrates that some consumers rely on conventional sources but are still reading news from internet sources, which is something to keep an eye out for in the future.

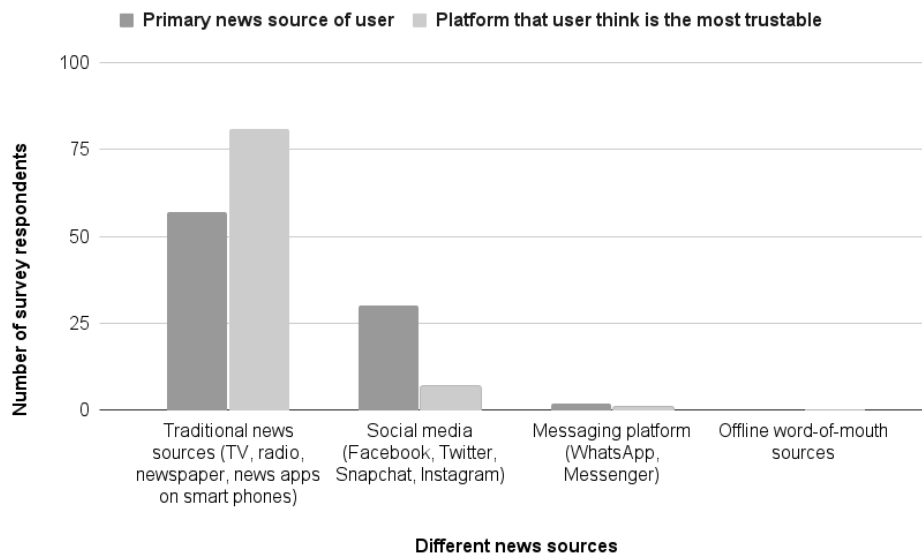


Figure 4.4: What were user responses towards (i) what are the primary sources of users to consume news and (ii) what medium does a user find most trustable when consuming news. We find that about 64% of the users consume news via traditional sources and 33.7% via social media. About 91% of the respondents trust traditional news sources for news consumption.

**Observation 2:** Additionally, we got a mixed response on the engagement activity of the users with their peers. As per the statistics shown in Figure 4.5, roughly 24.7% users less

often engage in news discussion. At the same time, only 23.6% of those surveyed discuss it daily.

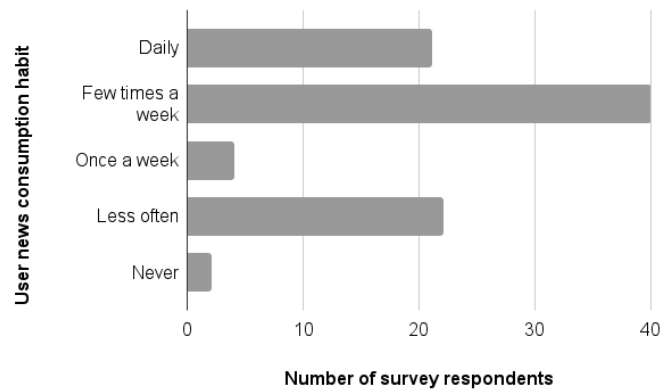


Figure 4.5: How frequently a user discusses the worlds happening in their immediate social circle. We find that only a small % of users, about 23.6%, talk about news daily.

### Inspecting Users' Awareness and Their Vigilance Towards of Fake News

We ask the participants the following questions to gauge their understanding of fake news. We start by asking if they comprehend the term *fake news*, if they can distinguish between true and false, and, more importantly, if they even care whether the news they read is accurate. Then, we also asked them what they thought contributed to the spread of false information online or in conventional venues. Additionally, which platform do they encounter the greatest volume of fake news being shared?

**Observation 3:** We observe that about 97.8% of respondents to our study said they were concerned about the veracity of the news they read. As shown in Table 4.2, only 75.3%  $[(31 + 8 + 29)/89]$  of participants understood the term fake news; among them, only 34.8%  $[31/89]$  were confident in their ability to tell the difference between true and false news.

	Can you differentiate between Real and Fake?			
		yes	no	maybe
	Do you			
	undertand			
	the term			
Fake News	yes	31	8	28
	no	0	6	1
	maybe	4	1	10

Table 4.2: The confusion matrix shows user responses towards understanding the term *fake news* and their ability to distinguish between the same. We received a total of 89 responses in the survey. We found that 75.3%  $[(31 + 8 + 29)/89]$  of participants understood the term fake news; among them, only 34.8%  $[31/89]$  were confident in their ability to tell the difference between true and false news.



**Observation 4:** Regarding the causes of the transmission of false information over the Internet, as opposed to traditional ways, about 53.9% of users believe that the intention of manipulators who share false information online is to spread propaganda against a person or an organization (see Figure 4.6). However, the motivation for disseminating misleading information via conventional methods differs slightly. Here, some participants think the goal is to change the reader’s perspective on a particular subject or point of view (see Figure 4.7).

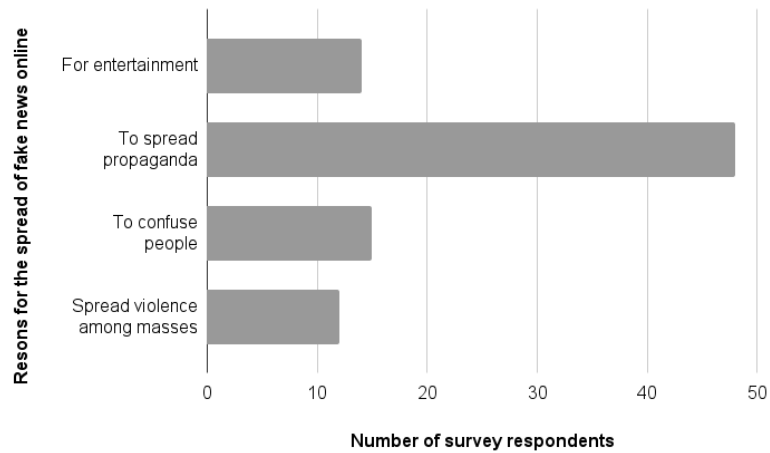


Figure 4.6: We asked what users believe to be the driving force behind the spread of false information online. We found that 53.9% of users believe that manipulators who share false information online intend to spread propaganda against a person or organization.

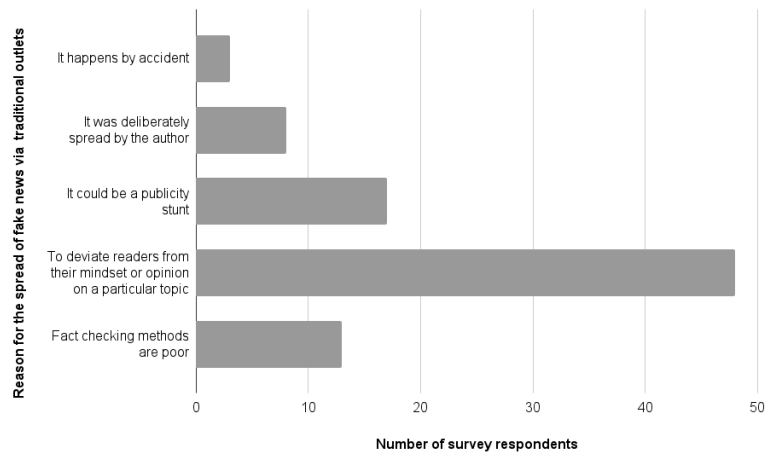


Figure 4.7: We asked what users believe to be the driving force behind spreading false information on traditional platforms. We found that participants think the goal is to change the reader’s perspective on a particular subject or point of view.

### Can users identify fake news effectively?

In the last segment of the questionnaire, we present a few samples to the participants and ask them to mark the veracity of the news. We also asked them which part of the news they focused on while reading and which one they thought would be fake. Further, what aids in detection? Is it the single modality or multiple modalities?

**Observation 5:** About 64% of the users focused on a combination of text and images to assess the veracity of the news. Whereas only 22.7% of the users considered headlines, roughly 4.5% considered only images.

**Observation 6:** The majority of respondents, i.e. 89.9% of those surveyed, thought detecting fake news was aided by multiple modalities.

#### 4.3.2 Inferences

1. From observation 1, we find that a negligible percentage of users consume news via WhatsApp. However, the prevalence of fake news via groups on WhatsApp has unleashed profound violence in the country.<sup>6</sup> We observe a mismatch in how people perceive their engagement online vs their actual behaviour. Therefore, creating behavioural studies that pinpoint and examine such inconsistencies could be helpful.
2. We learned through observation 2 that the number of users interested in knowing and discussing current affairs is low. This raises severe concerns since it increases the likelihood of someone believing fake news. We think that the government, educational institutions, and business organizations should encourage people to read about current events.
3. Many readers are concerned about the accuracy of the news they read, yet many are unable to spot the difference between real and fake news. It would be intriguing to study the characteristics of fake news, the factors that led to its creation, and how it has evolved through time. Learning such patterns would help us uncover reasons that make identifying fake news challenging for AI models and laypeople.

#### 4.3.3 Discussion

In this section, we covered the specifics of the survey, which sought to learn how people perceive false information and identify potential sources that might facilitate its dissemination. We also investigate if the availability of several modalities makes it simpler for individuals to recognise fake news. Here, we conclude that readers care about the accuracy of the news they read and that various modalities help readers comprehend and determine the veracity of the information. Next, we carefully review the literature on multimodal false news, concluding

---

<sup>6</sup><https://tech.hindustantimes.com/tech/news/inside-whatsapp-s-fake-news-problem-in-india-when-rumours-lead-to-killings-story-3BXP5eXAFb19aDEMrGm22M.html>

the aforementioned observations and focusing towards our second core contribution, i.e. designing simple baselines to effectively identify fake news via leveraging multiple signals.

## 4.4 SpotFake: A Multi-modal Framework for Fake News Detection

The rapid growth in fake news on social media is a very serious concern in our society. It is usually created by manipulating images, text, audio, and videos. This indicates a need for a multimodal system for fake news detection. Though there are multimodal fake news detection systems, but they solve the fake news problem by considering an additional sub-task. For instance, Wang et al. [276] built an end-to-end Event Adversarial Neural Networks (EANN) model for multimodal fake news detection, consisting of two components. The text part took a word embedding vector as input and generated text representation using a CNN [129]. Image representations are extracted from the VGG-19 model pre-trained on ImageNet [234]. Finally, the obtained representations are concatenated and fed in two fully connected neural network classifiers to perform event discriminator and fake news classification tasks. Inspired by [276] architecture, Khattar et al. [126] built Multimodal Variational Autoencoder (MVAE) model for fake news detection. The primary task of the method is to build an auto-encoder-decoder model, whereas the secondary task focuses towards fake news detection. The method utilizes bi-directional LSTMs and VGG-19 to extract text and image representations, respectively. The latent vectors produced by concatenating these two vectors are fed into a decoder for reconstructing the original samples. The same latent vectors are also used for the secondary task of fake news detection.

Though these multimodal systems perform well in detecting fake news, the classifiers have always been trained in tandem with another classifier. This increases training and model size overhead, increases training complexity and at times can also hinder the generalizability of the systems due to lack of data for the secondary task. To solve this issue, we introduce SpotFake- a multi-modal framework for fake news detection. Our proposed solution detects fake news without taking into account any other subtasks. It exploits both the textual and visual features of an article. Specifically, we made use of language models (like BERT) to learn text features, and image features are learned from VGG-19 pre-trained on ImageNet dataset. All the experiments are performed on two publicly available datasets i.e., Twitter and Weibo. The proposed model performs better than the current state-of-the-art on Twitter and Weibo datasets by 3.27% and 6.83% respectively.

### 4.4.1 Data

We utilize two publicly available datasets, Twitter and Weibo, to train SpotFake.

1. **Twitter MediaEval Dataset:** The dataset is released as the part of the challenge- The Verifying Multimedia Use at MediaEval [22]. The challenge aimed to find whether the information presented by the post sufficiently reflects reality. The dataset comprises of tweets and their associated images. It consists of 17,000 unique tweets related to various events. The training set consists of 9,000 fake news tweets and 6,000 real news tweets, and the test set containing 2,000 news tweets. A sample of fake images from the dataset is illustrated in Figure 4.3.
2. **Weibo Dataset:** The dataset is introduced in [116]. The fake posts are collected from the official debunking system of Weibo from May 2012-January, 2016. The tweets verified by Xinhua News Agency, an authoritative news agency in China, are considered for real posts. The dataset comprises 4,749 fake posts and 4,779 real posts partitioned into an 8:2 training and testing ratio.

#### 4.4.2 Methodology

In this section, we highlight our pre-processing setup and discusses the core architecture of SpotFake.

##### Data Filtration and Preprocessing

The tweet samples in [22, 116] cater to real-world events and might reflect some discrepancy. One such found in the datasets for evaluating SpotFake was the non-availability of the images for the tweet text. Hence, we eliminate about 18.9% and 1.23% of the data in the Twitter MediaEval and Weibo datasets, respectively. Next, since data is not synthetic and reflects the real world, there is a strong likelihood that the length of the text will differ significantly between samples. Figure 4.8 demonstrates the average text length in Twitter MediaEval and Weibo datasets. Observing the histograms, we select the final length that covers 95%

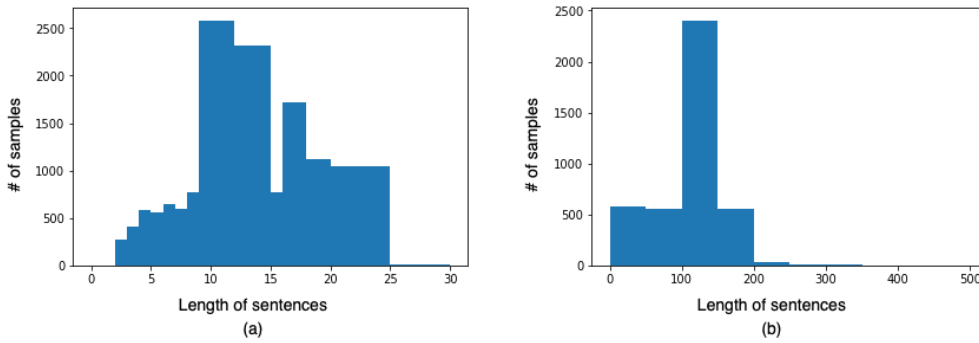


Figure 4.8: A histogram depicting the average length of sentences in the Twitter MediaEval (a) and Weibo (b) datasets, respectively. The final length value is decided to be the one where 95% of the sentences are below it. This is 23 tokens for the Twitter dataset and 200 characters for Weibo dataset.

of the data. For the Twitter MediaEval dataset, this equates to 23 tokens, whereas for the Weibo dataset, it equates to 200 characters. It is to be noted here that Weibo consists of samples in the Chinese language. Tokenizing sentences in such a language is performed at the character level, not the word level. After deciding the sequence length, we trim all the sentences that are longer than it and perform the padding operation (append with zeroes) for anything shorter. Further, we use a pre-trained VGG-19 network to process images in our proposed method. Hence, we resize and normalize the input to the same format the network was originally trained on, i.e. 224x224x3.

### Model Architecture

SpotFake has two sub-modules that extract textual and visual information, respectively. For our text feature extractor, we use pre-trained BERT-Base available on tfhub.<sup>7</sup> BERT stands for Bidirectional Encoder Representations from Transformers. It is an open-source framework to process natural language and uses surrounding text to establish context [62]. The base version of the model is trained on a large corpus (Wikipedia and Bookcorpus) in an unsupervised manner for language modelling to better understand the context of the input sentence. We took the pre-trained version and fine-tuned it for our multimodal fake news detection task. Further, BERT uses a wordpiece tokenizer that splits the word into full forms or word pieces (where one word can be broken into multiple tokens). It also uses special tokens to understand the text properly. One such token is the [CLS] token placed at the beginning of the text. The [CLS] token contains the numerical representation of the content sent at the input. There are two ways to obtain the content embedding via BERT. First, we can average the representations obtained for each word to form the content embeddings. Second, we can use the representations of the [CLS] token.

In our work, we input the pre-processed text from the previous step into the BERT module. Then, we utilize the embeddings of the [CLS] token to represent our text. We obtain vectors of length 768 which are then passed through two fully connected layers of size 768 and 32, respectively, to form the textual feature vector.

Similarly, we use VGG-19 [234] pre-trained the ImageNet database as our visual feature extractor. It is a deep CNN network that can understand the high and low-level features of an image. In our work, we input the re-sized images from the previous step to obtain intermediate representations of length 4096. The obtained vector representation is then passed through two fully connected neural network layers to form the visual feature vector.

The last sub-module of our proposed method is the fusion mechanism. We perform late fusion over the obtained single modality representations to form the compact multimodal representation. We call it news representation (as shown in the Figure 4.9). The represen-

<sup>7</sup>[https://tfhub.dev/google/bert\\_uncased\\_L-12\\_H-768\\_A-12/1](https://tfhub.dev/google/bert_uncased_L-12_H-768_A-12/1)

tations are further passed into a fully connected neural network classifier and then to the classification layer for determining whether the sample is real or fake.

#### 4.4.3 Experimental Setup

Everything is configurable in our model, from the number of hidden layers to the number of neurons and the dropout probabilities. A complete, exhaustive list of hyperparameters is given in Table 4.3. We perform iterations of random search on possible hyperparameter combinations to select the correct permutation of the hyperparameter. In each iteration, the number of possible permutations is reduced based on the performance of the previous

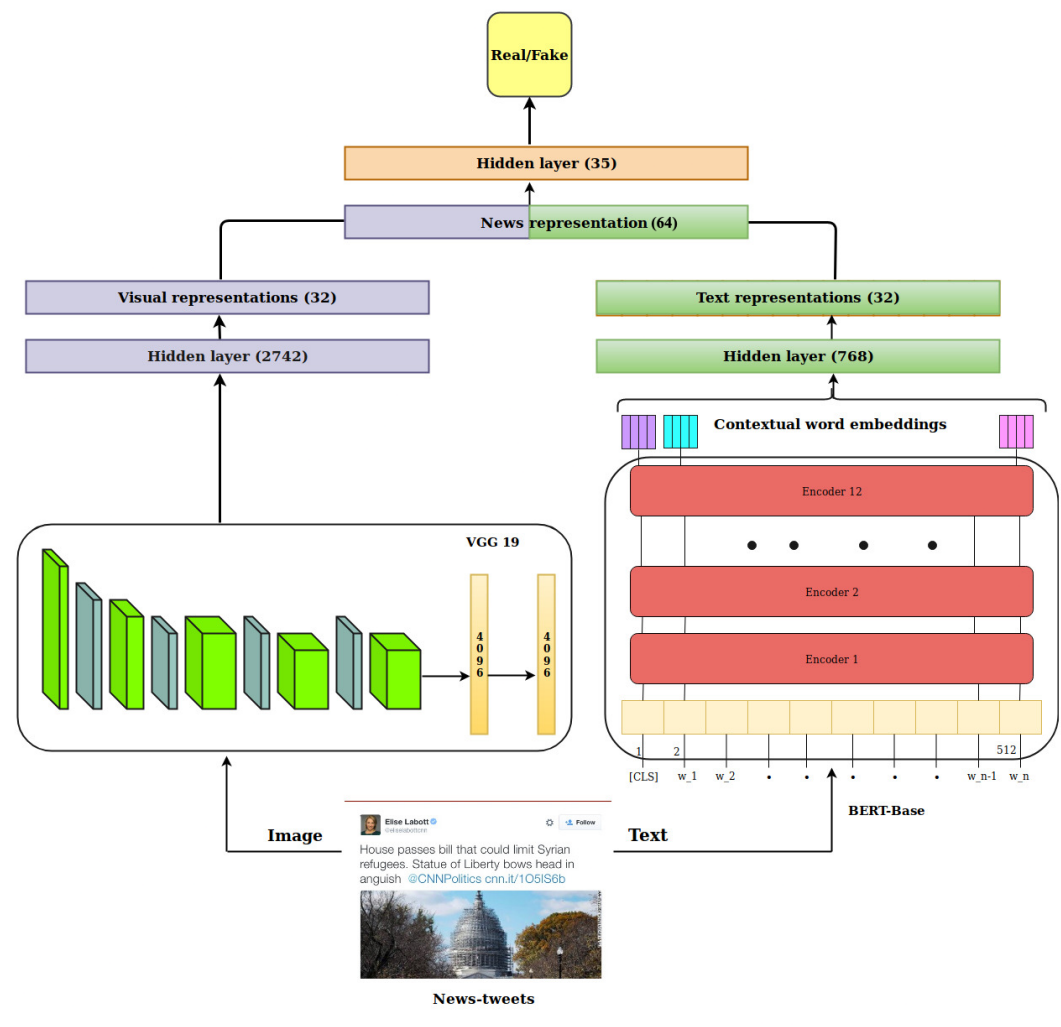


Figure 4.9: A schematic diagram of the proposed SpotFake model. Value in () indicates the number of neurons in a layer. SpotFake consists of two sub-module, text and visual feature extractor. The intermediate representations are fused to form the news vector that is then fed to the classification layer to determine the veracity of the sample.

parameters	Twitter	Weibo
BERT trainable	False	False
VGG trainable	False	False
dropout	0.4	0.4
# of hidden layers (text)	2	2
# of neurons in hidden layer (text)	768,32	768,32
# of hidden layers (image)	2	1
# of neurons in hidden layer (image)	2742,32	32
# of dense layers (concatenation)	1	1
# of neurons in dense layer (concatenate)	64	64
text length	23 words	200 char
batch size	256	256
optimizer	adam	adam
learning rate	0.0005	0.001

Table 4.3: An overview of hyper parameter setting used in SpotFake.

iteration. We use talos<sup>8</sup> library to conduct a random search and evaluate parameters. Talos is a Python library that completely automates model evaluation and hyperparameter adjustments. Next, we use the cyclical learning rate proposed in [244] to find an optimal learning rate for our models. It is a method for configuring, modifying, and adjusting the learning rate while training.

#### 4.4.4 Results

In this section, we are reporting the performance comparison of SpotFake with the existing state-of-the-art methods listed below. We evaluate the performance on accuracy %, precision, recall and F1-score of each class. The complete comparison results are shown in Table 4.4.

- Text: The baseline uses 32-dimensional pre-trained word-embedding weights of text content as input which is then fed to CNN to extract the textual features. Finally, an additional fully connected layer with a softmax function is used to predict whether the post is fake or real.
- Image: The baseline uses pre-trained VGG-19 network in tandem with a fully connected layer to extract visual features.
- VQA [14]: The original VQA model is modified to constitute only one-layer LSTM to perform binary classification task.

---

<sup>8</sup><https://github.com/autonomio/talos>

- NeuralTalk [267] : It is a method capable of generating captions for given images. The intermediate representations are obtained by averaging the outputs of RNN at each timestep which are then fed into a fully connected layer to make predictions.
- att-RNN [116]: The method uses an attention mechanism to fuse the textual, visual and social context features to perform fake news detection. For a fair comparison, we use a modified version of the method that excludes the sub-module responsible for extracting social context information.

Dataset	Model	Acc.	Fake News			Real News		
			Prec.	Rec.	F1-Score	Prec.	Rec.	F1-Score
Twitter	textual	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	visual	0.596	0.695	0.518	0.593	0.524	0.7	0.599
	VQA [14]	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	Neural Talk [267]	0.610	0.728	0.504	0.595	.534	0.752	0.625
	att-RNN [116]	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN- [276]	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	EANN [276]	0.715	NA	NA	NA	NA	NA	NA
	MVAE- [126]	0.656	NA	NA	0.641	NA	NA	0.669
	MVAE [126]	0.745	<b>0.801</b>	0.719	0.758	0.689	<b>0.777</b>	<b>0.730</b>
	SpotFake	<b>0.7777</b>	0.751	<b>0.900</b>	<b>0.82</b>	<b>0.832</b>	0.606	0.701
Weibo	textual	0.643	0.662	0.578	0.617	0.609	0.685	0.647
	visual	0.608	0.610	0.605	0.607	0.607	0.611	0.609
	VQA [14]	0.736	0.797	0.634	0.706	0.695	0.838	0.760
	Neural Talk [267]	0.726	0.794	0.713	0.692	0.684	0.840	0.754
	att-RNN [116]	0.772	0.797	0.713	0.692	0.684	0.840	0.754
	EANN- [276]	0.795	0.827	0.697	0.756	0.752	0.863	0.804
	EANN [276]	0.827	NA	NA	NA	NA	NA	NA
	MVAE- [126]	0.743	NA	NA	NA	NA	NA	NA
	MVAE [126]	0.824	0.854	0.769	0.809	0.802	<b>0.875</b>	<b>0.837</b>
	SpotFake	<b>0.8923</b>	<b>0.902</b>	<b>0.964</b>	<b>0.932</b>	<b>0.847</b>	0.656	0.739

Table 4.4: Classification results on Twitter and Weibo datasets. SpotFake is our proposed model, and we compare it with numerous unimodal and multiple modality baselines for a fair comparison. Our proposed method outperforms the state-of-the-art on Twitter and Weibo datasets by 3.27% and 6.83%, respectively.

- EANN [276]: It is an end-to-end framework that aims to capture event invariant features for fake news detection. The method extracts text and image features by employing Text-CNN [130] and pre-trained VGG-19 network [234] respectively. The prime motivation to keep an event discriminator is to exclude event-specific features and keep shared features among events to better classify a fake sample on a newly emerged event.



- EANN- : It is a variant of the EANN [276] model, excluding the event discriminator component. This version of EANN is trained end-to-end to perform the singleton task of fake news detection.
- MVAE [126]: The algorithm seek to establish correlation across the modalities by designing a multimodal variational autoencoder. This module aimed at reconstructing representations of both the modalities from the learned shared feature vector. This module is used in tandem with the classification module to detect fake news. The textual information is extracted via Bi-LSTMs and image features via VGG-19 pre-trained on ImageNet dataset [234].
- MVAE- : It is a variant of the MVAE [126] model, which does not include the sub-module responsible for extracting correlations across modalities.

### **Observations**

The strongest baselines for multimodal fake news detection are EANN [276] and MVAE [126]. Both models have two configurations each. EANN-/MVAE- is when fake news classifier is trained standalone. EANN/MVAE is when fake news classifier is trained in tandem with a secondary task. The secondary task in case of EANN is an event discriminator that removes the event-specific features and keep shared features among events. Whereas in MVAE, the secondary task is to discover the correlations across the modalities to improve shared representations. Though SpotFake is a standalone fake news classifier, we still outperform both configurations of EANN and MVAE by large margins on both the datasets. On the Twitter dataset, SpotFake achieves 12.97% and 6.27% accuracy gain over EANN- and EANN respectively. Performance gain on Weibo dataset over EANN- and EANN is 9.73% and 6.53% respectively. When compared to MVAE, on Twitter dataset, we outperform MVAE- and MVAE by 12.17% and 3.27% respectively. On Weibo dataset, the performance gain is 14.93% and 6.83%.

#### **4.4.5 Ethical Considerations**

We have taken utmost care that we follow the principles of ethical research. Participants are fully informed about the purpose, duration, and procedures of the research. Participants had the freedom to voluntarily participate and leave the research at any time and without penalty. Further, the news samples used in the survey are publicly accessible, ensuring transparency and adherence to copyright regulations.

#### **4.4.6 Limitations**

There are two main limitations to consider in this study. Firstly, during the survey, we observed an imbalance in the gender distribution, with a higher number of male participants compared to female participants. Additionally, the majority of survey respondents were

youths, comprising 97% of the total participants. Secondly, the text component of SpotFake is limited in its capacity to handle text that exceeds 512 characters. These limitations highlight the need for further exploration and potential improvements in sample diversity and text processing capabilities for more comprehensive research.

#### **4.4.7 Discussion**

This section discusses the nitty-gritty of our proposed architecture, SpotFake. Examining the existing state-of-the-art methods, we find that the complexities in the existing detection methods can be eliminated by designing simple baselines to identify fake news online. Hence, we design SpotFake, which leverages information from text and images to perform the detection task. Our proposed method outperforms the state-of-the-art on Twitter and Weibo datasets by 3.27% and 6.83%, respectively.

Further, news online is also accessible via the web portals of the news channels. Such news articles provide in-depth information about the event, accompanied by pictures and videos, to interest readers. The next area of emphasis is identifying false news for lengthy news items on online news web portals. In order to do this, we will talk about our third core contribution, SpotFake+, a technique that leverages transfer learning to spot fake news on the web, consisting of long-length articles.

### **4.5 SpotFake+: A Multimodal Framework for Fake News Detection via Transfer Learning**

Online news platforms are becoming exceedingly popular amongst consumers due to their ease of access and a vast selection of disparate sources. These platforms further democratise news distribution by making it exceedingly simple to publish. The flip side of this is the lack of proper editorial rigour, fact-checking and the presence of bad actors that have promulgated fake news to an equal extent. Hence, in this section we discuss SpotFake+, a multimodal approach that leverages transfer learning to capture semantic and contextual information from the news articles and its associated images and achieves the better accuracy for fake news detection. SpotFake+ is an advanced version of SpotFake [235] and one of the first attempts that performs a multimodal approach for fake news detection on a dataset that consists of full length articles.

#### **4.5.1 Data**

We utilize the FakeNewsNet repository [227] to train SpotFake+. Fakenewsnet is a multi-dimensional data repository that contains two comprehensive datasets, Politifact and Gossipcop. The authors utilize fact-checking websites like Politifact<sup>9</sup> and Gossipcop<sup>10</sup> to curate

---

<sup>9</sup><https://www.politifact.com/>

<sup>10</sup><https://www.gossipcop.com/>

the ground truth labels for the news samples. Politifact is a fact-checking website that rates the veracity of political assertions made by elected leaders. Politifact is used to compile both real and fake news examples for the political sector. Gossipcop, on the other hand, is a fact-checking website that verifies the accuracy of celebrity news reports. Gossipcop is employed to gather the fake samples, whereas E! Online<sup>11</sup> is utilised to collect the relevant real articles. The number of samples present in the dataset is given in Table 4.5. Each news sample has several features, including news content, social context, and spatial data. Our study uses news content features (linguistic and visual) to evaluate the veracity of the articles.

Dataset	Politifact	GossipCop
Real	624 (321)	16817 (10259)
Fake	432 (164)	5323 (2581)

Table 4.5: The number of samples in the FakeNewsNet repository. The values in the brackets indicate samples fit to use after data pre-processing.

#### 4.5.2 Methodology

In this section, we highlight our pre-processing setup and discusses the core architecture of SpotFake+.

##### Data Filtration and Preprocessing

The fakenewsnet repository consists of data samples curated from online news or fact-checking websites. Therefore, it may contain spurious elements within it. On closer inspection, we found inconsistencies in the dataset and wrote Python scripts to perform the cleaning process. We removed logos from the articles and eliminated samples with GIFs or no photos. Figure 4.10 shows a few images removed from the news samples during the cleaning process.



Figure 4.10: Image samples from the FakeNewsNet repository that are discarded after the cleaning process. The removed images are either the logos or GIFs present on the news website.

<sup>11</sup><https://www.eonline.com/ap>

## Model Architecture

The proposed SpotFake+ is a multimodal approach that successfully detects fake news in full-length articles. The schematic diagram of the model is shown in Figure 4.11.

Our method uses a transfer learning mechanism to understand the data well. Transfer learning is a machine-learning method that uses the information learned from one task to generalize it to another. The idea behind transfer learning is that a model trained on extensive data can be applied as a general framework to another task. Then, instead of starting from scratch, we can employ such learned feature maps. There are different ways to customize a pre-trained method. Our work employs the feature extraction mechanism to learn meaningful representations for the new samples. Further, on top of the pre-trained model, we build a new classifier that will be trained entirely from scratch to reuse the feature maps previously learnt for the dataset. In our work, we utilize the strength of the Transformer module, XLNet [295], to extract textual features. XLNet is an unsupervised language representation learning method that uses permutative language modelling to create a bidirectional contextualized representation of words. The model can understand sequence length beyond 512 due to added recurrence to the Transformer. In our proposed architecture, the title and content of the news sample are combined to provide the input for the pre-trained XLNet module. The resulting intermediate feature representations are then passed through several dense layers to form the final text embeddings. Our visual feature extractor is a VGG-19 model pre-trained on the ImageNet database [234]. It is a robust CNN network that can understand high-level and low-level characteristics in an image. In our work, we input the cleaned and filtered images from the previous step to obtain intermediate representations of length 4096. The resultant vector representation is fed through several dense layers to create the final visual feature vector. Lastly, we perform late fusion over the obtained unimodal representations to form the compact multimodal representation. The representations are further passed into a fully connected neural network classifier and then to the classification layer for determining whether the sample is real or fake.

### 4.5.3 Experimental Setup

The number of hidden layers, the number of neurons, and the dropout probability are all customizable in our model. Table 4.6 contains an entire list of all hyperparameters. All the codes are written in Python 3 using Keras with TensorFlow as the backend. To perform heavy computation, we use 1080Ti GPU and AMD ThreadRipper CPU.

### 4.5.4 Results

We contrast SpotFake+ with a representative list of state-of-the-art singular and multimodal fake news detection algorithms listed as follows:

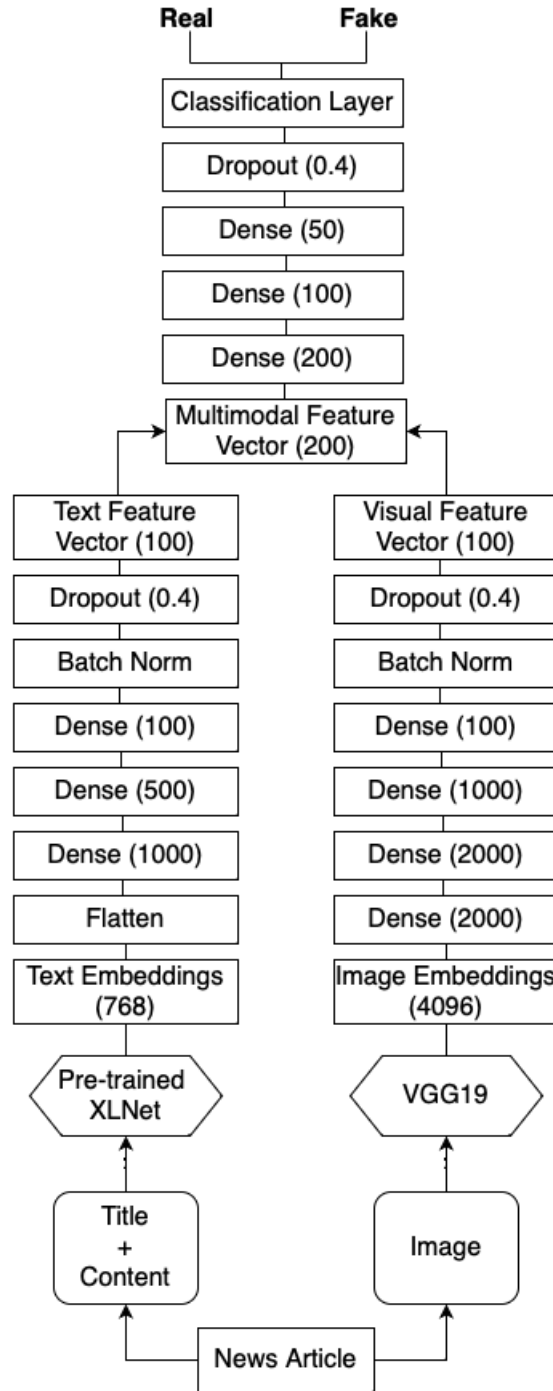


Figure 4.11: Our proposed SpotFake+ method for multimodal fake news detection. SpotFake+ consists of two sub-module, text and visual feature extractors. The intermediate representations are passed through numerous dense layers to form the unimodal representations, which are then fused to form the multimodal vector fed to the classification layer to determine the veracity of the sample. The values in () indicates the number of neurons in a layer.

<b>Parameters</b>	<b>Unimodal Text Model (XLNet + Dense layers)</b>	<b>Image Model (VGG19)</b>
Input feature dimension	768	4096
# of dense layers	3 (100, 500, 100)	3 (2000, 1000, 100)
Output feature dimension	100	100
Dropout	0.4	0.4
Activation	ReLu	ReLu
Optimizer	SGD	SGD
Batch size	32	32

Table 4.6: An overview of hyper parameter setting used in the two sub-modules of the SpotFake+.

- **Machine Learning Baselines:** We compare with several ML baselines like SVM, Naive Bayes and Logistic Regression. Each method is a supervised algorithm that needs labelled training data to perform the classification task.
- **Text-CNN [130]:** It is a deep learning algorithm that is capable of performing text classification. The algorithm uses a series of 1D convolutions and pooling layers to establish semantic relationship between the words of a text.
- **Social Article Fusion (SAF) [226]:** Social Article Fusion is a method that considers the news and social context features to perform fake news classification. It extracts news features via auto-encoder and social context features via recurrent neural networks. The intermediate feature representations are then fused together to form a single concatenated feature vector which is then fed to the classification layer.
- **XLNet + dense layers:** We utilize a pre-trained XLNet module to learn feature representations of the text. The intermediate representations are then passed through several dense layers before performing the classification task.
- **XLNet + CNN:** We utilize a pre-trained XLNet module to learn feature representations of the text. The intermediate representations are then passed through Text-CNN to understand phrase-level representations. The Text-CNN module constitutes three filter of varied sizes capable of capturing different relations in the text.
- **XLNet + LSTM:** We utilize a pre-trained XLNet module to learn feature representations of the text. The intermediate representations are then passed through the LSTM layer capable of memorizing the vital information to find the relevant pattern from the long-length text pieces.
- **VGG-19 [234]:** It is a deep convolutional neural network that consists of 19 layers. We use a version of the VGG-19 network pre-trained on the ImageNet database as the image classification baseline.

Modality	Models	Politifact	Gossipcop
<b>Text</b>	SVM	0.580	0.497
	Logistic Regression	0.642	0.648
	Naive Bayes	0.617	0.624
	CNN	0.629	0.723
	SAF (Social Article Fusion)	0.691	0.689
	<b>XLNet + dense layer</b>	<b>0.740</b>	<b>0.836</b>
	XLNet + CNN	0.721	0.840
	XLNet + LSTM	0.721	0.807
<b>Image</b>	VGG-19	0.654	0.800
<b>Multimodal Text+Image</b>	EANN	0.740	0.860
	MVAE	0.673	0.775
	SpotFake	0.721	0.807
	<b>SpotFake+ (XLNet + dense layer + VGG-19)</b>	<b>0.846</b>	<b>0.856</b>

Table 4.7: Classification results on the FakeNewsNet repository. SpotFake+ is our proposed model, and we compare it with numerous unimodal and multiple modality baselines for a fair comparison. Our proposed method outperforms the state-of-the-art by a relatively large margin.

- SpotFake [235]: The algorithm leverages the power of the language model, BERT, to extract contextual text information [62]. The image features are learned from the pre-trained VGG-19 network. The features obtained from both modalities are fused in an additive manner to build the desired news representation.

### Observations

A simple XLNET + dense layer model that looks at only text beat the best-recorded results, with 74% on Politifact and 83.6% on Gossipcop. SpotFake+ beats both the text-only and multiple-modality models, achieving 84.6% on Politifact and 85.6% on Gossipcop. The detailed analysis of the results is shown in Table 4.7. The loss function graphs are also plotted in Figure 4.12. The plot shows the divergence between training and test loss as a model is trained. From the graph, we can conclude that the performance of SpotFake+ is not a result of over or under-fitting. For both datasets, training loss and validation loss are close, with validation loss slightly greater than the training loss. In addition, we see a decline in the training and validation loss at the start and a flat training and validation loss from a certain point on until the finish.

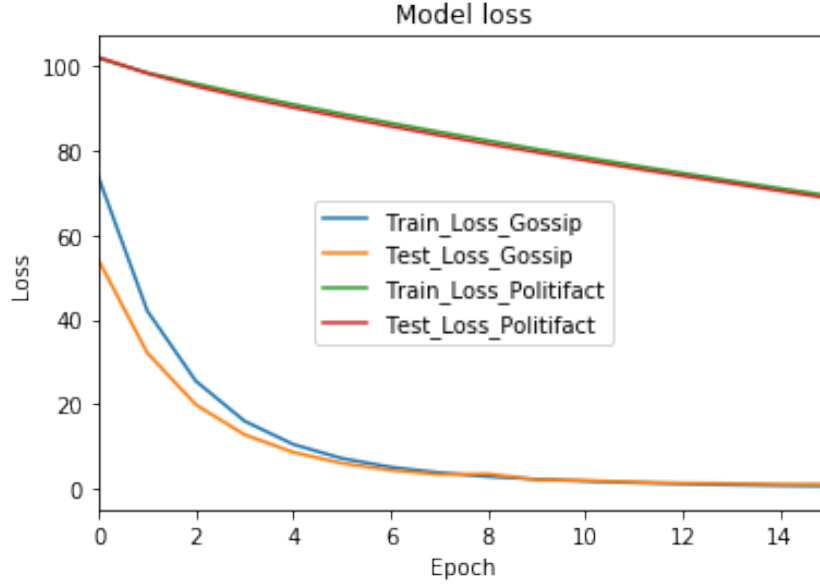


Figure 4.12: Visualisations of the loss functions produced by SpotFake+ using the GossipCop and Politifact datasets. From the graph, we can conclude that the performance of SpotFake+ is not a result of over or under-fitting. For both datasets, training loss and validation loss are close, with validation loss slightly greater than the training loss.

#### 4.5.5 Discussion

In this section, we present SpotFake+, which can classify a news article into two categories, real or fake. SpotFake+ is an advanced version of SpotFake that uses transfer learning to capture the textual and visual features within an article. We utilize the FakeNewsNet repository to evaluate the SpotFake+. The proposed method outperforms the performance shown by both single-modality and multiple-modality models.

## 4.6 Conclusion and Future Works

This chapter discusses numerous solutions to identify multimodal fake news on social media and the web. Prior to that, we conducted a user survey to understand how individuals view fake news. We also examine whether having multiple modalities makes it easier for people to spot fake news. As per the survey findings, individuals still favour traditional venues for news consumption. Users that read or discuss news are few in number. Awareness campaigns are required to educate readers on the value of being up to date on current affairs in order to avoid falling for false news. Furthermore, a sizable portion of the population felt concerned about the truthfulness of the news they read. They also believed that classifying fake news across multiple modalities would be more effective.

Next, we devised two solutions to detect fake news by leveraging multiple modalities. We first suggest SpotFake- a multimodal framework for fake news detection. The proposed method



leverages multimodal signals to identify fake news on social media. SpotFake exploits both the textual and visual features of news presented online. Specifically, we used the language model BERT to learn text features and image features learned from VGG-19 pre-trained on the ImageNet dataset. All the experiments are performed on two publicly available datasets, i.e. Twitter and Weibo. Next, we design SpotFake+, a multimodal framework for fake news detection via transfer learning to identify fake news disseminated over the web. SpotFake+ is an advanced version of SpotFake [235] that leverages transfer learning to capture semantic and contextual information from the news articles and its associated images and achieves better accuracy for fake news detection. One potential direction in the future can be to exploit the role of multiple images in the news article. All the existing works utilize the top image in tandem with the textual and social context features to perform multimodal fake news detection. Furthermore, experiments can be performed to study the contribution of each modality towards solving the problem.

## Chapter 5

# Exploring the Role of Multiple Images for Multimodal Fake News Detection

This chapter is partly a reproduction of a paper published at The ACM Multimedia Asia (ACM MM Asia) 2021.

### 5.1 Introduction

Typically, a news story online consists of text and visual cues to represent the information. Figure 5.1 demonstrates a news sample from the dataset used during evaluation. It consist of the headline, content, top-image and other corresponding images. Critical examination

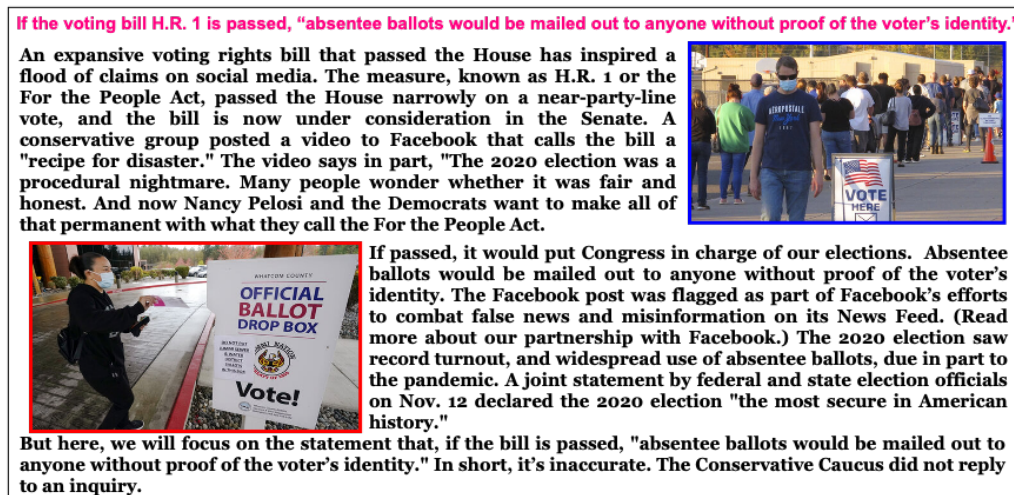


Figure 5.1: A sample of news article present on online media websites. The text written in *pink* and *black* color depicts the headline and content of the news, respectively. The image highlighted in *blue* and *red* represents the top image (first-image) and other-image present within the given news sample, respectively.

of news credibility in presence of such multiple cues becomes challenging. There are two factors at play. First, the news narrative is lengthy due to the assertions and evidence that support it; this makes it easier to introduce false information without getting noticed. Second, news on online media websites is aided with multiple visuals to make it look agreeable. This provides manipulators with several possibilities to reinforce their false tales with images. Prior research has attempted to identify fake news using information from both the text and image modalities [54, 126, 235, 237, 276, 307]. Additionally, in the chapter before, we also focused on designing simple yet effective baselines for multimodal fake news detection. All methods, as mentioned earlier, focused only on the *first-image* with the textual cue to perform multimodal fake news detection. However, we believe other visual signals in a news sample should also receive focus. Furthermore, the initial learning obtained for each modality is combined additively, ignoring the relationship across modalities for fake news classification. In this chapter, we discuss a solution that fixes the above-mentioned drawbacks by exploiting information from all the graphical cues aggregated with the textual details. We believe incorporating multiple images is beneficial for the following reasons: (i) understanding the story in a text often requires the reader to develop mental imagery skills [131, 314]. Images can facilitate the creation of such mental representations [68] and can result in deeper learning [161, 216, 217], (ii) images assist in the clarification of ambiguous relations in the text, often termed as “multimedia effect” [161], (iii) while words can be viewed as descriptive representations, images, in contrast, are depictive external representations, showcasing the meaning that the text represents [18].

## 5.2 Research Objective

We aim to investigate the role of multiple images in multimodal fake news identification. Upon examining the related literature, we find the strongest baselines for single-image and multi-image content-based multimodal fake news identification to be SAFE [307] and Giachanou et al. [82] respectively.

SAFE [307] investigates the relationship between multiple modalities present in a news sample to classify it as fake or real. To capture the relationship effectively, images are first converted into text using pre-trained image2sentence model [266]. Next, the relationship between the modalities is captured by performing a modified version of cosine similarity. Giachanou et al. [82] proposed a multimodal multi-image module that encapsulates information from multiple images in the form of tags and semantic features via a pre-trained VGG-16 network. Next, to establish similarity between the different components of the two modalities, cosine similarity score is calculated between text and image tags. Finally, textual and visual feature vectors are combined with the similarity score, in an additive manner to perform fake news detection.

We do, however, find certain inconsistencies in the current approaches. For example, in the research presented by Zhou et al. [307], the text features are extracted via a Text-

CNN [130], ignoring the contextual information. Whereas the image is converted into text via image2sent [266] model that might result in the loss of semantic information within an image. Further, no comparison is shown with the existing state-of-the-art methods to demonstrate the effectiveness of the proposed model. Conversely, work performed by Giachanou et al. [82] lacks the reasoning for utilizing multiple images for misinformation classification. Second, taking cues only from the headlines and ignoring the content might lead to information loss. Third, while capturing similarity, the top ten image tags are preferred over the image features. This might lead to inconsistent results as (i) extracted tags might fail to capture the semantic relationship across the images, (ii) incorporating only the top ten tags might not capture the information present in the image effectively, and (iii) extracted tags might be limited by the vocabulary of the pre-trained model used for extraction and can introduce external bias in the final representations.

To address the above-mentioned issues, we present an *Inter-Modality Discordance for Multimodal Fake News Detection* [236]. Our method aims to capture the synergies between the modalities for multimodal fake news detection based on inter modality discordance score. We hypothesize that fabrication introduced in any modality will lead to the dissonance between them, i.e. the obtained feature vectors from a fake (real) sample, when projected in a multimodal space, will be distant (closer) and portrays the negligible (significant) relationship between the involved modalities [46]. We examine the discordance score based on a modified version of the contrastive loss that enforces distinct features of a real sample to be closer to each other and farther for fake news. The designed method can also classify samples comprising only unimodal features as the modality-specific sub-modules can independently learn discriminative features via the imposition of the cross-entropy loss.

### 5.3 Data

We use the following publicly available datasets to perform multiple-image multimodal fake news detection task.

- FakeNewsNet Repository (raw version): Shu et al. [227] introduces a multi-dimensional dataset including news content, social context, and spatiotemporal information. Unlike previous datasets that consists of tweets [23, 116], this repository comprises of news articles belonging to either political or entertainment discipline. The fake and real news article pertaining to political domain are collected from Politifact<sup>1</sup> whereas fake and real samples for the entertainment domain are gathered from GossipCop<sup>2</sup> and E! Online<sup>3</sup> respectively. Table 5.1 includes the dataset statistics that are used in the studies.

---

<sup>1</sup><https://www.politifact.com/>

<sup>2</sup><https://www.gossipcop.com/>

<sup>3</sup><https://www.eonline.com/>

- FakeNewsNet Repository (clean version): Giachanou et al. [82] performs multimodal fake news detection by using a portion of dataset released by Shu et al. [227] i.e. GossipCop. Next, authors performed dataset cleaning in which all the news samples with non-news content images are removed by performing deduplication and manual intervention. In our study, for a fair comparison with the state-of-the-art, we use the cleaned version provided by the authors [82] that consist of 2,745 fake and 2,714 real samples having at least one image associated with them.

		# News Articles	# of Images
Politifact (raw)	Real	624 (399)	5,607 (5,027)
	Fake	432 (346)	5,423 (4,462)
	Overall	1,056 (745)	11,030 (9,489)
GossipCop (raw)	Real	16,817 (10,970)	4,05,367 (3,81,117)
	Fake	5,323 (4,223)	1,37,717 (1,25,361)
	Overall	22,140 (15,193)	5,43,084 (5,06,478)
GossipCop (clean)	Real	2,714 (952)	13,567 (1,718)
	Fake	2,745 (2,526)	44,306 (4,364)
	Overall	5,459 (3,478)	57,873 (6,082)

Table 5.1: The dataset statistics used during the experiments. Values in () signify the final count of samples used during experimentation. Politifact (raw) and Gossipcop (raw) are the data samples present in the FakeNewsNet repository [227]. Gossipcop (clean) is a clean version of the Gossipcop (raw) dataset presented by Giachanou et al. [82].

## 5.4 Methodology

In this section, we go over the basic layout of our proposed solution as depicted in Figure 5.2. The model performs multi-task operations, with the primary goal being multimodal fake news detection. It comprises four components, (i) inter-modality discordance score, (ii) text feature extractor, (iii) multiple-visual feature extractor and (iv) multimodal fake news detector.

### Problem Formulation

Assume we have a set of  $n$  news articles,  $N = \{(H_i, C_i, V_i, y_i)\}_{i=1}^n$ . Each news sample  $N_i$  consists of four elements, i.e. headline ( $H_i$ ), text content ( $C_i$ ), image-set ( $V_i$ ) and the corresponding ground-truth label  $y_i$ . We formulate the problem as a binary classification task where  $N_i$  can be categorized as either fake ( $y=1$ ) or real ( $y=0$ ). Specifically, we aim to combine complementary information from multiple modalities for fake news detection on online news websites. We hypothesize that leveraging information encapsulated in the training signals

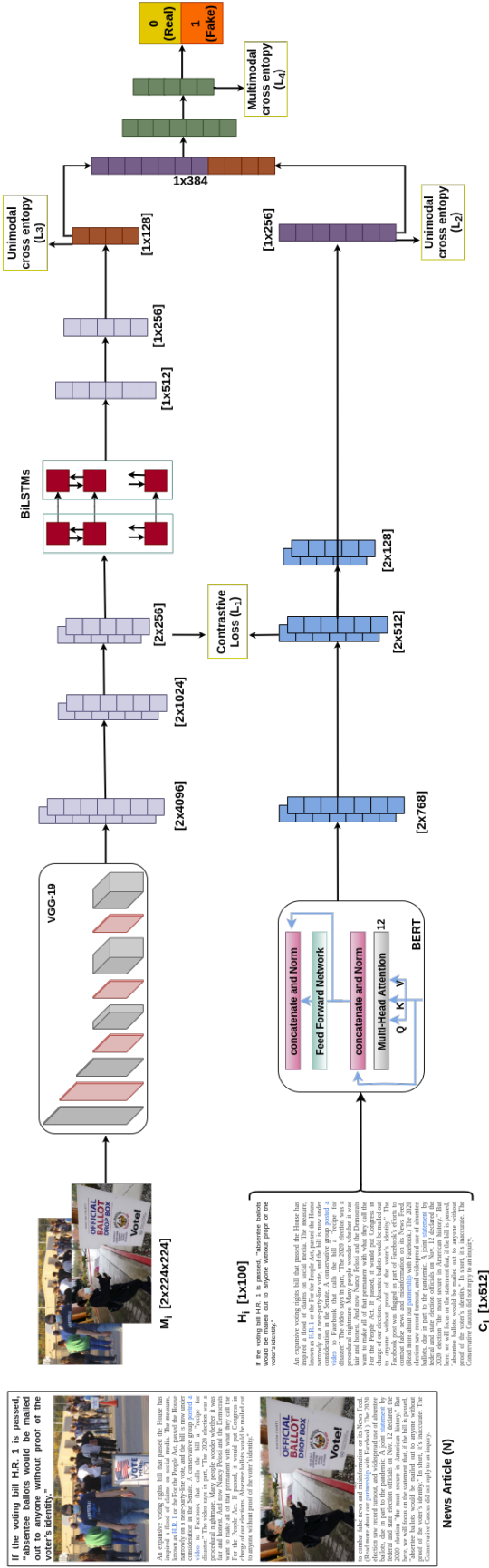


Figure 5.2: Illustration of our proposed model with the primary task being multimodal fake news detection. We introduce three auxiliary learning tasks, i.e. measuring inter-modality discordance score via contrastive loss, multiple visual feature extractor and textual feature extractor. The first number in [ ] depicts the count of the components in news sample  $N$ . For instance, the news sample has one headline ( $H_1$ ), content ( $C_1$ ), and two image components ( $M_1, M_2$ ). The second number in [ ] depicts the feature vector size obtained after passing through each layer of the proposed method. For instance, it is 512, 768, 512 and 128 for the text feature extraction module.

of the related atomic learning tasks will improve the generalization capability of the model. Hence, we present our approach as a multitask learning method that constitutes three atomic learning tasks finding a rich and robust representation of the input data to perform the desired primary task better. The first atomic task is the *Inter-modality discordance score*, ensuring that components of a real news article are pulled together in an embedding space while simultaneously pushing apart the components of a fake news article. The second component is the *unimodal multiple-visual feature extractor*, which uncovers hidden patterns within a set of sequential images to obtain the final discriminative rich embeddings. The final one is the *unimodal text feature extractor* that embodies the intra-modality relationship via granular fragment representation, independently from the headline and the content. Next, we discuss each component in detail.

### 5.4.1 Inter-modality Discordance Score

The first auxiliary task presented in our proposed method is calculating the discordance score. It captures the relationship (discordance) between various components present in a news article for multimodal fake news detection. More specifically, the idea is that the average distance between the different components of a fake news article is greater than the average distance between the different components of a real news article, in a multimodal space. We believe measuring discordance has the following implications. A recent study by Claire Wardle, *First Draft News Research Director*<sup>4</sup> presents a list of seven different types of fabricated content circulated in the online world. Though all these forms of fake news are created differently, some of them can be captured by measuring discordance between different components of a news article. For example, capturing relationships will help in the easy identification of fake stories where (i) headlines and visuals are not supporting the content, (ii) genuine content is circulated with false contextual information, (iii) both the content and image are real, but the context in which they appear frames a false story. Taking inspiration from [127], we measure inter-modality discordance score via a modified version of contrastive loss function. It is a form of metric learning that has shown significant improvement over the conventional cross entropy loss for supervised classification [47]. The objective is to predict relative distance between the inputs. The detailed outline to calculate inter-modality score is summarized in Algorithm 1 where  $(H_i, C_i, V_i)$  depicts the intermediate feature representations for the headline, content and image-set respectively.  $r_c$  denotes the centroid value and *distance* signify the average distance between the components of a news sample from the centroid. The distance metric chosen for measuring similarity is the euclidean distance (L2-norm).  $M$  indicates margin value. The function of margin is that, when average distance between the different components of a fake news article are distant enough, no efforts are wasted on enlarging that distance. However, when that distance is not

---

<sup>4</sup><https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79>

greater than  $M$ , then loss will represent a positive value, and net parameters will be updated to produce more distant feature vectors. The vice-versa happens for the real news articles.

---

**Algorithm 1:** Measuring Inter-modality Discordance Score (Training Phase)

---

**Input:**  $P = [H_i^R, C_i^R, \dots, V_i = \{I_1^R, \dots, I_k^R\}_{k=1}^l]_{i=1}^m, y \in (0, 1), M$

**Output:** Loss

**for** each  $P_i$  i.e.  $(H_i, C_i, \{I_1, \dots, I_k\})$  **do**

$r_c = \frac{1}{|P|} \sum P_i$ ;

$distance = \frac{1}{|P|} \sum_{i=1}^{|P|} \|r_{P_i}, r_c\|$ ;

**if**  $y=1$  **then**

$Loss = \max(0, M - distance)$ ;

**else**

$Loss = distance$ ;

**end**

**end**

---

#### 5.4.2 Unimodal Visual Feature Extractor

The second auxiliary task considered in the proposed method is the multiple visual feature extractor. The goal of the module is to obtain feature representations for a sequence of images. Though, there are numerous works from the past that have considered the *first-image* in tandem with text to solve multimodal misinformation classification [54, 237, 307]. However, little attention is drawn towards exploiting multiple images present in a news article. Taking inspiration from Giachanou et al. [82], we present a novel system that extracts sequential information from multiple images in a two-fold manner. Appending visual extractor as an auxiliary task unfolds various benefits such as, (i) the module will try to learn complex patterns from the multiple images to generate better feature representations and, (ii) the resultant feature vector for the images will not only assist in better performance of other auxiliary tasks but will also enhance the representations of the news vector that will perform multimodal fake news detection.

Let  $V_i = (I_1, I_2, \dots, I_k)$  represent a set of images present in the news article, where  $V_i$  denotes the image-set of a news sample and  $k$  denotes the corresponding count of images present within  $V_i$ . Since the count of image sequences in each news sample can be different, we perform padding keeping the sequence length to be the maximum count of images present in any sample of the training dataset. The pre-processed images are first passed through a VGG-19 network pre-trained on a ImageNet database. The second to last layer of the VGG-19 network serves as a feature embedding for each image present in the news article. Next, to capture temporal features from the intermediary sequential visual cues, we employ a Bidirectional Long-Short Term Memory (BiLSTM) cells. The continued representations then



obtained are passed through fully connected layers to match the length of vector dimensions with that of the resultant textual feature vector.

### 5.4.3 Unimodal Text Feature Extractor

The third auxiliary task introduced in the proposed method is the textual feature extractor. It extracts contextual representations from the headline and the content of a news sample. Context refers to information that helps the message of a literary text interpret accurately. Unlike Word2Vec [167] and GloVe [191] which are context insensitive, the word embeddings generated by Transformer [62] are context sensitive representations. Context sensitivity refers to giving different representations to same words according to the theme where they have been placed. For example, the word “bank” in the context of finance and in the context of river would carry different representation. Additionally, pre-trained models fails to produce the correct embeddings for the word tokens falling outside the training vocabulary. To address, the above mentioned shortcomings, we followed [62, 264] to design text representations for our proposed model. We believe there are two benefits to utilising such features. First, it will assist in better formation of the discriminative text feature representations that will capture both the semantic and the contextual meaning effectively. Second, such feature representations will in turn help in capturing the relationship (as discussed in section 5.4.1) between modalities efficiently.

In our work, each content piece ( $C_i$ ) of a news article is segregated into sentences. A sentence  $s$  is represented as a sequence of WordPiece tokens  $\{w_s^1, w_s^2, \dots, w_s^k\}$ , where  $w_s^k$  is aggregation of the token, position and segment representation for the  $k^{th}$  token present in a sentence  $s$ . Similarly, headline ( $H_i$ ) is divided into tokens for further processing. Motivated by Devlin et al. [62], the generated input sequence is passed to Transformers which is an attention based encoder-decoder type architecture. In our work, we use BERT (Bidirectional Encoder Representations from Transformers) architecture [62] that is deeply bi-directional and looks at the words by jointly conditioning on both left and right context in all layers. The context is pre-trained on Books Corpus and English Wikipedia to provide a richer understanding of language. BERT module is pre-trained with two unsupervised prediction tasks: (i) next sentence prediction that aims to predict whether sentence A is the next sentence of B and, (ii) masked language model that aims to randomly masks some percentage of the input tokens at random, and then predicts only those masked tokens. For our task, it transforms the input sequence of tokens  $\{w_s^1, \dots, w_s^k\}$  into an abstract continuous representation  $\{z_s^1, \dots, z_s^k\}$  that confines all the learned information of that input. Finally, the continuous representations obtained from the [CLS] token is considered as the resultant textual feature vector.

### 5.4.4 Multimodal Fake News Detector

In this section, we discuss the primary task of our proposed method i.e. multimodal fake news detection. It leverages information from the textual and multiple visual entities of a news

sample to form a multimodal feature vector. Although, we include auxiliary task that extracts modal-specific features from the news article, the necessity to add multimodal features is two fold, (i) capturing information from multiple modalities will help in creating a more robust system as compared to the ones build solely on unimodal features and, (ii) multimodal features will be more capable in discovering non-trivial patterns and relationship between data instances.

Next, to form multimodal feature vector, we need to perform multimodal fusion. The existing fusion strategies can be categorized as, (i) *early fusion*, that merges unimodal features to form joint representation before attempting to classify, (ii) *late fusion*, where individual modalities pass through powerful targeted unimodal processing pipeline to classify the content for each modality independently. The results obtained from the unimodal pipelines are then merged to perform the final classification and, (iii) *hybrid fusion*, a strategy that lies midway between the two extremes (i.e. early and late fusion). Here, the fusion point lies somewhere in the middle of the network.

In our work, we opted for early fusion over later fusion as the former fusion strategy is able to capture the cross-correlations between the data features, thereby providing an opportunity to improve the overall model performance as compared to the latter one. In addition, Gadzicki et al. [79] performed various experiments on wide range of datasets and concluded that early fusion performs far better than the other existing fusion strategies.

#### 5.4.5 Loss Functions

As stated earlier, we design the problem of multimodal fake news detection as a multitask learning method. The proposed method comprises of a primary task and three other related auxiliary (atomic) learning tasks. We employ a combination of loss functions from all the four tasks to better perform the desired task. It is to be noted that all four tasks are performed during model training but the primary task is considered when assessing the performance of the model.

To calculate inter-modality discordance, the training objective is that euclidean distance between the various components of a news sample, in a multi-modal space, is minimised for the real news articles and maximized for the false news samples. Taking inspiration from Khosla et al. [47], we use the modified version of contrastive loss, originally presented for training of Siamese networks [21] to calculate the inter-modality discordance score. This ensures the distinction between the positive and negative samples effectively. The loss function is represented in Equation 5.1.

$$L_1 = \begin{cases} \frac{1}{n} \sum_{i=0}^m d(r_m, r_c), & \text{if real sample} \\ \max(0, M - \frac{1}{n} \sum_{i=0}^m d(r_m, r_c)), & \text{otherwise} \end{cases} \quad (5.1)$$

Here,  $r_m$  depicts the embedding vector for the  $m$ -th component of a news article,  $r_c$  denotes the centroid,  $M$  depicts the margin, set as a hyper-parameter and,  $d(\cdot)$  denotes the

euclidean distance between the component of a news sample and its corresponding centroid value.

To capture discriminative unimodal features, we employ the cross-entropy loss to learn the modal independent representations in a robust fashion. The respective loss functions are represented in Equation 5.2 and 5.3, respectively. In addition, to model the cross-correlations between the entities, we perform a multimodal fusion on intermediate features to form the desired multimodal news vector. The corresponding loss function for the same is depicted in Equation 5.4.

$$L_2 = \frac{1}{n} \sum_{i=1}^n y_T^i \log \hat{y}_T^i + (1 - y_T^i) \log(1 - \hat{y}_T^i) \quad (5.2)$$

$$L_3 = \frac{1}{n} \sum_{i=1}^n y_I^i \log \hat{y}_I^i + (1 - y_I^i) \log(1 - \hat{y}_I^i) \quad (5.3)$$

$$L_4 = \frac{1}{n} \sum_{i=1}^n y_C^i \log \hat{y}_C^i + (1 - y_C^i) \log(1 - \hat{y}_C^i) \quad (5.4)$$

Hence, the final loss for the proposed method is the weighted sum of the four losses, i.e.,

$$L = \alpha L_1 + \beta L_2 + \gamma L_3 + \delta L_4 \quad (5.5)$$

where  $(\alpha, \beta, \gamma, \delta) \in [0, 1]$ . In our experiments, we set the value to be one. However, further hyper-parameter tuning can be performed on these values.

## 5.5 Experimental Setup

This section includes a summary of the dataset splitting, details regarding the evaluation metric employed, training and implementation specifics, and a list of the hyper-parameters used during training.

### Dataset Preprocessing and Split

We conduct extensive experiments on the two versions of the FakeNewsNet repository [227]. One is the original dataset presented by Shu et al. [227]. The other, presented by Giachanou et al. [82] is the cleaned version, obtained after dataset pre-processing on the [227]. Both, Giachanou et al. [82] and Singhal et al. [237] noticed that the various images associated with the news sample are non-news content images. Such images are either logo, gifs, advertisements or icons of the publishing house. To perform multimodal fake news detection, authors in [82] removed the unwanted images and performed their experiments. While evaluating the performance of our proposed model, we conduct experiments on all the versions of the datasets. The complete statistics of dataset used in our evaluation is shown in Table 5.1. We randomly split the dataset into 80% training and 20% testing. Table 5.2 depicts the distribution of fake and real samples in each split.

	Politifact (raw)	GossipCop (raw)	GossipCop (clean)
Train	316 (271)	8,752 (3,367)	766 (2,032)
Test	83 (75)	2,218 (856)	186 (494)
Overall	399 (346)	10,970 (4,223)	952 (2,526)

Table 5.2: The train-test split statistics of the datasets used during the experiments. Values in (.) signifies the count of fake samples.

### Metrics

We employ accuracy, precision, recall, and F1-score, standard assessment measures when conducting classification trials, to assess the performance of our proposed method.

### Implementation and Training Details

We use Pytorch version 3.7.9 to execute our proposed approach. All experiments use Nvidia GeForce RTX 2080Ti and 3090 GPUs with 11GB and 24GB of memory, respectively.

### Hyper-parameters

All the hyper-parameters are carefully adjusted in the validation set. An early stopping strategy is used to stop training if a certain number of threshold epochs does not improve validation performance. A learning rate of  $1e-5$  and a dropout value of 0.4 are employed during model training.

## 5.6 Results

In this section, we present a series of experiments to demonstrate the viability of our suggested approach. We particularly want to answer the following research questions:

1. Can the proposed method enhance multimodal fake news detection by incorporating multiple visual cues?
2. How effective is multimodal false news detection using the modality discordance hypothesis?

We next provide an overview of the baseline models used for comparison before delving into the questions in further detail.

### Baselines

We compare our proposed methodology with a representative list of state-of-the-art multimodal fake news detection algorithms listed as follows:

- LIWC [190]: It stands for Linguistic Inquiry and Word Count. The method is used to classify the text samples along the psychological dimensions. It identifies how much percentage of the words present in the text lies into any of the linguistic, psychological, and topical categories. Such analysis of the data is then fed as an input to the model for further analysis.
- VGG-19 [234]: VGG-19 is a variant of VGG model that comprises of 19 layers. It is generally used for image classification. Here, we use a fine-tuned version of VGG-19 as a baseline for images.
- Att-RNN [116]: The method is designed to utilize the textual, visual and social-context features for the multimodal fake news detection. The variant of the model used in the paper excludes the social-context information for a fair comparison.
- SAFE [307]: The objective of this model design is to capture the similarity among modalities to jointly exploit the multimodal information and excavate better representations for multimodal fake news detection. For this, a modified version of cosine similarity is introduced. The text and visual features are extracted by passing the initial representations through Text-CNN [130]. The intermediate representations for the images are obtained via image2sentence model.
- Multi-image Multimodal Method [82]: This is the first research that explores multiple images in tandem with text to perform fake news detection. To extract visual features from multiple images, tags information in combination with the features obtained via pre-trained VGG-19 network is used. Authors also exploit semantic information i.e. text-image similarity by calculating cosine similarity between them the modalities.
- L2: It is a variant of the proposed model when using only the textual information.
- L3: It is a variant of the proposed model when using only the visual information.
- L2+L3+L4: It is a variant of the proposed model without the inclusion of the similarity score, i.e. inter-modality discordance score.

We compare the performance of our proposed approach with the single-image and multi-image multimodal fake news detection state-of-the-art methods. Currently, SAFE [307] and Giachanou et al. [82] serves as the strongest baselines for the single-image and multiple-image respectively. Additionally, we also demonstrate the importance of each component in the proposed method by performing the ablation study.

#### **RQ1: Can the proposed approach enhance multimodal fake news detection by incorporating multiple visual cues?**

We contrast our suggested approach with the current state-of-the-art methods discussed in Section 5.6. A comparative table depicting the results and improvements of the proposed

methodology with the strongest baselines is shown in Table 5.3 and Table 5.4 respectively. We draw the following inferences, (i) from Table 5.3, we observe that our proposed method beats the text-only and image-only baselines for both the datasets, (ii) Table 5.4 shows that our proposed method beats the strongest baseline for multi-image multimodal fake news detection [82], (iii) Additionally, as shown in Table 5.3 for a comparison with the single-image multimodal fake news detection methods, our proposed method outperforms att-RNN [116] and SAFE [307] on the Politifact dataset. However, we observe inconsistent performance on the GossipCop (raw) dataset.

		LIWC <sup>†</sup>	VGG-19 <sup>‡</sup>	Att-RNN <sup>‡</sup>	SAFE <sup>‡</sup>	Proposed Method
<b>Politifact</b> <b>(raw)</b>	Acc.	0.822	0.649	0.769	0.874	<b>0.913</b>
	F1	0.815	0.720	0.826	0.896	<b>0.902</b>
<b>GossipCop</b> <b>(raw)</b>	Acc.	0.836	0.775	0.743	0.838	<b>0.850</b>
	F1	0.466	0.862	0.846	<b>0.895</b>	0.743

Table 5.3: Comparison of our proposed model with the text<sup>†</sup>, image<sup>‡</sup> and single-image multimodal<sup>‡</sup> fake news baselines. Our proposed method outperforms single and multiple modality baselines on the given datasets. However, we observe inconsistent performance on the GossipCop (raw) dataset.

		Giachanou et al. [82]	Proposed Method
<b>GossipCop</b> <b>(clean)</b>	Acc.	NIL	<b>0.880</b>
	F1	0.795	<b>0.915</b>

Table 5.4: Comparison of our proposed model with the multi-image multimodal fake news detection baselines.

#### Investigating the inconsistent performance of the proposed model on Gossipcop (raw)

Since previous studies [82, 237] have pointed out the presence of non-news images (i.e. logo, gifs, icons of the news websites and advertisements) within the news articles for the GossipCop (raw) dataset. We hypothesize that existence of such noisy images lead to decrease in the performance of our proposed method. To investigate the case, we first perform intersection on the GossipCop (raw) and GossipCop (clean) datasets to get the resultant set comprising of non-news images. We observe that on an average for a sample, 77.77% of the images are non-news in the raw dataset. We further examine what amount of noise is passed through the model. We found that on an average for a sample, there is a 0.8 probability i.e. atleast two out of three images passed to the model will be noisy. This probability further shoots up to 0.97 for atleast one out of three images passed to the model. In conclusion, our proposed model is designed to capture the discordance between the modalities. Since there

exists no relationship between the incoming noisy images and text, our model fails to learn any representative pattern about the news article leading to inconsistent results.

## RQ2: Efficacy of modality discordance hypothesis towards multimodal fake news detection

In order to answer RQ2, we perform two experiments. **First**, we visualize the average distance between the components of a news sample via Kernel Density Estimate (KDE) plots. Figure 5.3 shows the distribution of discordance score during the training and testing phase for GossipCop (clean) and Politifact respectively.

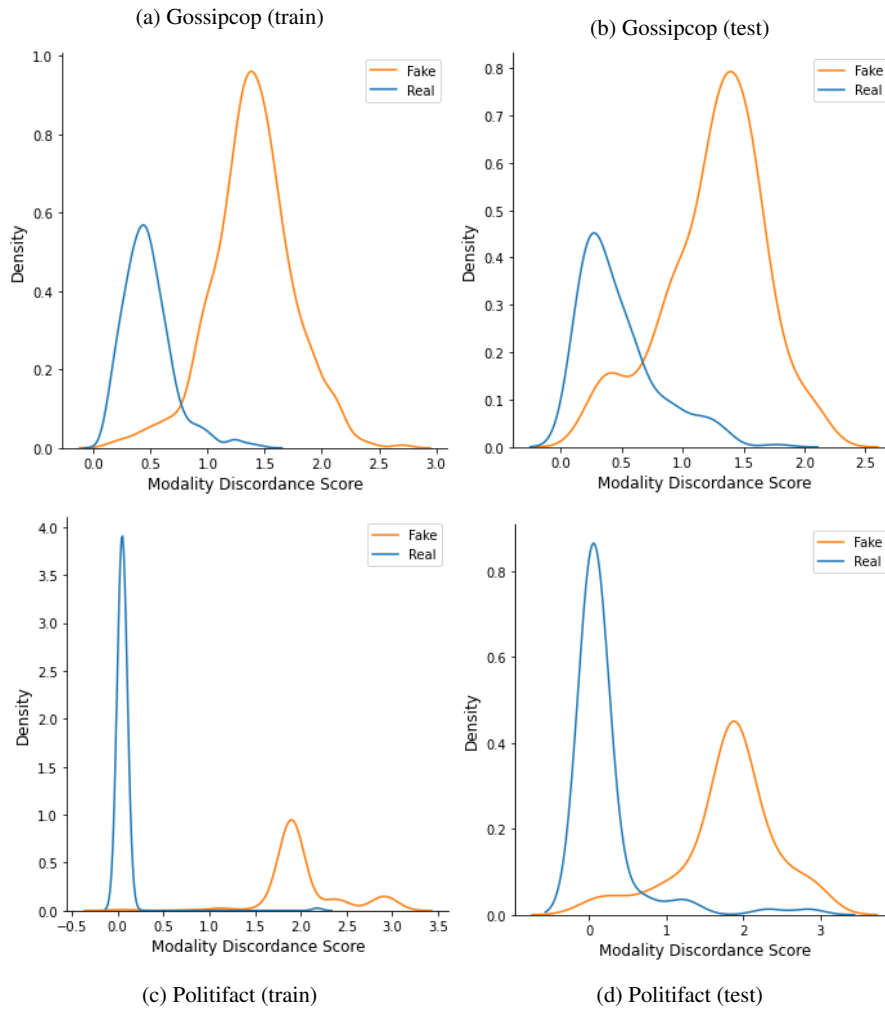


Figure 5.3: We are measuring modality discordance scores on train and test sets of GossipCop (clean) and Politifact, respectively. We observe that, on average, the mean distance between real and fake news components is 0.476 and 1.39, respectively. The density plots are narrow, depicting that the variance in the output of a class is low. We also notice that the intersection of the area under the curve for real and fake news is minimal, indicating a clear separation between the classes.

As stated in Algorithm 1, the margin value signifies the radius around the embedding space of a sample. We hypothesize that the average distance between the components of a real (fake) news sample lies closer (farther) to the radius ( $M=1$ ). We observe that on an average, the mean distance between the components of real and fake news is 0.476 and 1.39, respectively, as shown in Figure 5.3 (a). The validity of our results are solidified by the fact that peaks in our density plots are narrow which shows that the variance in the output of a class is low. Additionally, the intersection of area under the curve for real and fake news is minimal indicating a clear separation between the classes. All the aforementioned observation are consistent across datasets and training-validation splits, as depicted in Figure 5.3 (b-d).

**Second**, we perform an ablation study to examine the performance of our proposed model with its different variants. The results are depicted in Table 5.5. Here L2 and L3 signifies the performance of the proposed model when using only text and visual features respectively. On the other hand, L2+L3+L4 represents the multimodal framework in absence of the inter-modality discordance score. The general performance of the different variants on both the datasets is: Proposed Method > L2+L3+L4 > L2 > L3. From Table 5.5, we observe an improvement in the L2+L3+L4 variant on addition of L1. This shows that adding inter-modality discordance score in the multimodal detection method aids in better fake news detection.

		<b>L2</b>	<b>L3</b>	<b>L2+L3+L4</b>	<b>Proposed Method (L1+ L2+L3+L4)</b>
<b>Politifact (raw)</b>	Acc.	0.898	0.828	0.906	<b>0.913</b>
	F1	0.896	0.821	0.889	<b>0.902</b>
<b>GossipCop (clean)</b>	Acc.	0.861	0.684	0.863	<b>0.880</b>
	F1	0.906	0.785	0.908	<b>0.915</b>

Table 5.5: Comparison of the proposed model with its different variants. L2 and L3 signify the variant comprising only the text and visual features. Whereas L2+L3+L4 represents the multimodal framework in the absence of the inter-modality discordance score. We observe an improvement in the L2+L3+L4 variant with the addition of L1 (inter-modality discordance component).

## 5.7 Limitations

While developing the solution, we learned about some general limitations of multimodal fake news detection. **First is the need for robust validation.** Though our paper surpasses the performance of the current state-of-the-art methods, this may only be true over time across the geographies and events. Due to the volatile nature of how fake news is fabricated, it is possible for neural networks to learn statistical cues that might not generalize for future



incoming data. For example, a model trained on 2020 data can be heavily biased towards the pandemic-related fake news that might not perform well further down the line. We believe that to generate more robust fake news models; they should be cross-validated across time, geographies, events and other parameters. However, this is not possible in the current situation due to a lack of suitable datasets. **Second is the need for comparison.** One major limitation observed during our study was the unavailability of state-of-the-art methods to assess the performance of the proposed method. This might be because different methods have opted for different evaluation metrics. For example, for single-image fake news methods, SpotFake [237], SAME [54], and SAFE [307] were the baselines. All three models lack a comparative study with other baselines and use different evaluation metrics to demonstrate the performance. We believe fixing the performance metric across all the methods will help create more robust systems.

## 5.8 Conclusion and Future Works

This chapter presents a solution that looks into the role of several images in multimodal fake news identification. We introduce Inter-modality Discordance for Multimodal Fake News Detection- a method that leverages information from the text and multiple visual features of a news sample and investigates the relationship between them via a modified version of contrastive loss. The count of image sequences in each news sample can be different; hence the method can incorporate news samples with varied image counts. The proposed method can also classify samples comprising only unimodal features as the modality-specific sub-modules can independently learn discriminative features via the imposition of the cross-entropy loss. Extensive experiments on two real-world datasets demonstrate the strong performance of our proposed method.

Solutions discussed so far focus towards content-based fake news detection leveraging information from text and images. In Chapter 4, we designed simple yet effective baselines for multimodal fake news detection. This chapter focused on leveraging the other visual cues within the news apart from text and the top image. We also incorporated a strategy to learn relationships between different modalities while deciding. However, we hypothesize that not all modalities play an equal role in deciding the veracity of the news. Therefore, it might be interesting to look for a solution that analyses the significance of the modality in decision-making.

## Chapter 6

# Extracting Intra and Inter Modality Relationship for Multimodal Fake News Detection

This chapter is partly a reproduction of a paper published at the International Workshop on Multimodal Understanding for the Web and Social Media (MUWS), co-located with The WebConf (WWW) 2022 [240].

### 6.1 Introduction

*A husband divided his assets in half while settling the divorce case with her ex-wife.* At first read, the text might look believable. Now, when we read the same piece of information but with an image, as shown in Figure 6.1, we might question the credibility of the news. The image show objects (i.e. car and laptop) cut into halves. In an ideal situation, dividing assets into two does not mean how it is shown. The example is a marketing gimmick by a German lawyers' group and is a typical case of a satirical story.<sup>1</sup> Further, viewing the image would



Figure 6.1: A sample of tweet from the MediaEval dataset [23]. The corresponding text reads, ‘Husband Gave His Unfaithful Ex-Wife Half Of Everything He Owned – Literally.’ The intra-modality feature extractor, one of the sub-components of our proposed technique, curates the fine-grained salient representations for the text, capturing words like *Half* from the caption and the image segments highlighted by the blue colored boxes.

also significantly impact the decision if someone was asked to determine the veracity of the news. Hence, we hypothesize that not all modalities play an equal role in the decision-making process of any particular sample. In this chapter, we aim to discuss a solution that works on the principle of strong and weak modalities to identify the authenticity of the news. A modality is strong when assigning a high probability to the correct class. A higher probability implies a more informative signal and stronger confidence. The underlying principle is introduced in [149] and has thus far been utilized in several study fields [2, 170]. However, existing methods for multimodal fake news detection and those discussed in the previous chapters do not work on the foundation of strong and weak modalities [54, 126, 235, 237, 276, 307]. Instead, methods capture high-level information from different modalities and jointly model them to determine the authenticity of the news. Such methods assume that both text and image modality play an equal role in determining the veracity of the news. Additionally, feature extraction occurs globally, ignoring the salient pixels containing meaningful information. For instance, Figure 6.1 highlights essential segments of the image and text containing details—however, the approaches mentioned before use Text-CNN or VGG-19 to extract visual information. Since a complete image is passed through the network to generate the representations, unwanted (redundant) information in the form of background is also included. Similarly, there is a need to extract contextual dependencies for the textual features.

This chapter discusses an architecture that utilizes a multiplicative multimodal method [149] to capture inter-modality relationships. The proposed method suppresses the cost of a weaker modality by introducing a down-weight factor in the cross-entropy loss function. The down-weight factor associated with each modality highlights the average prediction power of the remaining modalities. So, if the other modality has higher confidence in predicting the correct class, the cost associated with the current modality is suppressed and vice versa. We also capture the intra-modality relationship that generates fragments of a modality and then learn fine-grained salient representations from the fragments. For image modality, we perform bottom-up attention [13] to extract the image patches. The complex relationship between the patches is then encoded via the self-attention mechanism [264]. We obtain the final visual representation by performing an average pooling operation over the fragment representations, resembling the bag-of-visual-words model. We use a wordpiece tokenizer to generate text fragments for text modality. Taking inspiration from [62], we use a Transformer module, BERT, to extract contextual representations. We then pass the obtained embeddings via a 1D-convolution neural network to extract the phrase-level information. We obtain the resultant text representation by passing intermediate learned representations via a fully connected layer.

---

<sup>1</sup><https://www.cnn.com/2015/06/22/german-lawyers-claim-theyre-behind-viral-divorce-story.html>

## 6.2 Research Objective

Given multiple input modalities, we hypothesize that not all modalities may be equally responsible in the decision-making process on any particular sample. Hence, in this chapter, we aim to design a method that effectively identifies and suppresses the cost of a weaker modality and extracts relevant information from the strong modality on a per-sample basis. We also plan to establish an intra-modality relationship by extracting fine-grained image and text features. Specifically, we aim to answer the following evaluation questions:

- **RQ1:** Is the proposed model improving multimodal fake news detection by leveraging intra and inter-modality relationships?
- **RQ2:** How effective are the extracted fragments and self-attention representations in improving multimodal fake news detection?
- **RQ3:** Can the proposed model identify the modality that aided in easy recognition of falsification in a particular news sample?

## 6.3 Dataset

We use two publicly available datasets to perform multimodal fake news detection,

- **MediaEval Benchmark (Twitter) Dataset:** The dataset is released as a part of the Verifying Multimedia Use task that took place as a part of MediaEval Benchmark in 2015 [23]. It comprises of 16,521 unique tweets with corresponding images covering 11 real-world events. There are 12,740 tweets in the training partition divided into 7,032 fake tweets and 5,008 real tweets. The testing partition comprises 2,564 fake tweets and 1,217 real tweets.
- **Weibo Dataset:** The dataset is introduced in [116] where fake posts are collected from the official debunking system of Weibo from May 2012-January 2016. The tweets verified by Xinhua News Agency, an authoritative news agency in China, are considered for real posts. The dataset comprises 4,749 fake posts and 4,779 real posts partitioned into an 8:2 training and testing ratio.

## 6.4 Methodology

Assume we have a set of  $n$  news articles,  $S = \{S_i^T, S_i^I\}_{i=1}^n$ . Each news sample  $S_i$  consists of two elements, content ( $S_i^T$ ) and the corresponding image ( $S_i^I$ ). We encapsulate the cross-modal synergies to investigate multimodal fake news detection as a binary classification task where  $S_i$  can be categorized as either fake ( $y=0$ ) or real ( $y=1$ ).

Every content piece ( $S_i^T$ ) comprises of  $k$  sentences,  $\{S_i^{Ta}\}_{a=1}^k$ . Each sentence  $S_i^{Ta}$  is further tokenized into  $\{w_{i1}, w_{i2}, \dots, w_{ik}\}$  sub-words using a subword algorithm. Similarly, every

image ( $S_i^I$ ) is segregated into a finite set of  $\{m_i^1, m_i^2, \dots, m_i^{36}\}$  fragments via a bottom-up attention module [13]. Next, fine-grained intra-modality relationship for each modality is captured by passing the intermediate representations through multi-head self-attention layers. Finally, the continuous representations,  $\{z_s^{i:k}\}$  for each of the text piece are passed through a one dimensional convolution followed by a fully-connected layer to extract the text representations. The final image embeddings are obtained by performing average pooling over the continuous intermediate representations. To correctly classify a news sample, we employ a statistically efficient method that aims to suppress the learning from the modality that independently incorrectly classified the sample. After, every modality  $\{S_i^T, S_i^I\}$  makes its own independent decision with its modal-specific model, a multiplicative fusion method is designed to mitigate the information gained from the weak modality by introducing a down-weight factor.

Our proposed framework, as shown in Figure 6.2, consists of two components, the Intra-modality relationship extractor that gathers segment information from all the modalities independently and derives the global relationship among each fragment extracted for each modality; and an Inter-modality relationship extractor that designs a robust method to derive complementary information across modalities via multiplicative fusion. Next, we explain each component in detail.

#### 6.4.1 Self Attention

An attention module can be defined as a mapping function that takes in  $n$  inputs and returns  $n$  outputs. Every input comprises three representations: key, query, and value. These representations interact and decide which to focus on more. Our method seeks to process the obtained image and text fragments independently. We use self-attention, a special case of the attention mechanism, to encode the interaction between fragments of images or texts. All three input representations, i.e., queries, keys and values, are equal in self-attention. Taking inspiration from [264], we perform the attention function via Transformers. A Transformer module embodies two sub-layers, the multi-head self-attention sub-layer and the position-wise feed-forward sub-layer. In the multi-head attention sub-module, the attention mechanism runs multiple times in parallel. Each attention head attends to a part of the sequence uniquely, and finally, all independent outcomes are combined and linearly reshaped to obtain the desired projection size. For instance, let us assume that we have a finite set of fragments  $\{f_1, f_2, \dots, f_b\}, f_i \in R^{1 \times d}$ , where  $b$  depicts the total number of fragments available for a modality and  $d$  is the representation size. Combining all these together resulted in a matrix  $F = [f_1; \dots; f_b] \in R^{b \times d}$ . Mathematically,

$$MultiHead = [head_1 \otimes, \dots, head_i] W^O \quad (6.1)$$

$$head_i = attention(FW_i^Q, FW_i^K, FW_i^V) W^O \quad (6.2)$$

$$attention(Q, K, V) = softmax(QK^T / \sqrt{d_k}) V \quad (6.3)$$

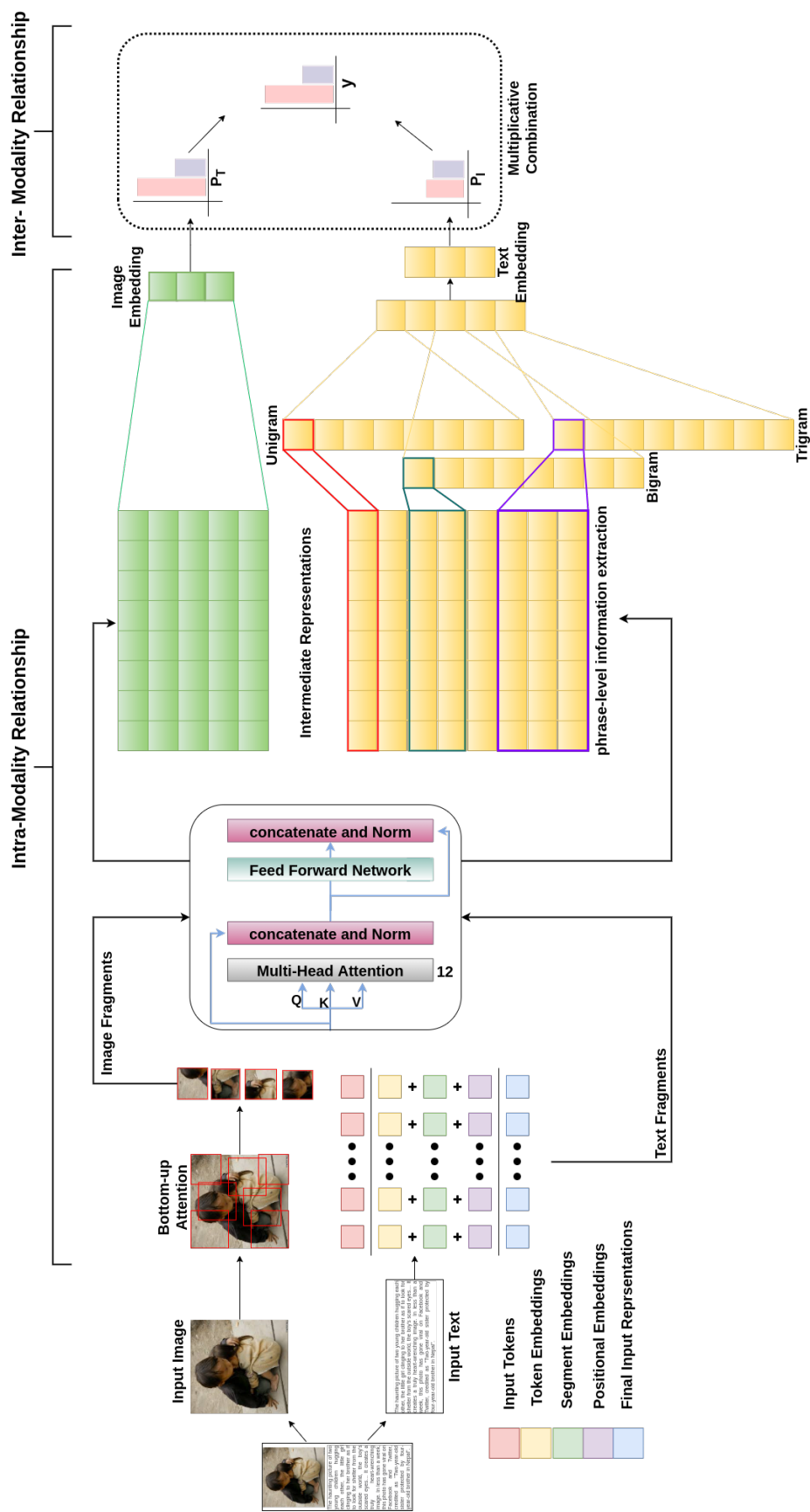


Figure 6.2: The framework of our proposed model. It comprises of two sub-modules, one extracting intra-modality relationship across modalities and other capturing the inter-modality relationship with an emphasises on the modality that shows least resistance towards fake news classification.

where  $Q$ ,  $K$ ,  $V$  are Query, Key-Value pair and  $W$  are all learnable parameter matrices. Next, we apply a position-wise feed-forward sub-module on each fragment independently and identically to rearrange fragment embeddings in the preferred dimension:

$$FFN(g) = \max(0, gW_1 + b_1)W_2 + b_2 \quad (6.4)$$

$(g, b_1, b_2) \in R^{1 \times d_g}$ ,  $(W_1, W_2) \in R^{d_g \times d_g}$ . At last, to promulgate position information to higher layers, we apply residual connections followed by layer normalization around each sub-layer.

#### 6.4.2 Text Embeddings

Context refers to information that helps the message of a literary text interpret accurately. Unlike Word2Vec [167] and GloVe [191] which are context insensitive, the word embeddings generated by Transformer [62] are context sensitive representations. Context sensitivity refers to giving different representations to same words according to the theme where they have been placed. For example, the word “bank” in the context of finance and in the context of river would not carry the same representation. Additionally, pre-trained models fails to produce the correct embeddings for the word tokens falling outside the training vocabulary. To address, the above mentioned shortcomings, we follow [62] to design text representations for our proposed model. Each content piece ( $S_i^T$ ) of a news article is segregated into sentences. A sentence  $s$  is represented as a sequence of WordPiece tokens  $\{w_s^1, w_s^2, \dots, w_s^k\}$ , where  $w_s^k$  is aggregation of the token, position and segment representation for the  $k$ -th token present in a sentence  $s$ . Motivated by [62, 264], we pass the generated input sequence to Transformers which is an attention based encoder-decoder type architecture. In our work, we make use of Transformer encoder that maps an input sequence of tokens  $\{w_s^1, \dots, w_s^k\}$  into an abstract continuous representation  $\{z_s^1, \dots, z_s^k\}$  that confines all the learned information of that input. For instantiation, we adopt the BERT (Bidirectional Encoder Representations from Transformers) architecture [62] that is deeply bi-directional and looks at the words by jointly conditioning on both left and right context in all layers. The context is pre-trained on Books Corpus and English Wikipedia to provide a richer understanding of language. We pass the continuous representations obtained via BERT for each of the textual fragments,  $Z_s = [z_s^1, \dots, z_s^k]$  through a 1D convolution neural network to capture the hidden local context information of the sequential features. The convolutional layer is used to produce a feature map,  $F_s = \{f_s^i\}_{i=1}^{k-h+1}$  from a sequence of continuous inputs  $\{z_s^{i:(i+h-1)}\}_{i=1}^{k-h+1}$  via a filter  $w_s$ . Each local input is a group of  $h$  continuous words represented as,

$$f_s^i = \sigma(w_s \cdot z_s^{i:(i+h-1)} + b_s),$$

$$z_{i:(i+h-1)} = \text{concat}(z_i, z_{i+1}, \dots, z_{i+h-1}),$$

where  $w_s, z_s^{i:(i+h-1)} \in R^{hd}$ ,  $b_s \in R$  is a bias,  $\sigma$  is ReLU activation function and,  $w_s, b_s$  are the parameters learned within the convolution neural network. After obtaining the convolution outputs, we apply max-pooling operation on the obtained feature map for dimensionality

reduction,  $\hat{f}_s = \max\{f_s^i\}_{i=1}^{k-h+1}$ . The text representations are then derived via  $s = W_s \hat{f}_s + b_s$ , where  $\hat{f}_s \in R^n$ ,  $W_s \in R^{qn}$ ,  $b_s \in R^q$ . Specifically,  $n \in \{1, 2, 3\}$  depicts the three window sizes chosen for encapsulating the phrase level information at uni-gram, bi-gram and tri-gram level. Finally, we obtain the resultant text feature vector by passing  $s$  through a fully connected layer followed by L2 normalization.

### 6.4.3 Image Embeddings

Following [125, 141], for extracting the news image features, we primarily focus on extracting the objects and other salient regions using a pre-trained detector. The reasons for abandoning the classical method for image feature extraction are twofold. First, the embedding representations obtained from the last pooling layer of VGG/CNN successfully preserve the spatial information but fails to capture the semantic relationship [110, 175, 274]. Second, the classical approach divides an image equally at the spatial level, which leads to fragments containing redundant background information. Filtering out unnecessary fragments demands additional computation and amendments in the algorithmic design.

Given an image  $I$ , we employ bottom-up attention model [13] pre-trained on Visual Genome [133] to extract a fixed-sized set of  $l$  image patches,  $V = \{v_1, v_2, \dots, v_l\}$ ,  $v_i \in R^d$  such that each image feature encodes a salient region and is represented by a pooled convolutional feature vector. The bottom-up attention module makes use of Faster R-CNN [209], a two-step object detection framework that identifies image patches belonging to certain classes and localize them with bounding boxes. The first stage, identified as a Region Proposal Network, aims at predicting object bounds and objectness scores at each spatial position. In the second stage, region of interest pooling is used to capture feature map for each bounding box and to classify the image within the proposed region. Next, we add a position-wise fully connected layer to transform the image features into a required dimension space for further processing i.e.  $\{y_1, y_2, \dots, y_l\}$ ,  $y_j \in R^{l \times d}$ . We then pass the intermediate representations obtained for each of the image fragment through a self-attention layer introduced in sub-section 6.4.1 to capture complex relation among the image patches. With such a mechanism, each output fragment can attend to all input fragments, and the distance between each fragment is just one. The output obtained after passing through multi-head self-attention module followed by the layer normalization ( $LN$ ) is,  $O = [o_1; \dots; o_l] \in R^{l \times d}$  where,  $O = LN(Y + (MultiHead(Y)))$ . Then, the position-wise feed-forward and layer normalization is applied to get a set of continuous representations,  $Z = \{z_i\}_{i=1}^l$ , where  $z_i = LN(o_i + FeedForward(o_i))$ . Finally, the obtained image embeddings are condensed into a dense representations by performing average pooling followed by L2 normalization to procure the resultant image feature vector.

### 6.4.4 Multimodal Fusion

In our work, we aim to capture the interaction among different modalities to better perform the task in hand. An intrinsic method to combine complementary information is to aggregate



signals from different modalities and design learning models over the concatenated features. The idea has been incorporated in numerous existing multimodal techniques including early and late fusion [91, 92], hybrid fusion [17] and fusion methods enumerated from deep learning methods [177, 178, 282]. In such methods, intermediate representations are collated together and are jointly modeled to make a decision. These techniques are termed as additive approaches due to the type of aggregation operation performed. However, there are some practical constraints in integrating synergies across modalities using existing additive approaches: (i) Additive methods assume that every modality is potentially useful and should be combined, (ii) The neural network models built on top of the aggregated features are not able to determine the quality of each modality and its contribution toward the desired task, on per sample basis. For instance, fake manipulations can be created by fabricating text, images or both modalities. This clearly indicates that fake content creators can induce forgery in any modality to deceive the readers. Given multiple input modalities, an ideal algorithm should be robust to noise from weak modalities and harvest relevant details from the stronger modalities on per sample basis.

In this work, we perform multiplicative multi-modal fusion that [150] addresses the above mentioned challenges. Specifically, the technique explicitly models the fact that on any particular sample not all modalities may contribute equally. Let every modality present in a news sample make its own independent decision i.e.  $P_T = [p_T^1; p_T^0]$ ,  $P_I = [p_I^1; p_I^0]$ , where  $P_T, P_I$  denotes the text and image predictions respectively. Typical, additive combination would have resulted in,

$$l_{cross\_entropy}^y = - \sum_{i=1}^M \log(p_i^y) \quad (6.5)$$

where  $l^y$  is a class loss as it is part of the loss function associated with a particular class. To mitigate the challenges, we utilize a down-weight scaling factor,

$$q_i = [\prod_{j \neq i} (1 - p_j)]^{\beta/(m-1)} \quad (6.6)$$

where  $\beta$  is the hyper parameter used to control the strength of down-weighting. The down-weight factor is responsible to suppress the predictive power for the modality that incorrectly classifies the sample. For instance, if  $p_i$  shows confident predictions for the correct class then down-weight factor will be small value, suppressing cost for other modalities ( $j \neq i$ ). Intuitively, when the current modality is giving a favourable prediction, other modalities need not to be equally useful. Larger the value of the down-weight factor, stronger suppressing effect on that modality and vice versa. Thus, when performing fake news classification task, we leverage the benefits of extracting complementary information from the given piece of information using the multiplicative method that have resulted in the modification of loss function as,

$$l_{multiplicative}^y = - \sum_{i=1}^M q_i \cdot \log(p_i^y) \quad (6.7)$$

## 6.5 Results

In this section, we present a series of experiments to demonstrate the viability of our proposed approach. We particularly want to respond to the following research questions: (i) Is the proposed model improving multimodal fake news detection by leveraging intra and inter-modality relationships? (ii) How effective are the extracted fragments and self-attention representations in improving multimodal fake news detection? (iii) Can the proposed model identify the modality that aided in easy recognition of falsification in a particular news sample? We next provide an overview of the baseline models used for comparison before delving into the questions in further detail.

### Baselines

We compare our proposed methodology with a representative list of state-of-the-art multimodal fake news detection algorithms listed as follows:

- Text-CNN [130]: It is a deep learning algorithm that is capable of performing text classification. The algorithm uses a series of 1D convolutions and pooling layers to establish a semantic relationship between a text’s words.
- BERT [62]: It is a transformer-based machine learning technique capable of extracting contextual meaning from the text. We used a version of BERT pre-trained on Wikipedia and Brown Corpus.
- VGG-19 [234]: It is a deep convolutional neural network that consists of 19 layers. It is used for classifying images. We used a version of the VGG-19 network pre-trained on the ImageNet database.
- EANN [276]: It is an end-to-end framework that aims to capture event invariant features for fake news detection. The method extracts text and image features by employing Text-CNN [130] and pre-trained VGG-19 [234] networks. The prime motivation to keep an event discriminator is to exclude event-specific features and keep shared features among events to better classify a fake sample on a newly emerged event.
- MVAE [126]: The algorithm seeks to establish correlation across modalities by designing a multimodal variational autoencoder. The module reconstructs representations of both modalities from the learned shared feature vector. This module is used in tandem with the classification module to detect fake news. The textual information is extracted via Bi-LSTMs and image features via VGG-19 pre-trained on the ImageNet dataset [234].
- SpotFake [235]: The algorithm leverages the power of language models to extract contextual text information [62]. The image feature is generated from the pre-trained

VGG-19 network. The features obtained from both modalities are fused in an additive manner to build the desired news representation.

We also include various variants of the proposed method to show effectiveness of the extracted fragments and fragment self-attention representations in multimodal fake news detection.

- Proposed w/o Text: It is a variant of the proposed method when using only visual information.
- Proposed w/o Image: It is a variant of the proposed method when using only textual information.
- Proposed w/o multiplicative fusion: It is a variant of the proposed method that fuses information from both modalities in an additive manner. Taking cues from the previous multimodal approaches [126, 235, 276], we used the late fusion strategy to perform the desired task.

We compare the performance of our proposed approach with the single-image and multi-image multimodal fake news detection state-of-the-art methods. Additionally, to understand the contribution of the components to the overall system, we test the performance of the suggested strategy by removing certain components.

**RQ1: Is the proposed model improving multimodal fake news detection by leveraging intra and inter-modality relationships?**

The question aims to examine the performance of the proposed model with the existing state-of-the-art models described previously. We use two unimodal text-based baselines comprising deep learning (Text-CNN) and transformer (BERT) based techniques. VGG-19 is used as a unimodal image-based baseline. We also use several multimodal fake news detection State-of-the-arts methods (EANN, MVAE, SpotFake) for a fair comparison. Results shown in the Table 6.1 indicate that our proposed method outperforms the baselines on accuracy and F1-score for Twitter and Weibo, respectively. SpotFake [235] is the strongest baseline on multimodal fake news detection, and our proposed method outperforms it by a fair margin of an average of 3.050% and 4.525% on the accuracy and F1-score, respectively.

**RQ2: How effective are the extracted fragments and self-attention representations in improving the multimodal fake news detection?**

The RQ2 aims to measure how effective the extracted fragments and self-attention representations are in improving multimodal fake news detection. To answer the question, we compare proposed model with its different variants. The question aims to establish effectiveness of each sub-module in the method. We compare the performance of our proposed approach with the single-image and multi-image multimodal fake news detection state-of-the-art methods.

Additionally, to understand the contribution of the component to the overall system, we test the performance of the suggested strategy by removing certain components. For instance, Proposed w/o Multiplicative measures the effectiveness of the fusion strategy described in Section 6.4.4. On average, we encounter a drop of 1.8% and 2.7% on the accuracy and F1-score, respectively, on removing the multiplicative fusion module. Similarly, to examine the effectiveness of the extracted fragments for the text and image modality, we evaluate the performance of the Proposed w/o Text and Proposed w/o Image, respectively.

### RQ3: Can the proposed approach indicate the modality that aided in easy recognition of falsification in a particular news sample?

We conduct a qualitative examination of the proposed methodology to validate its efficacy in identifying the modality that easily recognizes falsification in a particular news sample. We took a random subset from the test set to examine the accuracy of the obtained  $q$  score. A few example studies are displayed in Figure 6.3. Further, we also cross-examine the outcomes with human intervention in the loop to validate the inferences. The observations are as follows:

- Figure 6.3 (a) is a typical case of *False Context* where truthful information (Eiffel Tower lit up) is shared with the false contextual information (Barbaric attacks in Lahore). Sources<sup>2</sup> claim the information to be false, with no connection between Paris and Pakistan.
- Figure 6.3 (b) is a classic example of *Fabricated Content* where the content is created to deceive or do harm. Both text and image provide strong confidence in detecting

<sup>2</sup><https://www.scoopwhoop.com/The-Photo-Showing-Eiffel-Tower-Lit-Up-In-Green-Is-Not-For-Victims-Of-Lahore-Blast/>

	Twitter				Weibo			
Baselines	Acc	Prec.	Rec.	F1	Acc	Prec.	Rec.	F1
Text-CNN <sup>†</sup>	0.614	0.599	0.612	0.594	0.794	0.791	0.800	0.792
BERT <sup>†</sup>	0.607	0.595	0.601	0.594	0.861	0.860	0.870	0.859
VGG-19 <sup>‡</sup>	0.558	0.572	0.573	0.558	0.654	0.502	0.502	0.501
EANN <sup>‡</sup>	0.648	0.697	0.630	0.634	0.782	0.790	0.780	0.778
MVAE <sup>‡</sup>	0.745	0.745	0.748	0.744	0.824	0.830	0.822	0.823
SpotFake <sup>‡</sup>	0.777	0.791	0.753	0.760	0.8923	0.874	0.810	0.835
<b>Proposed<sup>‡</sup></b>	<b>0.831</b>	<b>0.836</b>	<b>0.832</b>	<b>0.830</b>	<b>0.900</b>	<b>0.882</b>	<b>0.823</b>	<b>0.847</b>

Table 6.1: Comparison of our proposed model with the unimodal text<sup>†</sup>, image<sup>‡</sup> and multimodal<sup>‡</sup> fake news detection baselines. Our proposed model beats the strongest baseline, SpotFake by an average of 3.05% and 4.525% on accuracy and F1-score, respectively.

	Variants	w/o Text	w/o Image	w/o Multiplicative	Proposed
MediaEval Benchmark (Twitter)	Acc	0.703	0.626	0.813	<b>0.831</b>
	Prec.	0.707	0.622	0.814	<b>0.836</b>
	Rec.	0.707	0.621	0.812	<b>0.832</b>
	F1	0.705	0.621	0.812	<b>0.830</b>
Weibo	Acc	0.736	0.794	0.873	<b>0.900</b>
	Prec.	0.608	0.802	0.824	<b>0.882</b>
	Rec.	0.588	0.791	0.815	<b>0.823</b>
	F1	0.595	0.791	0.820	<b>0.847</b>

Table 6.2: Comparison of our proposed model with its different variants. We can see an improvement of 1.8% and 2.7% in the accuracy measures on Twitter and Weibo datasets, respectively, on adding intra modality relationship module (*Proposed*) to the base detection method (*w/o Multiplicative*). Similarly, to examine the effectiveness of the extracted fragments for the text and image modality, we evaluate the performance of the *Proposed w/o Text* and *Proposed w/o Image*, respectively.

the veracity. Hence, the down-weight factor ( $q$ ) assigned by the model to the text and image is 0.7 and 0.4, respectively.

- Figure 6.3 (c) depicts an image of a girl claiming that she is selling chewing gum on the streets of Jordan. The story is an instance of *False Connection*. It is a case where no truth is established between the headline of the content, its image, or caption. On closer inspection, we observe a happy emotion depicted in the image that is irrelevant to the war-like situation. Moreover, our model also shows stronger confidence in the Image modality by assigning a  $q$  score of 0.03 and 0.87 to the text and image, respectively.
- Figure 6.3 (d) highlights a clear case of a doctored photo presented with genuine information to deceive the readers. Since imaging modality shows more substantial confidence towards prediction, the model’s performance highlights same. The example is a variant of fake news, often termed as *Manipulated Content*.

## 6.6 Conclusion

This chapter presents a novel framework that leverages intra and inter-modality relationships for multimodal fake news detection. Our proposed method comprises of two sub-modules. The first sub-module, *intra-modality feature extractor*, is responsible for extracting fine-grained salient image and text features. We perform bottom-up attention and wordpiece tokenizer to extract image and text fragments. The complex relationship between the image patches is then encoded via the self-attention mechanism [264]. Similarly, the obtained text

fragments are passed through a 1D-convolution neural network to extract the phrase-level information. The second sub-module, *inter-modality relationship extractor*, fuses multimodal features multiplicatively. We hypothesize that not all modalities play an equal role in the decision-making process in any particular sample. We identify the weak modality and suppress its cost, which in turn encourages the emergence of important patterns from the informative (strong) modality. With this, on a per-sample basis, weaker modalities are not forced to contribute to the prediction process. As a result, during training, the model learns to choose more reliable modalities over less reliable ones. We utilize a multiplicative multimodal method [149] that suppresses the cost associated with a weaker modality by introducing a down-weight factor in the cross-entropy loss function. The down-weight factor associated with each modality highlights the average prediction power of the remaining modalities. So, if the other modality is more confident in predicting the correct class, the cost associated with the current modality is suppressed and vice versa. Experimental results on the two publicly available datasets show that our proposed method outperforms state-of-the-art methods by an average of 3.050%

## 6.7 Limitations and Future Works

We discovered two concerns when evaluating the effectiveness of the proposed strategy. First is the validity of the q score. If multiple modalities are involved in a sample, our method assigns weak or strong label based on the q score. To validate the q score in our work, we recruited a group of three annotators to perform labelling on a portion of the test set to indicate the modality that helped them determine the accuracy of the news. It might be challenging to scale the proposed method in real-time as we would need more



Figure 6.3: Different variants of fake news detected by our proposed model. Our proposed sub-module, *inter-modality feature extractor* (discussed in Section 6.4.4) utilizes a multiplicative multimodal method to identify strong modality on a per-sample basis to determine veracity of the news sample.

ground-truth labelling of the samples at the modality level, which can be time-consuming, laborious and costly. Hence, as a society working towards eliminating fake news, we should consider designing datasets presenting veracity labels at the modality level. The benefit is that the granular level representation would aid in identifying different variants of the fake news as discussed in Chapter 2. Second is the need for humans in the loop to determine the veracity of the news. The model may make erroneous choices if it only examines the granular-level representations of the samples. This is because there are cases of fake news where the involved modalities depict correct information, but the combination represents a fake story. Additionally, the fabrications employed to create false news are inconsistent. We would need to train the models frequently to teach them new patterns. Humans may act as an external knowledge source for correcting incorrect predictions as they can understand the facts well and map the learnings to determine the veracity.

## Chapter 7

# Inspecting Fake News

This chapter is partly a reproduction of papers published at the Communications of the ACM, 2022 [238] and The International Conference on Web and Social Media (AAAI-ICWSM), 2022 [242].

### 7.1 Overview

The proliferation of fake news is accelerating, and it has significantly impacted all events worldwide, including the 2016 U.S. elections [28], global pandemic (COVID-19) [5, 15] and 2019 Indian General elections [208]. There have been continuous efforts by researchers worldwide to halt the dissemination of false information. The methods covered in Section 3 and those proposed in Chapters 4-6 demonstrate near-to-perfect performance towards the detection task. However, in reality, such solutions fail to cope with the changing dynamics of fake news. We hypothesize that the performance promises of the current state-of-the-art multimodal fake news detection methods are significantly overstated. The overestimation might be due to the systematic biases in the dataset or the models that leverage such preferences ignoring the actual cues for detection.

This and the following chapter covers the second major theme of the thesis, i.e. inspecting Multimodal Fake News. We begin the inspection of fake news from an information viewpoint, where the goal is to study the changing dynamics of fake news over time. We select India as the region for the study since the production and circulation of fake content there is a rising problem.<sup>1</sup> Though there are numerous efforts towards automatic fact-checking and fake news detection, we believe that such solutions might have a limited impact on solving the issue in India because the first language (mother tongue) of Indians is diverse and not restricted to English. As a result, we might encounter the production and distribution of fake content in regional languages.

Towards this end, we begin our research by looking for resources catering to fake news samples from India. We find one by Sharma et al. [222] that proposes IFND: Indian Fake

---

<sup>1</sup><https://indianexpress.com/article/india/214-rise-in-cases-relating-to-fake-news-rumours-7511534/>



News Dataset, comprising of the following attributes: (i) title, (ii) date and time, (iii) source of the news, (iv) link to news, and (v) label. The dataset consists of 37,809 and 7,271 real and fake news samples. The real news is collected from Tribune<sup>2</sup>, Times Now News<sup>3</sup>, The Statesman<sup>4</sup>, NDTV<sup>5</sup>, DNA India<sup>6</sup>, and The Indian Express<sup>7</sup>. The fake news samples are curated from Alt News<sup>8</sup>, Boomlive<sup>9</sup>, Digit Eye<sup>10</sup>, The Logical Indian<sup>11</sup>, News Mobile<sup>12</sup>, India Today<sup>13</sup>, News Meter<sup>14</sup>, Factcrescendo<sup>15</sup>, TeekhiMirchi<sup>16</sup>, Daapan, and Afp<sup>17</sup>. In another attempt, Dhawan et al. [64] proposed FakeNewsIndia to examine the fake news incidents in India. The team curated 4,803 fake news stories from June 2016- December 2019 from 6 fact-checking websites: Times of India, Alt News, Afp, India Today, pIndia, and Factly. The dataset comprises the following attributes, title, author, text, video, date-time, and website. The authors have also curated 5,031 tweets and 866 Youtube videos present in the dataset. Though the datasets have made an effort to create resources that cater Indian region, it still faces a few limitations, (i) The IFND dataset is highly imbalanced. No assurance about the authenticity of sources is provided, (ii) In the FakeNewsIndia dataset, the sample count is low. Data curation is also performed for a short period, (iii) Both the curated datasets consist of samples in English, missing the data in regional languages, (iv) There are numerous attributes present in a website, but both the papers are limited to some specific features. This might lead to information loss, and (v) None of the datasets had the *investigation\_reasoning* attribute proposed in our solution. To the best of our knowledge, there is no existing repository for fact-checked or fake news stories that curate data covering a large spectrum of different regional languages.

## 7.2 Research Objective

The amplification of fake news is becoming rampant in India too. Debunked information often gets republished with a replacement description, claiming it to depict some different incidence. To curb such fabricated stories, it is necessary to investigate such deduplicates and false claims made in public. Most studies on automatic fact-checking and fake news

---

<sup>2</sup><https://www.tribuneindia.com>

<sup>3</sup><https://www.timesnownews.com>

<sup>4</sup><https://www.thestatesman.com>

<sup>5</sup><https://www.ndtv.com>

<sup>6</sup><https://www.dnaindia.com>

<sup>7</sup><https://indianexpress.com>

<sup>8</sup><https://www.altnews.in>

<sup>9</sup><https://www.boomlive.in>

<sup>10</sup><https://digiteye.in>

<sup>11</sup><https://thelogicalindian.com>

<sup>12</sup><https://newsmobile.in>

<sup>13</sup><https://www.indiatoday.in>

<sup>14</sup><https://newsmeter.in>

<sup>15</sup><https://english.factcrescendo.com>

<sup>16</sup><https://www.youtube.com/channel/UCODbDJz3vs2Cru9xuq1HbUA>

<sup>17</sup><https://www.afp.com/en>

detection are restricted to English only. However, for a country like India, where only 10.67% of the literate population speaks English<sup>18</sup>, the role of regional languages in spreading falsity cannot be undermined. This chapter aims to present a data repository of fake news stories across India and curate instances covering multilingual aspects.

Hence, we present FactDrill, the first large-scale multilingual fact-checking dataset for regional Indian languages. Further, using the dataset, one can investigate the changing dynamics of fake news in a multilingual setting in India. The resource would aid in examining the fake news at its core, i.e. investigating the different kinds of stories being disseminated, the modalities or combinations used to create the fabricated content and the platform used for dissemination.

### 7.3 FactDrill: A Data Repository of Fact-Checked Social Media Content to Study Fake News Incidents in India

We collect an exhaustive dataset across seven months covering eleven low-resource languages and the two major lingua franca of the country, i.e. Hindi and English. Our proposed FactDrill dataset consists of 9,058 samples belonging to English, 5,155 samples to Hindi and the remaining 8,222 samples are distributed across various regional languages, i.e. Bangla, Marathi, Malayalam, Telugu, Tamil, Oriya, Assamese, Punjabi, Urdu, Sinhala and Burmese. FactDrill is unique due to the following reasons:

- **Multilingual Information:** There are 22 official languages in India. The 2011 Census of India<sup>19</sup> shows that the languages by the highest number of speakers (in decreasing order) are as follows: Hindi, Bengali, Marathi, Telugu, Tamil, Gujarati, Urdu, Odia, Malayalam, and Punjabi. Though the current datasets are in English, the above statistics indicate a need to shift fake news from English to other languages. Hence, FactDrill consists of news samples that span over 13 different languages spoken in India.
- **Investigation reasoning:** With the FactDrill dataset, we present an attribute that explains how the manual fact-checkers carry out the investigation. Figure 7.1 shows a screenshot of the attribute taken from a sample of the Boom website.<sup>20</sup> We believe providing such information can give insights about the news story like, (i) social media account or website that posted the fake content (highlighted in yellow), (ii) platform that first encountered the fake content (highlighted in orange), (iii) links to the archive version of the post if the original content is deleted (highlighted in green), (iv) tools used by fact-checkers to investigate the claim (highlighted in pink), and (v) links to the supporting or refuting reports related to the claim (highlighted in blue). Such insights

<sup>18</sup>[https://en.wikipedia.org/wiki/2011\\_Census\\_of\\_India](https://en.wikipedia.org/wiki/2011_Census_of_India)

<sup>19</sup>[https://en.wikipedia.org/wiki/Multilingualism\\_in\\_India](https://en.wikipedia.org/wiki/Multilingualism_in_India)

<sup>20</sup><https://www.boomlive.in>

have the potential to drive the research towards studying the ‘*Nature of fake news production*’ in general. The attribute is exclusive to the FactDrill dataset.

### FACT-CHECK

BOOM found that the viral forward is from a satire article that was published by 'Daily Expose', a website that has previously put out misinformation on COVID-19 vaccines.

Taking a hint from the forward as it states that its from "Daily Expose", we searched with the relevant keywords, we found the original article on the website. The headline of the article is, "SATIRE – In an alternative universe Bill Gates has called for the withdrawal of all Covid-19 Vaccines"

Daily Expose has now put an editorial note which reads, "Note from The Editor – when we first published this article we should have made it clear at the beginning that it was satire rather than at the end. We did not do this and we apologise..."



Click [here](#) to view an archive

Several fact-checkers including PolitiFact and Full Fact have fact-checked The Expose for misinformation around COVID-19 vaccines. PolitiFact had previously fact-checked the same viral forward in August 2021 when was being shared with the false claim.

Additionally, we did not find any such speech by Bill Gates against COVID-19 vaccines or any credible news reporting the claims made in the viral post.

Figure 7.1: An excerpt from our proposed FactDrill dataset depicting the *investigation\_reasoning* attribute. The attribute is exclusive of the FactDrill dataset and is not present in any existing fact-checking datasets. The attribute provides minute details of the fact-checking process, i.e. social media account or website that posted the fake content (highlighted in yellow), the platform that first encountered the fake content (highlighted in orange), links to the archive version of the post if the original content is deleted (highlighted in green), tools used by fact-checkers to investigate the claim (highlighted in pink), and links to the supporting or refuting reports related to the claim (highlighted in blue).

- Multi-media and multi-platform information: Fake news can be published in any form and on any social and mainstream platform. The curated dataset incorporates the information about media (images, text, video, audio, or social media post) used in fake news generation and the medium (Twitter, Facebook, WhatsApp, and Youtube) used for its dissemination.
- Multi-domain information: The previous fact-checking dataset covers information on specific topics only. For example, Emergent [76] only captures the national, technolog-

ical, and world related happening in the US whereas [9, 275] include health, economic, and election-related issues. In our proposed dataset, we have curated information from the existing fact-checking websites in India, giving us leverage to capture news stories of different topics and cover events that happened during the time frame.

### 7.3.1 Data Curation

We curate the first large-scale multilingual Indian fact-checking data to the best of our knowledge. Figure 7.2 shows the complete data curation process. In this section, we discuss the key steps of data curation, i.e. data collection, data extraction and data annotation.

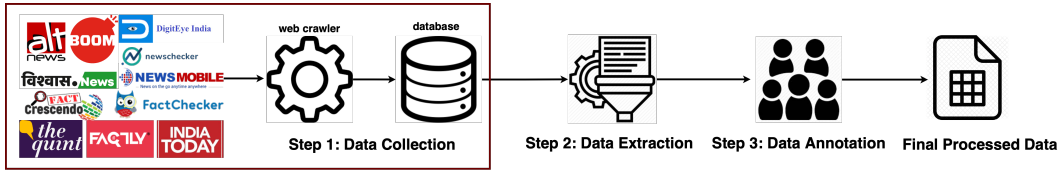


Figure 7.2: Our proposed dataset curation pipeline. Step 1 describes the data collection process. This is followed by Step 2, describing the data extraction methodology, and Step 3, discussing the data annotation and evaluation process.

#### Step1: Data Collection

Though fact-checking services play a pivotal role in combating fake online content, little is known about whether users can rely on them. To corroborate trust among the audience, fact-checking services should ensure transparency in their processes, organizations, and funding sources. To look out for trusted Indian fact-checking websites, we came across International Fact-Checking Network (IFCN).

IFCN is owned by the Poynter Institute of Medical Studies, located in St. Petersburg, Florida. It was set in motion in September 2015. The primary objective of establishing IFCN was to bring together fact-checkers across the globe under one roof. It also provides mandatory guidelines for fact-checking organizations to gain the license. The legally registered organizations routinely scrutinize the statements made by public figures. The statements can be text, visual, audio and other formats mainly related to public interest issues. On the other hand, organizations whose opinions are influenced by the state, any other influential identity, or a party are generally not admitted to the grant. To be eligible for an IFCN signatory, an organization is critiqued by independent assessors on 31 criteria. The assessment is then finally reviewed by the IFCN advisory board to ensure fairness and consistency across the network. There are about 82 Verified signatories of the IFCN code of principles, among which 11 are based in India.<sup>21</sup> To ensure the authenticity and verifiability of the curated data,

<sup>21</sup>As per 2020 statistics.

we have considered those Indian fact-checking sources that are IFCN rated verified.

The prime benefit of gathering data from fact-checking websites is that we can read the reasoning behind the veracity of a news sample. The detailed description of the investigation gives valuable insight to the reader about how and why the viral claim was false. With this objective in mind, we decided to collect data from the existing fact-checking websites that aim to debunk fake information across India. Table 7.1 provides an overview of the fact-checking websites considered for data curation. The table highlights the key features of a particular website in the form of (i) organization establishment year, (ii) languages debunked by the website, and (iii) domain covered.

Website	Establishment	Languages Supported	Domain
Alt News	Feb 2017	English, Hindi	Politics, Science, Religion, Society
Boom Live	Nov 2016	English, Hindi, Bangla, Burmese	General
DigitEye India	Nov 2018	English	General
FactChecker	Feb 2014	English	General, Modified
Fact Crescendo	July 2018	English, Hindi, Tamil, Telugu, Kannada, Urdu, Oriya, Assamese, Punjabi, Bengali, Marathi, Gujarati, Malayalam,	General, Coronavirus
Factly	Dec 2014	English, Telugu	General, Coronavirus
India Today		English	General
News Mobile	2014	English	General
NewsChecker		English, Hindi, Marathi, Punjabi, Gujrati, Tamil, Urdu, Bengal	General
Vishvas News		English, Hindi, Punjabi, Odia, Assamese, Gujrati, Urdu, Tamil, Telugu, Malayalam, Marathi	Coronavirus, Politics, Society, World, Viral, Health
The Quint- Webqoof		English, Hindi	General, Health

Table 7.1: An overview of the fact-checking sources considered during the data collection. Empty cells indicate that the information is not available on the website.

## Step 2: Data Extraction

We set up a data extraction system that uses a Python library, BeautifulSoup<sup>22</sup> to extract data from web pages. Our system checks the sources for new data once in 24 hours. This thesis presents a study on samples curated from 2013 to 2020. By the end of the data curation process, we had 22,435 news samples. Among them, 9,058 samples belong to English, 5,155 samples to Hindi and the remaining 8,222 samples are distributed in various regional languages, i.e. Bangla, Marathi, Malayalam, Telugu, Tamil, Oriya, Assamese, Punjabi, Urdu, Sinhala, and Burmese.

The screenshot shows a fact-checking article on the Vishvas News website. The article title is "Fact Check: This Picture Of Delhi CM Kejriwal Helping A Child Wear Mask Is From 2019; No Relation To Covid 19". The article is dated January 12, 2022. The article text states that a picture of Delhi CM Arvind Kejriwal helping a child wear a mask is from 2019 and has no relation to COVID-19. The article includes a link to the original picture and a link to the tweet. The article also includes a conclusion stating that the viral picture of Delhi Chief Minister Arvind Kejriwal helping a child wear a mask is not from the Corona period. The article is tagged with "CoronaVirusFacts", "Fact Check", and "Fact Check".

Annotations on the screenshot include:

- website\_name**: Points to the Vishvas News logo.
- article\_link**: Points to the URL in the browser address bar.
- unique\_id**: Points to the article's unique identifier.
- publish\_date**: Points to the date "January 12, 2022".
- domain**: Points to the website's domain.
- title**: Points to the article title.
- image**: Points to the main image of the article.
- claim**: Points to the text describing the viral claim.
- investigation**: Points to the text describing the investigation process.
- tweet**: Points to a tweet from Arvind Kejriwal.
- links\_in\_text**: Points to links within the article text.
- bold\_text**: Points to bold text within the article.
- tags**: Points to the article tags.

Figure 7.3: We present a screenshot from a fact-checking website (Vishvas News) to depict different attributes present in the proposed FactDrill dataset.

<sup>22</sup><https://pypi.org/project/beautifulsoup4/>

## Dataset Attributes

We curate numerous features from unstructured data. Figure 7.3 presents a sample showcasing all the attributes. We have categorized the extracted feature set into various classes like meta-features, textual features, media features, social features, and event features. The final processed data is shown in Figure 7.4.

1. **Meta Features:** We consider those attributes as meta\_features that tell us about the sample, like (i) *website\_name* that denotes the name of the source from where the sample is collected, (ii) *article\_link* that denotes the name of the source from where the sample is collected, (iii) *unique\_id*, an attribute acts as the primary key for data storage and (iv) *publish\_date* which signifies the date on which the fact-checking websites published the article.
2. **Textual Features:** A fact-checked article can be segregated into title and content. The content attribute can further be divided into claim and investigation. All three attributes together form the textual features in FactDrill. Since the curated data from the website is highly unstructured, information in claims and investigation is generally present in the content part of the data. This information is extracted from the content attribute using human intervention. This is discussed in detail in Section 7.3.1. The following are the textual features in the FactDrill: (i) *title* of the article, (ii) *content* that acts as the body of the article that consists of information in the form of claim and investigation, (iii) *claim*, an attribute that gives the reader background information about *what was said in the related post* and, (iv) *investigation* help readers in understanding *why the fact-checkers concluded a particular post to be fake*. The complete inspection process is discussed in detail, with tools and techniques used to explore.
3. **Media Features:** The claim viral on any social media platform or mainstream media might have many modalities. Similarly, the investigation to conclude the status of any viral news can also be backed by numerous supporting claims containing information in multimedia form. The set of attributes that fall under the multimodal feature set is as follows: (i) *image\_links* depicting the links of all other images that will either belong to the original claimed images group or are presented in support of the viral claim are put under this feature as a list object, (ii) *video\_links*, for those samples where prime media used for fabrication is video, the link to the original video is provided by fact-checkers to back their investigation. This attribute stores all such links, (iii) the *audio\_links* attribute presents all the supporting audio links related to the viral claim, (iv) To provide complete justification to what was said in the investigation report, fact-checkers provide different media links in support of their investigation. All such links are present in the attribute named *links\_in\_text*. However, to identify where a specific link is mentioned in the fact-checked article, an attribute named *bold\_text* is used for easy identification and matching of the corresponding text from the article.



4. Social Features: The attribute stores the tweet ids present in the sample. The tweet ids can be the post that (i) needs to be investigated or (ii) is present in support of the fake claim. We can extract the complete information from the tweet thread with this attribute.
5. Event Features: The set of features in this group gives information about the event to which a news sample belongs. These include topic and tag attributes. For instance, the Boom article titled: ‘False: Chinese Intelligence Officer Reveals Coronavirus Is A Bioweapon’ had the following tags (*Coronavirus China*, *COVID-19*, *Coronavirus outbreak*, *Bioweapon*, *Biological warfare*, *China*, *Intelligence Officer*) associated with it. This kind of information helps identify the topic of the article.

Meta Features		Social Features	Event Information
<b>website_name:</b> alt_news_english <b>article_link:</b> <a href="https://www.altnews.in/2018-image-of-muslims-offering-prayers-in-up-shared-as-recent-from-tamil-nadu/">https://www.altnews.in/2018-image-of-muslims-offering-prayers-in-up-shared-as-recent-from-tamil-nadu/</a> <b>unique_id:</b> 2018-image-of-muslims-offering-prayers-in-up-shared-as-recent-from-tamil-nadu <b>publish_date:</b> 02-05-2020 00:00:00		<b>tweet_id:</b> 1256281042131881984	<b>domain:</b> Religion
<b>Textual Features</b> <b>Title:</b> 2018 image of Muslims offering prayers in UP shared as recent from tamil Nadu <b>Claim:</b> An image of a group of men, wearing a skull cap, offering namaz is getting shared on social media. It is being shared with the claim that a group of 700 men offered namaz amid the lockdown in Tamil Nadu's Vellore district. Facebook user Vijay Ajay posted the image with the claim in Tamil on April 30. The image is credited to 'Alamy stock photo' and 'Jaya Murgan'. The text in Tamil reads, "மேலூர் மாவட்டம் திருப்பத்தூர் டவுனில், ஜும்மா மசூதி தொகுதில் கடந்த இரண்டு நாட்களாக, நாளிரவு 1 மணிக்கு நடு ரோட்டிலேயே சுமார் 700 பேர்கள் தொழுகையில் ஈடுபட்டு வருகின்றனர். காலத்துறையினை உயரதிகாரிகளின் உத்தரவிக்கு கட்டுப்பட்டு, அவர்களுக்கு எந்த தொந்தரவும்!!!! மற்றும் அங்கு இரவு பணியில் இருக்கும் காவலர்கள் எவரும் தமது தொலைபேசியில் புகைப்படமோ அல்லது வீடியோவை எடுக்கக்கூடாது!!!! என்ற உயர் அதிகாரியின் உத்தரவுக்கு கட்டுப்பட்டு கையை பிசைந்து நிற்கின்றனர்.சம்மந்தப்பட்ட காவல்துறையை சேர்ந்த நண்பர் ஒருவரின் மகக்குழந்தை நனதி : Jaya Murgan நனதி : alamy stock photo". Another Facebook page with the name Bjp Coimbatore Thondamuthur Assembly posted the image with the same claim. <b>Investigation:</b> With a Google reverse image search, Alt News found that the image is from Uttar Pradesh's Allahabad city, now known as Prayagraj. According to the stock photo agency Alamy, the image was shot on May 17, 2018. "Muslims offer night prayers called taraweeh during Ramadan month in Allahabad, Ramadhan is the ninth month of the Islamic calendar, and the month in which the Quran was revealed. Fasting during the month of Ramadan is one of the Five Pillars of Islam. The month is spent by Muslims fasting during the daylight hours from dawn to sunset," reads the caption. The image is credited to a person named Prabhat Kumar Verma. Tamil Nadu's Tirupathur district police also took to Twitter and called out the misinformation circulated on social media. "Picture shot in #allahabad is falsely shared in social media as one taken in #tirupathur district.FIR has been registered against the miscreants," said the tweet. In conclusion, a 2018 image from Uttar Pradesh was falsely shared to claim that Muslims offered namaz in Tamil Nadu's Vellore district despite an ongoing nationwide lockdown. Earlier this week, a video from Mumbai shot on March 23 outside a mosque in Dongri was shared with the false claim that it showed Muslims gathering in Ahmedabad's Jamalpur area amidst the lockdown.		<b>Media Features</b> <b>image_links:</b> ['https://www.altnews.in/wp-content/uploads/2020/05/2020-05-02-18_51_29-3-Facebook.png?resize=380%2C638', 'https://static.xx.fbcdn.net/images/emoji.php/v9/9b/1.5/16/162f.png', 'www.altnews.in/wp-content/uploads/2020/05/2020-05-02-18_41_55-3-Facebook.jpg?resize=511%2C627', <b>video_links:</b> None <b>audio_links:</b> None <b>links_in_text:</b> ['https://www.facebook.com/story.php?story_fbid=599996507340457&id=100025903388426', 'https://www.facebook.com/142425632589930/posts/1591460311010781/', 'https://twitter.com/hashtag/allahabad?src=hashtag_click', 'https://twitter.com/hashtag/tirupathur?src=hashtag_click', 'https://twitter.com/hashtag/allahabad?src=hashtag_ref_src=twsrc%5Btfw', 'https://twitter.com/hashtag/tirupathur?src=hashtag_ref_src=twsrc%5Btfw', 'https://twitter.com/hashtag/fakenews?src=hashtag_ref_src=twsrc%5Btfw', 'https://twitter.com/hashtag/tnpolice?src=hashtag_ref_src=twsrc%5Btfw', 'https://t.co/9Xkt102ba0', 'https://twitter.com/sp_tirupathur/status/1256281042131881984?ref_src=twsrc%5Btfw', 'https://www.altnews.in/march-23-video-from-maharashtra-shared-as-muslims-gathering-at-mosque-in-ahmedabad-amid-lockdown/', 'https://www.instagram.com/altnews/'], <b>bold_text:</b> ['posted', 'Bjp Coimbatore Thondamuthur Assembly', '#allahabad', '#tirupathur', '#allahabad', '#tirupathur', '#fakenews', '#tnpolice', 'pic.twitter.com/9Xkt102ba0', 'May 1, 2020', 'shared']	

Figure 7.4: A excerpt from the dataset displaying different attributes present in a sample of the proposed *FactDrill* dataset. The feature list is paced under different headers namely, meta features, text features, social features, media features, and event information. The attributes are discussed in Section 7.3.1.

### Step 3: Data Annotation

This section addresses the three key questions that facilitate the data annotation process.

**Description of the Annotation Tasks:** We divide the complete annotation process into two tasks. In Task 1, annotators have to mark the sample as fake or non-relevant. The non-relevant subset includes samples (i) that were investigated to be true, (ii) articles containing general fact information that news websites usually publish<sup>23</sup> and, (iii) weekly-wrap up

<sup>23</sup><https://www.boomlive.in/technologies-will-tackle-irrigation-inefficiencies-agricultures-drier-future/>



articles that increase the chance of duplication in the dataset. In Task 2, the annotators are provided with three attributes: content, claim and investigation. The *content* attribute is already divided into claim and investigation using a keyword-based heuristic method. The role of the annotator is to check whether the segregation performed is correct or not. If not, the annotator has to place text under the correct header.

**Annotation Process:** We hired language experts to annotate the samples. For Hindi and English languages, each sample is annotated by two annotators. However, due to the limited expertise in the regional languages, each sample is annotated by a single annotator. The annotators are provided with annotation guidelines that include instructions about each task, a definition of the attributes that need to derive from the text and a few examples. The annotators first study the document and work on a few examples to familiarize themselves with the task. They are then given feedback on the sample annotations, which helped them refine their performance on the remaining subset.

Website	Language	Inter Annotator Agreement Score		# Samples
		Task 2 (Percent Agreement)	Task 1 Cohen's Kappa /Gwet's AC(1) AC(2)	
Alt News	English	0.78	0.48	2058
	Hindi	0.76	0.53	1758
Boom	English	0.42	0.66	909
	Hindi	0.90	0.53	880
DigitEye	English	0.86	0.56	147
FactChecker	English	0.31	0.15	156
Fact Crescendo	English	1.00	<b>1.00</b>	256
	Hindi	0.99	<b>1.00</b>	264
Factly	English	0.92	0.76	971
India Today	English	0.95	0.44	788
News Mobile	English	0.71	0.29	1543
Vishvas News	English	0.94	<b>0.91</b>	254
	Hindi	0.98	<b>0.90</b>	1369
Webqoof	English	0.86	0.47	1771
	Hindi	0.95	<b>0.97</b>	328

Table 7.2: Inter-annotator agreement for the two tasks. The values in bold indicates that Gwet's AC(1) and AC(2) scores are calculated for the samples. We observe a mix of moderate (0.41-0.60) and substantial (0.61-0.80) agreement for the majority of the samples.

**Annotation Evaluation Metric:** To evaluate the performance of Task 1, we calculate the inter-annotator agreements using Cohen's Kappa [163]. We observe a mix of moderate

and substantial agreement for most samples. Table 7.2 summarizes Cohen’s kappa measures. Although Cohen’s Kappa performs exceptionally well when a dichotomous decision is involved and takes care of chance agreement, it fails badly when annotators show near 100% agreement. This phenomenon is termed as ‘the paradoxes of Kappa’. During our evaluation, we observe high agreement between annotators for 1000 samples. To solve the ‘the paradoxes of Kappa’ issue, we used Gwet’s AC(1) and AC(2) statistic [98]. It overcomes the paradox of high agreement and low-reliability coefficients. Table 7.2 summarizes Gwet’s score for the subset. To evaluate the performance for Task 2, we checked for matched ordinal positions for each annotated sample. The final inter-agreement score is computed using the percent agreement for two raters [258]. It is calculated by dividing the total count of the matched sample by the total number of samples in the data.

### 7.3.2 Basic Dataset Characterization

We begin by providing a statistical overview of our proposed dataset.

#### Summary Statistics

Figure 7.5 (a) shows the distribution of samples across languages in our proposed FactDrill dataset. The diffusion of samples in the regional interface is majorly dominant by Bangla,

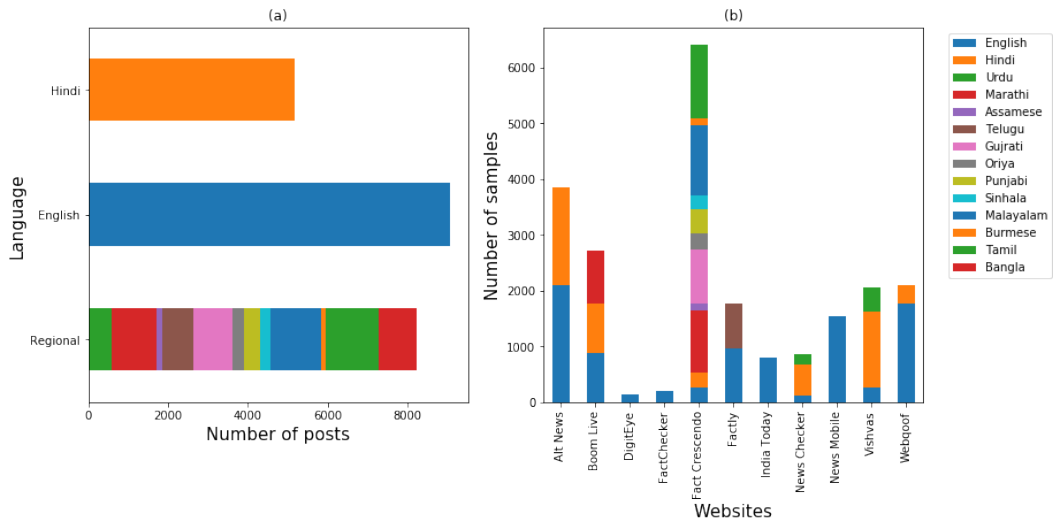


Figure 7.5: (a) shows the distribution of different languages in our proposed FactDrill dataset, (b) shows the spread of data across different fact-checking websites. It also depicts the language supported by each website.

Malayalam, Urdu and, Marathi language. Figure 7.5 (b) represents the number of samples belonging to the fact-checking websites. Among them Fact Crescendo website rules in debunking fake news dissemination in different languages.

## Popular Fake Events in India

We analyze the topic distribution of fact-checking articles in different languages, i.e. English, Hindi and Regional languages. All the tags and domain knowledge for the Hindi and regional languages are present in English only. From Figure 7.6, we can conclude that political activity is essential for fake news creation in all languages. With the onset of 2020, the world witnessed a global pandemic, i.e. Coronavirus, that has affected peoples' lives and given rise to the infodemic of fake content. Unsurprisingly, the second popular topic for fake news creation in India was Coronavirus, followed by health and religion.



Figure 7.6: Topic Distribution in English, Hindi and Regional languages (left to right). The figures clearly show that most fake news dissemination across the country is centred on the political domain.

## Circulation of Fake News in India

Figure 7.7 (a) shows the percentage increase in production of fake news over the years. The fact-checking trend came to India in 2013, majorly debunking news in the English language. As and when fake news dissemination in English got popular, we saw it intruding into other languages too. This steady shift to other languages was observed in 2017 and 2018. We observe sharp peaks and drops in the graph that will be an exciting study in the future. Figure 7.7 (b) shows the year-wise distribution of samples in the dataset. The graph shows a

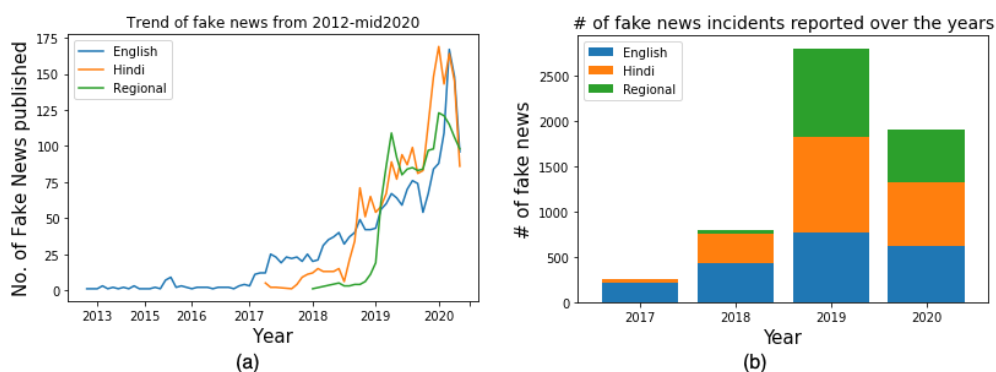


Figure 7.7: Circulation of fake news over the years in India. Figure (a) demonstrates sharp peaks and drops in the graph that will be an exciting study in the future. Figure (b) shows the year-wise distribution of samples in the dataset. During the data collection stage, the last sample collected was in June 2020. Hence, the sample count is shown till June.

steady increase in fake news production, with a major peak observed in 2019. For both these observations, the data considered for 2020 is till June.<sup>24</sup>

### 7.3.3 Use Cases

There are various threads of fake news research that can be initiated with the help of FactDrill dataset.

- **Understanding the nature of fake news:** Various efforts have been made to date to eliminate fake news on the Internet. There are two primary drawbacks to such approaches, (i) The system performs well on trained samples but fails drastically on real-world data, and (ii) the performance of classifiers varies considerably based on the evaluation archetype and performance metric [30]. We believe there is a need to study the nature of fake news before attempting to detect it. Towards this end, we present a dataset that provides a detailed investigation of the fake sample that includes (i) the modality faked in the news, (ii) ‘how’ the sample was concluded to be false and (iii) tools used for conclusion.
- **Suppressing Fake News Dissemination at an Early Stage:** Fact-checkers are making a constant effort to debunk false information online. However, we still witness duplicates and republish content online. This demonstrates that fact-checking initiatives are not reaching the general public. With the FactDrill dataset, we can develop technologies stationed at different social media platforms; such systems can use information from the debunked news sample and make it available to the audiences on the platform.
- **Bias among fact-checkers:** Fact-checking is tedious. Different websites aim to debunk news on different topics. There can be websites that aim at exposing a particular kind of information. It will be interesting to look for biases in the fact-checking pattern and its related effects.
- **Modelling Temporal Progression:** FactDrill dataset consists of data that spans from the year 2013 to the year 2020. It can serve as an excellent source to study the evolution of fake news over the years.
- **Event-centric Studies:** FactDrill dataset comprises news stories that span different events across the timeline. For instance, it had fake news stories busted during the CAA, NRC Bill, and COVID-19 pandemic, to name a few. The proposed dataset can be used to study the impact of fake news dissemination during such popular events in the country.
- **Exploring the Multilingual Fake News Direction:** FactDrill dataset comprises news stories spanning over thirteen languages spoken in the country. The proposed dataset will

---

<sup>24</sup>During the data collection stage, the last sample collected was in June 2020.

help in designing automatic detection and language identification systems. Moreover, data in multiple languages can further open up research opportunities in the Natural Language Processing (NLP) domain.

- Challenge Proposal: We want to extend our work as a challenge proposal to dig deep into studying the fake news patterns in India.

### 7.3.4 FAIR Principles and Limitations

Our proposed FactDrill dataset adheres to the four FAIR data principles: Findable, Accessible, Interoperable and Reusable, as follows: (i) The complete dataset is publicly available at the following link: <https://precog.iiit.ac.in/requester.php?dataset=FactDrill> making FactDrill dataset easily Findable and Accessible, (ii) We provide the proposed dataset in an xlsx (Excel Spreadsheet) format that can be viewed and parsed easily. Additionally, we give a thorough description of each attribute in the FactDrill dataset in Section 7.3.1. Further, the dataset can be exported to other data formats like CSV (Comma Separated Values), making the FactDrill dataset Interoperable and Reusable.

Next, there are a few limitations to running FactDrill continually: (i) Time and again, websites change their web layout, breaking the Python scripts running in the backend, (ii) Pattern matching algorithms can be challenging as there is no specific demarcation of the content into claim and investigation by the website authorities, and (iii) Processing raw data in different languages is tedious, and finding experts in the different regional languages becomes hard.

### 7.3.5 Conclusion

This thesis presents FactDrill: a data repository of fact-checked social media content to study fake news incidents in India. To the best of our knowledge, this is the first large-scale multilingual Indian fact-checking data that provides fact-checked stories for thirteen languages spoken in the country. We believe such a dataset can aid researchers in exploring the fake news spread in regional languages. Researchers could also look out for the dissemination of fake content across the different language silos. The FactDrill dataset comprises 22,435 samples from the IFCN-rated Indian fact-checking websites. Fourteen features associated with each sample are grouped under meta, textual, media, social, and event features. We also present a new attribute to the feature list, i.e. *investigation reasoning*, and explain its relevance and need in the current fact-checking mechanism.

## 7.4 Validating Fake News Detection Methods

In the preceding section, our focus was primarily on examining the informational aspect of fake news. However, in this section, our attention turns towards a comprehensive examination of fake news datasets and the methods employed to study them. Despite extensive efforts to develop solutions for combating fake news online, none of the existing methods claim

to provide real-time solutions. This limitation can be attributed to two main factors. Firstly, these methods often struggle to adapt to the rapidly changing dynamics that characterize fake news. Secondly, the perceived effectiveness of these methods may be influenced by systematic biases that exist within the dataset and models. Consequently, these biases can cause a reliance on specific preferences without adequately considering actual cues that are crucial for accurate detection.

#### 7.4.1 Research Objective

Research investigating the composition of text datasets has revealed that advanced models can exhibit a tendency to prioritize “shortcut” and rely heavily on token distributions rather than truly comprehending language. When models are trained on datasets that contain a significant proportion of highly effective cues, there is a risk that they may learn to associate these cues with specific labels, thereby disregarding the linguistic information present. In order to assess this phenomenon, we evaluate fake news datasets using several metrics to identify potential cues or linguistic artifacts that the model may rely upon. Additionally, we employ dataset ablations to validate the accuracy of these cues and their impact on the model’s performance.

#### 7.4.2 Methodology

To test the validity and effectiveness of our detection methods, we are conducting cues and ablation studies. In the cues studies, we analyze the presence of specific cues or linguistic artifacts in the dataset that the models may rely upon for detecting fake news. This analysis allows us to determine the extent to which these cues contribute to the detection accuracy. Additionally, we employ ablation studies to evaluate the impact of removing certain cues or linguistic elements from the dataset on the performance of the detection methods. These studies enable us to assess the robustness and reliability of the methods by examining how their performance is affected when key cues or linguistic features are altered or eliminated. By conducting these studies, we aim to gain insights into the underlying mechanisms of the detection methods and validate their efficacy in identifying fake news.

##### Cues

Unigram, Bigram, and Trigram sized tokens of each sample are extracted and evaluated based on the following metrics.

1. *Applicability*: Given a token unique to a single class, the *Applicability* of the token is the count of data samples which contain that token. Mathematically, the *Applicability* for a token  $k$  is defined in Equation 7.1. Intuitively, this metric speaks to potential cues for the model to associate with that class.

$$\alpha_k = \sum_{i=1}^n 1 \left[ \exists j, k \in T_j^{(i)} \wedge k \notin T_{\neg j}^{(i)} \right] \quad (7.1)$$

Here,  $T_j^{(i)}$  is set of all tokens in subset for data point  $i$  with label  $j$ .

2. **Productivity:** Given a token unique to a single class, the *Productivity* of the token is the proportion of data samples for which the model predicts the correct answer, relative to the *Applicability* of the token. Therefore, the *Productivity* of a token is a metric for how useful the model would find it to be. *Productivity*  $\pi_k$  can be calculated as defined in Equation 7.2.

$$\pi_k = \frac{\sum_{i=1}^n 1 \left[ \exists j, k \in T_j^{(i)} \wedge k \notin T_{\neg j}^{(i)} \wedge y_i = j \right]}{\alpha_k} \quad (7.2)$$

3. **Coverage:** Coverage is simply the proportion of *Applicability* relative to the total number of rows in the dataset.

$$C_k = \frac{\alpha_k}{n} \quad (7.3)$$

4. **Strength:** The strength of a cue is defined as the product of its *Coverage* and *Productivity*.

$$S_k = C_k \times \pi_k \quad (7.4)$$

## Data Ablation

As discussed, cues across labels can distract models from learning context or meaningful information from the data. To verify the extent to which this has occurred, we perform data ablation tests.

1. **Scramble Word Order:** The tokens or words for each sample in the test set are randomly shuffled. A drop in test set accuracy indicates that the model depends upon the sequential nature of words/cues to at least some extent. If the test set performance does not change significantly, this would potentially indicate that the model has effectively learned a bag-of-words style classification.
2. **Shuffle Labels:** The class labels are randomly shuffled, and the model is retrained. If test set performance does not change significantly, this indicates that the model has not learned to associate contextual cues with each label. This is the only ablation test that involves model retraining instead of perturbing inputs of the previously trained model.
3. **Partial Input:** Tokens from the samples are removed with a probability of (0.5), but the sequential order of the remaining tokens is not disturbed. If test set performance does not drop, this indicates that the model depends upon a subset of tokens within each sample to make a prediction. In general, for a model to be learning linguistic information from the dataset, we would expect the following: (i) The data should contain fewer cues of considerable strength and (ii) Model performance should drop appreciably with each ablation.

With the above assumptions, we perform a robust set of experiments and analysis using the methods described in this section, as explained next.

### 7.4.3 Results

In this section, we will discuss the classification results obtained from our analysis, as well as the findings from the cues study.

#### Classification Model

In our experimental setup, we leverage the FakeNewsNet repository [227] to facilitate our research. To accommodate the requirements of the Productivity metric (Equation 7.2), which relies on model predictions, we select the N-gram Convolutional Neural Network (CNN) as our chosen model for conducting the experiments. The performance of CNN on the FakeNewsNet repository is shown in Table 7.3. The choice of model allows us to effectively analyze and evaluate the cues introduced by the metric and assess their impact on the overall productivity measurement.

Datasets	Accuracy	F1-Score
Politifact	90.38%	0.9367
Gossip	87.73%	0.9276

Table 7.3: The table presents the performances of the N-gram Convolutional Neural Network (CNN) on the FakeNewsNet repository, highlighting the effectiveness of the model in detecting fake news. The CNN achieves an accuracy of 90.38% on the Politifact dataset and 87.73% on the Gossipcop dataset.

#### Evaluation of Fake News Datasets

Subsequently, we proceed to extract cues from each dataset and calculate their Productivity, Coverage, and Strength. Cues with non-zero Applicability (Equation 7.1) or Coverage (Equation 7.3) primarily serve as characteristic cues for each label. In the case of the FakeNewsNet repository, a significant portion of cues demonstrates an inverse relationship between Productivity and Coverage. This indicates that most cues are either widely applicable (high Applicability) or highly influential for the model’s predictions (high Productivity). In Table 7.4, we present the top ten cues by strength for each class across the datasets. Interestingly, the strongest cues for both the real and fake classes appear to be rather random in nature.

#### Model Sensitivity Analysis (Ablation Study)

The presence of cues across different labels has the potential to divert the attention of models from effectively learning contextual or meaningful information from the data. In line with the approach proposed by Heinzerling [104], we conduct data ablation tests to evaluate the impact of these cues on model performance. The results are shown in Table 7.5.



Dataset	Class	Top 10 Cues
GossipCop	Real	<p>excerpt read  excerpt read online  news submit  wake breaking  breaking news submit  wake breaking news  read online wake  online wake breaking  watch clip  brush</p>
	Fake	<p>visit source  content published entertainment  users news  completely factual  posts represent  guarantee reporting  content accuracy  report concerns content  imdb's opinions guarantee  inc takes responsibility</p>
Politifact	Real	<p>challenge  fair  progress  oil  somebody  defici  debt  nt think  rates  d</p>
	Fake	<p>antitrump  confirming  bombshell  according reports  trump tower  black lives  reveals  daily mail  Michael flynn  devin</p>

Table 7.4: The top 10 cues by strength for each class across datasets. The strongest cues for both the real and fake classes appear to be fairly random, suggesting that the model considers a wide range of cues, possibly including irrelevant ones, when making decisions.

The first ablation test involves scrambling the word order within each sample in the test set, thereby disrupting the sequential nature of the words or cues. As shown in Table 7.5, we observe a notable decrease in test set accuracy, indicating that the model relies, to some extent, on the sequential arrangement of words or cues for classification. Conversely, if the performance had not significantly changed, it would have suggested that the model has learned to classify in a bag-of-words style, without considering word order.

Another ablation test entails shuffling the class labels randomly and retraining the model. As depicted in Table 7.5, we observe a decrease in test set accuracy, indicating that the model has learned to associate specific contextual cues with each label. Unlike the previous tests that perturbed inputs of the pre-trained model, this particular test involves retraining the model with shuffled class labels.

Lastly, we perform a partial input ablation, where tokens from the samples are randomly removed with a probability of 0.5, while maintaining the sequential order of the remaining tokens. The results in Table 7.5 demonstrate a minimal decrease in test set accuracy, suggesting that the model depends on a subset of tokens within each sample for making predictions.

Overall, the findings from the ablation tests presented in Table 7.5 provide valuable insights into the model’s reliance on cues and linguistic information within the dataset. The noticeable drop in performance observed with each ablation indicates a strong dependence of the model on cues. This sensitivity of the models to the ablations further suggests that they consider the entirety of the message rather than relying solely on statistical artefacts, highlighting their reliance on linguistic information for making accurate predictions. These results contribute to understanding the model’s behaviour and emphasize the importance of considering linguistic cues in fake news detection.

<b>Ablation</b>	<b>F1-Score</b>	
	<b>Politifact</b>	<b>GossipCop</b>
<b>No Ablation</b>	0.9367	0.9276
<b>Shuffle Label</b>	0.7468 (-0.1899)	0.8822 (-0.0454)
<b>Scramble Words</b>	0.8379 (-0.098)	0.8825 (-0.045)
<b>Partial Input (20% of original)</b>	0.9299 (-0.0067)	0.9125 (-0.015)

Table 7.5: The ablation study results highlight the model’s reliance on cues and linguistic information within the dataset. The significant drop in performance observed with each ablation indicates a strong dependence of the model on these cues. This suggests that the model heavily relies on specific patterns and features present in the data to make accurate predictions.

#### 7.4.4 Conclusion

This section focuses on the validation of fake news datasets and detection methods. We hypothesize that models tend to rely on token distributions rather than truly understanding the language. To investigate this phenomenon, we employ various metrics, such as productivity, applicability, coverage, and strength, to evaluate the fake news datasets. The goal is to identify potential cues or linguistic artifacts that the model may rely upon. Additionally, we utilize dataset ablation techniques, including scrambling word order, shuffling labels, and employing partial input strategies, to assess the accuracy of these cues and their impact on the model's performance. The results demonstrate a noticeable decline in performance with each ablation, indicating a strong dependency of the model on these cues.

#### 7.4.5 Future Works

This section highlights the discrepancies in existing fake news datasets and detection methods, emphasizing the need for further investigation to establish robust and generalizable fake news detection models. However, limited research has been conducted in the area of multimodal fake news detection. Therefore, there are several potential future directions that can be explored to address this gap in the field.

#### Quality Assessment of Multimodal Fake News Datasets

The examination of existing multimodal fake news datasets should focus on several key measures to ensure their quality and reliability. These measures include:

- **Duplicate Entries:** It is essential to inspect the datasets for duplicate entries, as the presence of duplicates can skew the training process and lead to biased results. Identifying and removing duplicate entries will help improve the dataset's integrity and prevent overrepresentation of certain instances.
- **Noisiness of the Dataset:** Noise in the dataset can negatively impact the performance of fake news detection models. It is important to assess the level of noise present in the dataset, such as misleading or incorrectly labeled instances, to determine its impact on the model's effectiveness. Cleaning or reducing noise in the dataset can enhance the reliability of the detection methods.
- **Relevance to Specific Events or Topics:** Multimodal fake news datasets should be evaluated to determine whether they cover a diverse range of events or topics. Ensuring the inclusion of samples that represent a wide spectrum of events and topics is crucial for training models that can generalize well across different contexts.
- **Dataset Balance:** The balance of the dataset, i.e., the distribution of fake news and genuine news samples, is an important factor to consider. Imbalanced datasets can lead to biased model performance, where the model may prioritize the majority class

and overlook minority classes. Therefore, it is important to examine the dataset's balance and consider techniques such as oversampling or undersampling to address any imbalances.

- **Dataset Shift:** Dataset shift refers to the discrepancy between the training and testing data distributions. It is important to identify if dataset shift exists in multimodal fake news datasets, as this can impact the generalizability of the trained models. Techniques such as domain adaptation or transfer learning may be explored to mitigate the effects of dataset shift.
- **Analyzing Visual Cues:** The role of images within the multimodal dataset needs to be investigated, including understanding what aspects the images highlight and whether they accurately represent the corresponding news articles. In the context of social media content, where multiple news articles may share the same images, the role of images becomes even more significant. Examining the relationship between images and the associated news content can provide insights into the impact of visual cues on fake news detection.

By thoroughly inspecting existing multimodal fake news datasets based on these measures, researchers can ensure the quality, relevance, and generalizability of the data, thereby improving the effectiveness of fake news detection models.

### **Opening the black-box: Understanding Model Biases**

In order to assess the trustworthiness of an algorithm, it is crucial to understand the factors that trigger its decisions. By identifying and visualizing the patterns that detectors consider during decision-making, we can gain insights into their decision-making processes. This can be achieved through the following approaches:

- **Examining Model Layers and Attention:** By analyzing the model layers, we can determine which parts of the text or image the algorithm pays the most attention to. This provides valuable insights into the features or cues that are considered important for making decisions. Visualizing the attention mechanisms of the model can help us understand the specific patterns or elements that influence its predictions.
- **Evaluating Extracted Patterns:** It is important to assess the quality and genuineness of the patterns extracted by the algorithm. This involves examining whether the identified patterns are reliable indicators of the target class or if they are merely noise. By performing statistical bias tests, such as the Productivity metric, we can evaluate the patterns for any biases or inaccuracies. This step ensures that the algorithm is not relying on spurious patterns or artifacts in the data.

By conducting these analyses, we can gain a deeper understanding of the decision-making process of the algorithm and determine the reliability and trustworthiness of its predictions. It

allows us to assess whether the algorithm is truly capturing meaningful and relevant patterns or if its decisions are influenced by biases or noise in the data. Ultimately, this investigation helps to enhance the transparency and accountability of the algorithm and contributes to the development of more robust and reliable detection methods.

### **Validating Reported Results**

The detection of fake news has emerged as a critical research area, necessitating the development of robust and reliable detection methods. However, the evaluation of these methods poses several challenges, including the need for effective validation measures. One potential direction in the future could be to perform a comprehensive assessment of potential measures to validate fake news detection methods during evaluation.

- **Basic N-folds:** This measure involves splitting the dataset into multiple folds and performing cross-validation. The dataset is divided into training and testing sets, and the model is trained and evaluated on each fold separately. This helps assess the generalizability of the detection method across different subsets of the data and reduces the impact of any specific biases or variations in the dataset.
- **Time Invariance Cross Validation:** Time invariance cross-validation aims to evaluate the detection method's performance across different time periods. This involves dividing the dataset into different time intervals and training and testing the model on different time segments. By examining the consistency of the method's performance over time, we can assess its ability to maintain accuracy and effectiveness as the nature of fake news evolves.
- **Predicting Never Before Encountered Domains:** This measure focuses on evaluating the detection method's performance on news articles from domains that were not present in the training data. By training the classifiers on a randomly sampled subset of domains and testing them on articles from previously unseen domains, we can assess the method's ability to generalize to new and unfamiliar sources.
- **Forecasting into the Future:** This approach involves training classifiers on fake and mainstream news articles written before a specific time point and testing them on articles written after that time. By simulating real-world scenarios where the detection method needs to classify news articles in real-time, this measure assesses the method's ability to adapt to new and evolving types of fake news.

Through these measures, researchers and practitioners can gain insights into the reliability, generalizability, and adaptability of detection methods in identifying fake news.

## Chapter 8

# Intervention Strategies: Enhancing Data Accessibility via System Design

This chapter is partly a reproduction of work showcased at ACM-W India Celebration of Women in Computing (AICWiC), 2020 [241] and the Communications of the ACM (CACM), 2022 [238].

### 8.1 Overview

In the preceding chapters, our focus was primarily on two core aspects of the thesis: designing multimodal fake news detection baselines and inspecting fake news to gain insights into its evolving dynamics. This chapter deals with the third core aspect of the thesis, i.e. Intervening in Fake News. We aim to design intervention method that enable readers to identify fake news. After going through numerous solutions towards multimodal fake news detection, we understand it not to be a trivial task. Further, after inspecting fake news scenarios in India, we realized that much of the fake news circulating online had been resurfaced and debunked by fact-checking websites before. Building upon this insight, we propose a solution in this chapter that leverages the efforts of fact-checking organizations to combat the proliferation of fake news online and make debunked information readily accessible to the public. By harnessing the work done by fact-checking organizations, we aim to provide readers with the means to effectively identify and navigate through the vast landscape of fake news. Through this intervention approach, we aspire to contribute to the mitigation of fake news and promote a more informed and discerning online environment.

### 8.2 SachBoloPls: A Realtime Identification of Fake News Online

We design SachBoloPls- a system that validates news on Twitter in real-time. SachBoloPls (‘Speak the Truth Please’) is an effort to curb the proliferation of debunked fake news online. It is an effort to make audiences aware of such fact-checking organizations and educate them

about false viral claims. When a user invokes SachBoloPIs on Twitter, the system collects the original news tweet of the thread and matches it with the database. Results are then formed into a tweet and tweeted back to the original line. There are three components of the design that are independent of each other. Thus, the proposed prototype can be extended to social media and instant messaging platforms like Instagram, WhatsApp, Facebook, and Telegram. SachBoloPIs can also incorporate regional languages making it a viable tool to fight against fake news across India.

### 8.3 Prototype Design

Following are the working steps of SachBoloPIs:

#### Module 1: Continuous Data Collection

In the sub-module, we have implemented a data crawling process that retrieves data from Indian fact-checking websites at regular intervals of twelve hours. To accomplish this, we have developed Python scripts that utilize the BeautifulSoup library. These scripts are responsible for extracting relevant data from the websites and subsequently saving each curated sample as a JSON file in our MongoDB instance. The data crawling process involves accessing the fact-checking websites, parsing the HTML content, and extracting the desired information, such as article titles, URLs, claims, fact-checking verdicts, and supporting evidence. This curated data is then organized and stored in the form of JSON files, ensuring its structured representation and ease of access. By implementing this data crawling sub-module, we are able to maintain an updated and comprehensive database of information from Indian fact-checking websites. This database serves as a valuable resource for our intervention methods and enables the accessibility of debunked information to aid readers in identifying and combating fake news effectively.

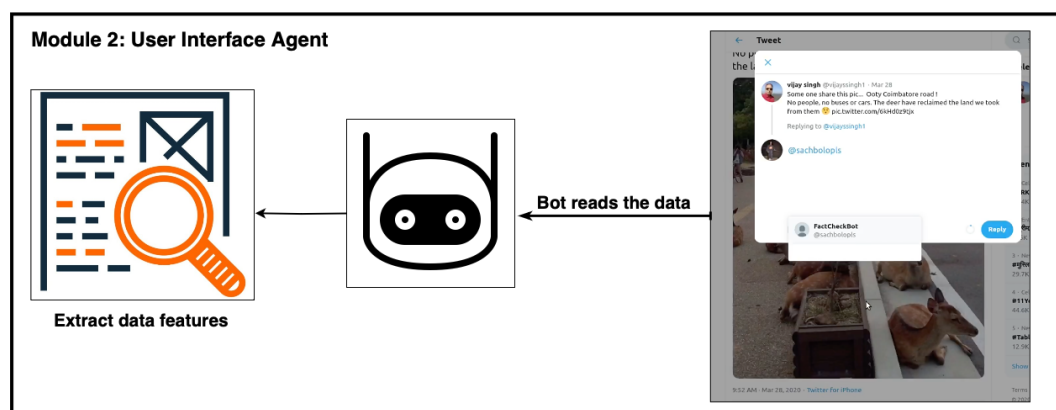


Figure 8.1: The second sub-module of SachBoloPIs, i.e. User Interface Agent, which gets activated when a user calls it. The module is responsible for reading data and extracting features of the query on which SachBoloPIs is invoked.

## Module 2: User Interface Agent

The SachBoloPls system utilizes the Twitter API to track platform mentions. By monitoring tweets that mention *@sachbолоpls*, users can engage with the bot service provided by SachBoloPls to evaluate the veracity of questionable content. Upon invoking the bot, we collect and curate the tweet ID of the tagged tweet and the ID of its source tweet. Leveraging the Twitter API, we extract information associated with the source tweet ID, i.e. meta-features, content-related data, and network features. Figure 8.1 provides a high-level representation of how the sub-module functions.

## Module 3: Multimodal Search and Ranking System

We call the multimodal search and ranking module that matches the curated tweet information with the samples in the database. We perform word-watching of the tweet content with the title of the news sample and its associated tags. Further, we apply the Facebook PDQ algorithm to find similar images in the database. The combination of text and image match aids in finding the news sample that closely resonates with the queried sample. In case of failure, the module returns ‘no match’ found. On finding the match, our Python script pings the source tweet id with a reply constituting the link of the article that shows that the queried news is fake. Figure 8.2 provides a high-level explanation of how the sub-module functions.

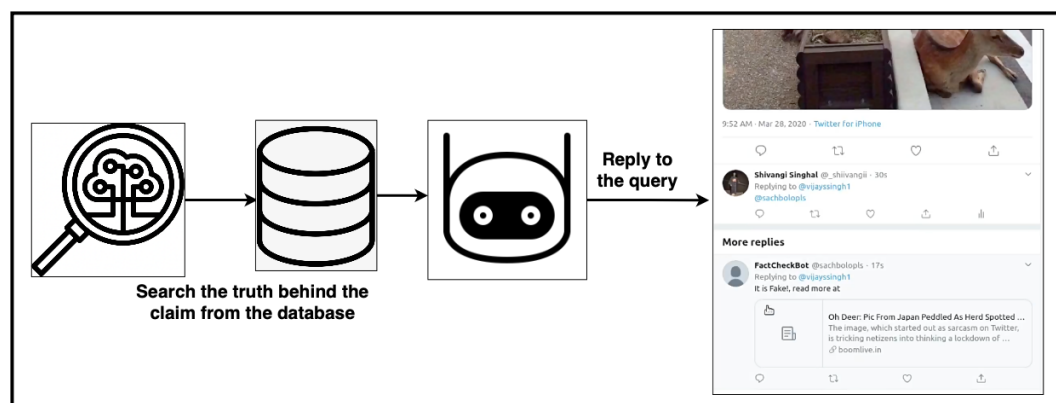


Figure 8.2: The third sub-module of SachBoloPls, i.e. Multimodal Search and Ranking System, which performs pattern matching over the queried data and the database to find the relevant news article, if it has been debunked before by the any of the agencies. In failure, the system returns a ‘not found’ message.

## 8.4 Implications

The work has following implications:

- Fact-checking Organizations: Though there has been a significant effort by fact-checkers across the globe to debunk the false viral claims, the efforts seem to go in



vain. For instance, while curating FactDrill [242], we realized that a large chunk of fake news is a re-surfaced version. What is more concerning is that even though multiple fact-checking organizations have refuted these news stories, the public continues to find them persuasive and falls victim to their influence. To address the issue, we designed SachBoloPls, an initiative that utilizes fact-checking efforts to make the masses aware of the already debunked fake news. Through this initiative, we strive to bridge the gap between fact-checking efforts and public awareness, ultimately mitigating the impact of fake news and fostering a more informed society.

- **General Audience:** SachBoloPls serves as a comprehensive platform where users can access information about all debunked viral claims. The prototype is designed to support thirteen languages, making it accessible to a diverse audience. Moreover, the application is user-friendly and requires no technical expertise, ensuring ease of use for individuals of all generations.
- **Researchers:** SachBoloPls aggregates data from fact-checking websites every twelve hours, providing the most extensive collection of debunked, fact-checked viral claims and information about their various iterations found online. This wealth of data offers researchers an opportunity to study fake news and its evolving dynamics, gaining insights into the factors driving such changes.
- **Government Organizations:** The SachBoloPls module consists of three independent components, allowing for easy expansion to other social media and instant messaging platforms such as Instagram, WhatsApp, Facebook, and Telegram. By doing so, we can effectively combat the proliferation of debunked fake news on a broader scale. Additionally, SachBoloPls can incorporate regional languages, making it a valuable tool in the fight against fake news throughout India. Furthermore, this system can serve as a model for international implementation, enabling the monitoring of online debunked news worldwide.

## 8.5 Limitations

Due to the discontinuation of free access to the Twitter API on February 09, 2023, our proposed SachBoloPls is currently non-functional. However, as stated earlier, each component in SachBoloPls is an independent module, allowing for flexibility in shifting our focus to WhatsApp as an alternative platform. Additionally, maintaining continuous data collection poses significant challenges. Websites frequently modify their web layouts or implement firewalls, disrupting our data collection scripts. Keeping track of eleven different websites across fourteen languages becomes a laborious task. Furthermore, the IFCN network regularly adds new websites, necessitating the regular update and inclusion of these additions into our database. These ongoing tasks demand considerable effort and attention to ensure the effectiveness and accuracy of the SachBoloPls system.

## **8.6 Conclusion**

The thesis introduces SachBoloPls ('Speak the Truth Please'), an initiative to mitigate the widespread dissemination of fake news online and ensure accessible access to debunked information for the public. Time and again, we have witnessed fake stories resurfacing the online media and the masses falling prey to them, a clear sign of negligence from the Indian audiences toward fact-checking efforts. It is crucial to inform and educate audiences about the existence of fact-checking organizations and raise awareness about the prevalence of false viral claims. SachBoloPls work towards fostering a more informed and discerning public, better equipped to distinguish between real and fake news.

## Chapter 9

# Conclusion, Limitations and Future Works

The thesis studied the domain of multimodal fake news from the viewpoint of multiple modalities. We address three fundamental challenges. First, our research focused on devising different methods to *Identify*, a.k.a., detect fake news online by extracting different feature sets from the given information. By designing foundational detection mechanisms, our work accelerates research innovations. Second, our research closely *Inspect* the fake stories from two perspectives. First, from the information point of view, we present a collection of fake news that targets India. With the help of such a data repository, one can inspect fabricated content to (i) identify the different patterns of false stories disseminating over the web, (ii) the modality used to create the fabricated content and (iii) the platform used for dissemination. Next, from the model point of view, we inspect detection mechanisms used in prior work and their generalizability to other datasets. Lastly, we focus on designing *Intervention* method to

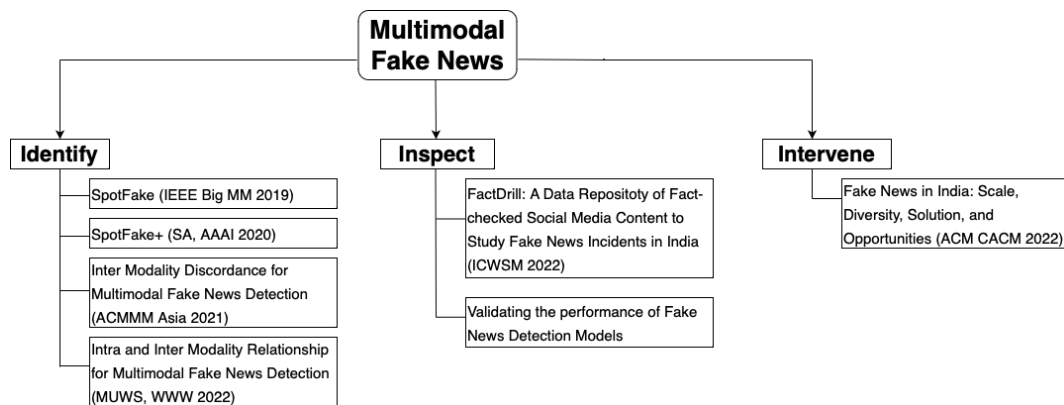


Figure 9.1: A summary of the research done for the PhD. The thesis studies the domain of multimodal fake news from the viewpoint of multiple modalities. We address three fundamental challenges, i.e. Identify, Inspect and Intervene.

expand the intuitive understanding of fake news among online readers. We propose practical

implications for social media platform owners and policymakers. Figure 9.1 captures a high-level schema of the problems addressed in this thesis. We believe our suggestions and findings will help scholars, organizations, and the public better comprehend the present fake news situation. This chapter discusses the implications of our work on multimodal fake news, its shortcomings and suggestions for future study.

## 9.1 Summary

The main contribution of the thesis is summarized as follows:

### 9.1.1 Designing Multimodal Fake News Detection Baselines

In 2017, Jin et al. [116] made the first attempt toward multimodal fake news detection. The paper proposed a content-based multimodal fake news detection method that uses a recurrent neural network with an attention mechanism to combine text and social context features. It uses VGG-19 pre-trained on the Imagenet database to generate representations for images present in tweets. Several other works performed a similar task, discussed in Chapter 3.1.3. However, there are a few inconsistencies with the existing literature:

1. None of the approaches extracts contextual information from the text. Each method captures the syntactic and semantic features of the text.
2. There are multimodal fake news detection systems in the literature, but they solve the fake news problem by considering an additional sub-task like an event discriminator [276] and finding correlations across the modalities [126]. The results of fake news detection are heavily dependent on the subtask, and in the absence of subtask training, the performance of fake news detection degrades by 10% on average.

Proposed Solution:

1. We introduce SpotFake- a multimodal framework for fake news detection [235]. Our proposed solution detects fake news without taking into account any other subtasks. It exploits both the textual and visual features of a news sample. Specifically, we used language models (like BERT) to learn contextual representations for the text and image features learned from VGG-19 pre-trained on the ImageNet dataset.
2. We introduce SpotFake+ [237], a multimodal approach that leverages transfer learning to capture semantic and contextual information from the news articles and its associated images and achieve better accuracy for fake news detection.

### 9.1.2 Identifying the role of Multiple Images

Numerous studies have been performed on multimodal fake news detection, but limited attention is drawn to addressing the role of multiple images. Moreover, none of the works

establishes the relationship between multiple components of a news article. Upon examining the related literature, we find the strongest baselines for single-image and multi-image content-based multimodal fake news detection to be SAFE [307] and Giachanou et al. [82], respectively. Both methods demonstrate the following drawbacks:

1. In the research presented by Zhou et al. [307], (i) the textual features are extracted via a Text-CNN [130] ignoring the contextual information, (ii) the image is converted into text via image2sent [266] model. Next, cosine similarity is calculated to explore the relationship between the two modalities. We believe converting an image into text might result in a loss of semantic information within an image, and (iii) no comparison is shown with the existing state-of-the-art methods to demonstrate the effectiveness of the proposed model.
2. On the other side, work performed by Giachanou et al. [82] lacks the reasoning for utilizing multiple images for multimodal fake news classification. Second, taking cues only from the headlines and ignoring the content might lead to information loss. Third, while capturing the similarity, the top ten image tags are preferred over the image features. This might lead to inconsistent results as (i) extracted tags might fail to capture the semantic relationship across the images, (ii) incorporating only the top ten tags might not capture the information present in the image effectively, and (iii) extracted tags might be limited by the vocabulary of the pre-trained model used for extraction and can introduce external bias in the final representations.

Proposed Solution:

- We present an inter-modality discordance-based multimodal fake news detection method [236]. It captures intra-modality relationships by extracting sequential information from the text and multiple images. In addition, it also forms a multimodal representation of the news article to explore the hidden, latent patterns. Our work also introduces a novel application of contrastive loss for measuring the discordance between the components. Enforcing all such losses in conjunction enables better feature extraction and robust learning to achieve state-of-the-art performance on multi-image multimodal fake news detection.

### **9.1.3 Extracting Intra and Inter Modality Relationship**

We find the following drawbacks upon examining the literature in Chapter 3.1.3.

1. Each method discussed before extracts visual information via Text-CNN or VGG-19. A complete image is passed through the network to generate the representations. Image contains unwanted (redundant) information in the form of background that can be excluded.

2. Existing methods for multimodal fake news detection do not work on the principles of weak and strong modality [54, 126, 235, 237, 276, 307]. Instead, methods capture high-level information from different modalities and jointly model them to determine the authenticity of the news. The feature extraction also occurs globally, ignoring the salient pixels containing meaningful information. However, reports<sup>1</sup> show the existence of different versions of fake news due to manipulations performed in the different modalities.

Proposed Solution:

- We present a novel framework that leverages intra and inter-modality relationships for multimodal fake news detection [240]. The method comprises two modules, (i) Capturing inter-modality relationship, where we present a novel architecture that uses a multiplicative multimodal method to capture the inter-modality relationship between modalities. Using the multiplicative multimodal method, we aim to leverage information from a more reliable modality than a less reliable one on a per-sample basis, (ii) Capturing intra-modality relationship, where we intend to extract the fine-grained salient representations for image and text. The resultant feature vectors capture rich contextual dependencies present within its components.

#### 9.1.4 Resource Creation for Indic languages

India witnessed a 214% rise in cases relating to fake news in 2019.<sup>2</sup> For instance, during the 2016 Indian banknote demonetization, multiple fake news reports about spying technology added to the banknotes went viral [43]. Around 5000 social media handles were suspended by Indian security and intelligence agencies during the CAA protests.<sup>3</sup> The dissemination of fake content via WhatsApp was prevalent during the 2019 Indian general election [196]. The proliferation of fake news in India is massive, and there is a dire need to consider solutions explicitly catering to the Indian region. There has been little effort made to study the menace of fake news in India, still it faces a few limitations.

1. The IFND dataset [222] is highly imbalanced. No assurance about the authenticity of sources is provided.
2. In the FakeNewsIndia dataset [64], the sample count is low. Data curation is also performed for a short period.
3. Both the curated datasets [64, 222] consists of English samples, missing the data in regional languages.

---

<sup>1</sup><https://www.pagecentertraining.psu.edu/public-relations-ethics/introduction-to-the-ethical-implications-of-fake-news-for-pr-professionals/lesson-2-fake-news-content/types-of-fake-news/>

<sup>2</sup><https://bit.ly/3u8sYTc>

<sup>3</sup>Around 5,000 Pak social media handles spread fake news on CAA". Outlook India. IANS. 16 December 2019. Retrieved 21 December 2019

4. There are numerous attributes present in a website, but both the papers [64, 222] limits to some specific features. This might lead to information loss.

Proposed Solution:

- We present FactDrill [242], a dataset containing 22,435 fact-checked social media content, to study fake news incidents in India. The dataset comprises news stories from 2013 to 2020, covering 13 languages in the country. There are 14 different attributes present in the dataset.

### 9.1.5 Data Accessibility: A System Design

In India, the surge in Internet penetration accompanied by digital illiteracy has resulted in the rise of fake news online [185]. Internet penetration in India has risen from 137 million Internet users in 2012 to over 600 million in 2019.<sup>4</sup> Though audiences have access to information, their inability to comprehend the nuances and implications of fake news has resulted in the growth of fake news dissemination in the online world. One possible pathway would be to design intervention policies to measure the respondents' ability to identify fake news. The other way to tackle the situation would be to introduce mediums for the accessibility and utilization of fact-checkers efforts. To this, we design SachBoloPls [241], a system that validates news on Twitter in real-time. It is an effort to make audiences aware of fact-checking organizations and educate them about false viral claims. There are three components of the design that are independent of each other. Thus, extending the working module to other social media platforms like Instagram, WhatsApp, Facebook, and Telegram gives us leverage. SachBoloPls can also incorporate regional languages making it a viable tool to fight against fake news across India.

## 9.2 Limitations

Research on fake news has made significant strides in understanding its impact and dynamics. However, there are several limitations and challenges associated with studying fake news that researchers need to consider:

- Definition and conceptualization: Fake news lacks a universally agreed-upon definition. Despite our attempts to elucidate the concept in Chapter 2, researchers continue to hold diverse interpretations and understandings of what encompasses fake news. This variance in perspectives has resulted in inconsistent research findings and challenges when comparing across different studies.
- Sample representativeness: In Chapter 4, we administered a survey to understand users perspective about fake news. However, it is important to acknowledge that the study

---

<sup>4</sup>Mohan, Shriya (26 April 2019). 'Everybody needs a good lie'. Business Line. Retrieved 28 August 2019.

faced limitations, primarily related to the representativeness of the sample. Gathering a representative sample of fake news stories or users who engage with fake news is challenging. The internet is vast, and fake news can spread rapidly across platforms, making it challenging to capture the entire landscape accurately. Limited sample size or biased sampling can affect the generalizability of research findings.

- Fake news is evolutionary: Several multimodal datasets, such as the MediaEval Twitter Benchmark dataset [23], the Weibo dataset [116], and the FakeNewsNet repository [227], have captured fake news incidents circulating online. However, it is crucial to note that these datasets have temporal limitations. For example, the MediaEval Twitter Benchmark dataset was collected in 2015 and 2016, the Weibo dataset was captured in 2017, and the samples in the FakeNewsNet repository were collected in 2018. These datasets predominantly focus on specific events or capture data during a specific period, rendering them prone to obsolescence over time. It is essential to recognize that the performance of detection algorithms trained on such datasets may not accurately reflect their effectiveness in real-time identification of false news. Models trained on datasets capturing the 2016 US elections, for instance, might not perform well in identifying fake news related to the 2019 COVID Pandemic. As new forms of fake news continue to emerge, researchers face the challenge of keeping up with the evolving landscape. This constant evolution necessitates ongoing research efforts to develop and refine detection algorithms capable of effectively identifying and combating fake news in real-time scenarios.
- Lack of interdisciplinary collaboration: Studying fake news requires interdisciplinary collaboration across fields such as psychology, communication, sociology, computer science, and data science. The lack of collaboration between these disciplines can limit the holistic understanding of fake news and hinder the development of effective countermeasures.
- Platform limitations: The availability of comprehensive and real-time data from social media platforms and search engines has been constrained by privacy concerns and the use of proprietary algorithms. In particular, the removal of free access to the Twitter API in 2023 has posed challenges for researchers in understanding the dynamics and spread of fake news on these platforms. Researchers heavily relied on social media data to analyze patterns, trends, and the dissemination of fake news. The loss of free access to the Twitter API restricts researchers' ability to gather large-scale, real-time data, hindering their capacity to study fake news comprehensively.

### 9.3 Future Works

As technology continues to advance and new forms of fake news emerge, researchers are compelled to explore innovative approaches and promising directions to deepen their



understanding and develop effective countermeasures. Several key trends and avenues of research are expected to shape the future landscape of fake news research.

- **Algorithmic Transparency:** The opaque algorithms used by social media platforms and search engines play a significant role in the spread and visibility of fake news. Future research will focus on advocating for algorithmic transparency and understanding the algorithms' influence on information dissemination. By shedding light on algorithmic biases and their impact on the filter bubble effect, researchers can propose guidelines and policies to mitigate the unintended consequences of algorithmic decision-making.
- **Technological Countermeasures:** As technology advances, so do the techniques used to create and spread fake news. Future research can focus on developing advanced technological tools and solutions, such as AI-powered fact-checking and verification algorithms or user-friendly browser extensions, on effectively detect and combat fake news. For instance, after inspecting fake news scenarios in India, we realized that much of the fake news circulating online is a resurfaced version and has been debunked before by fact-checking websites. SachBoloPls [241] is an effort to curb the proliferation of fake news online and make debunked information accessible to people. Expanding the design to more online platforms is a potential future development. Further, we can upgrade the data collection module to incorporate other fact-checking websites.
- **Vulnerable Populations:** Certain demographics, such as older adults or individuals with lower digital literacy, may be more vulnerable to fake news. Investigating the specific vulnerabilities and information needs of these populations can guide the development of targeted interventions and educational programs to empower and protect them from fake news.
- **Psychological and Cognitive Factors:** Fake news exploits various psychological and cognitive biases, such as confirmation bias and the illusion of truth effect, which make individuals more susceptible to fake news. Future research can delve deeper into these factors to develop effective cognitive inoculation strategies and interventions that help individuals recognize and resist fake news.
- **Long-term Impact Assessment:** Evaluating the long-term effects of fake news on individuals, public trust, and democratic processes is an important yet challenging task. Future research will emphasize longitudinal studies to understand the persistence of fake news beliefs and their consequences over time. By examining the societal impact of fake news, researchers can provide insights into the long-term strategies needed to build resilience against fake news.

# Appendix A

## Survey Questionnaire

Reference Chapter: Designing Simple Baselines for Multimodal Fake News Detection (Chapter 4, Section 4.3)

Real vs Fake News

16/06/23, 10:29

### Real vs Fake News

The purpose of the survey is to know whether humans are capable enough to distinguish between Real and Fake News

\* Indicates required question

1. Q1. What is your age? \*

Mark only one oval.

- ☐ Below 21 years old  
☐ 21-29 years old  
☐ 30-39 years old  
☐ Above 41 years old

2. Q2. Are you: ? \*

Mark only one oval.

- ☐ Male  
☐ Female  
☐ Other: \_\_\_\_\_

3. Q3. How often do you discuss what's happening in the news with your family or friends on any platform (e.g., in person, on the phone, on social media, etc.)?

*Mark only one oval.*

- ☐ Daily
- ☐ Few times a week
- ☐ Once in a week
- ☐ Less often
- ☐ Never

4. Q4. What is your primary news source? \*

*Mark only one oval.*

- ☐ Traditional news sources (Eg: TV, radio, newspaper, News Apps on smart phones)
- ☐ Social media (eg., facebook, twitter, snapchat, instagram, etc.)
- ☐ Messaging platform (such as whatsapp, Messenger)
- ☐ Offline word-of-mouth sources

5. Q5. Which of the following medium, do you think is more trust-able for consuming news? \*

*Mark only one oval.*

- ☐ Tradition media like Newspapers, TV, News App on smart phones
- ☐ Online Social Media like Facebook and Twitter
- ☐ Messaging Platform: Whatsapp
- ☐ Offline word-of-mouth sources

6. Q6. Do you know what is termed as fake news? \*

*Mark only one oval.*

- ☐ Yes  
☐ No  
☐ Maybe

7. Q7. Do you know when are you reading fake news? \*

*Mark only one oval.*

- ☐ Yes  
☐ No  
☐ Maybe

8. Q8. Does it matter to you whether or not the news you are reading is considered real news or fake news?

*Mark only one oval.*

- ☐ Yes, it matters to you  
☐ No, it does not matter to you

9. Q9. Why do you think will be the reason for spreading fake news online? \*

*Mark only one oval.*

- ☐ For entertainment  
☐ To spread propaganda or an event  
☐ To confuse people  
☐ To spread violence among masses

10. Q10. Why do you think will be the reason for spreading fake news through traditional outlets?

*Mark only one oval.*

- ☐ It happens by accident
- ☐ It was deliberately spread by the author
- ☐ It could be a publicity stunt
- ☐ To deviate readers from their mindset or opinion on a particular topic
- ☐ Fact checking methods are poor

11. Q11. On which platform, you encounter more fake news? \*

*Mark only one oval.*

- ☐ Traditional Outlets
- ☐ Online Social Media Outlets

12. Q12. Which of the following news type is easy to distinguish between Real and Fake News?

*Mark only one oval.*

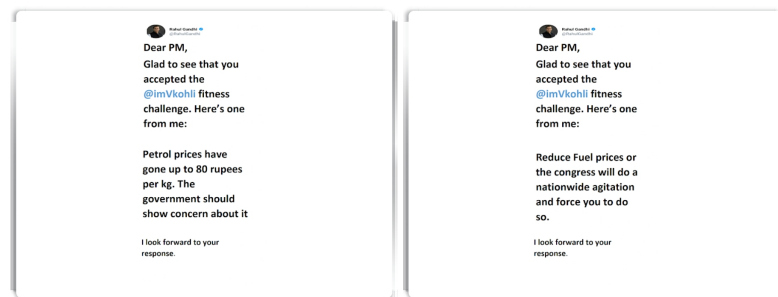
- ☐ A news article containing only Text
- ☐ A news article containing only Image
- ☐ A news article containing both Text and Image

13. Q13. From the given news article, identify which image is fake? \*

### Petrol price hike: Rahul Gandhi gives PM Modi a fuel challenge

Congress president Rahul Gandhi dared the prime minister to accept a “fuel challenge”, asking him to reduce the spiralling fuel prices or face a nationwide stir by his party.

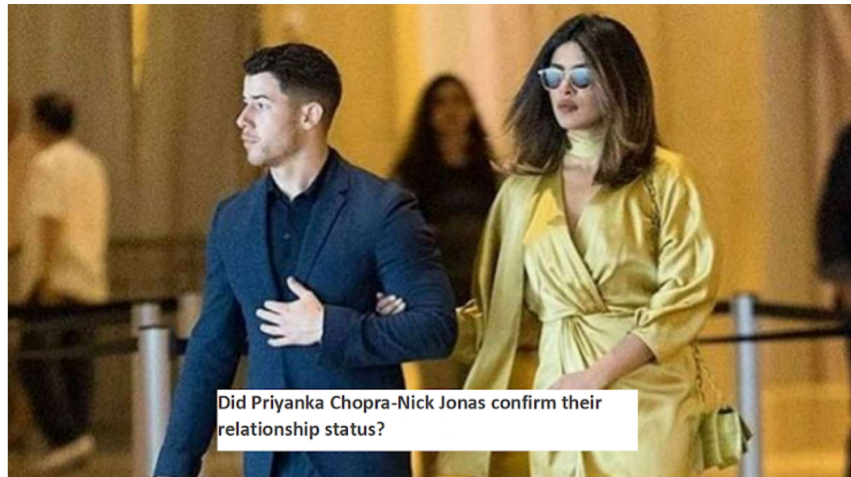
Mark only one oval.



☐ Image 1

☐ Image 2

14. Q14. For the given image, which news article, do you think could be fake? \*



Mark only one oval.

Priyanka Chopra has been much in the headlines lately for her romantic involvement with American singer Nick Jonas. Recently, Priyanka's mother, Madhu Chopra opened up about the same. In a recent media interaction, when Madhu was quizzed if Priyanka is getting married, she said, "No." In earlier interviews Madhu Chopra had mentioned that she can't imagine her daughter with a foreigner, adding that it would be easy to adapt if both sides belong to the same culture.

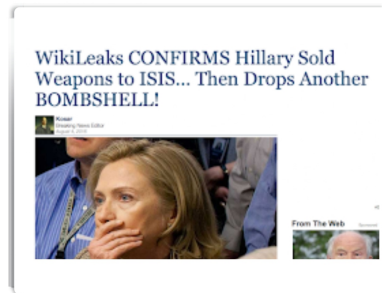
☐ Option 1

A lot has been said about Priyanka Chopra and Nick Jonas rumored relationship. The two have been spotted on many occasions together. Recently there were reports that the two got engaged on Priyanka Chopra's birthday. Reportedly Nick closed down a Tiffany store in New York City to buy an engagement ring. Now according to a report by a tabloid, Priyanka and Nick might tie the knot on his birthday on September 16, 2018. If the reports turn out to be true then this will surely be a double celebration for the couple.

☐ Option 2

15. Q15. Identify the news article related to US elections that looks fake? \*

Mark only one oval.



☐ Option 1



☐ Option 2



16. Q16. From the given viral news, identify those that you think are real? \*

Check all that apply.



☐ Option 1



☐ Option 2



☐ Option 3

17. Q17. After reading the above articles, what could a fake news comprise of? \*

Check all that apply.

- ☐ They were poorly written
- ☐ Hedline does not match the content
- ☐ Lack of validity
- ☐ Too vague
- ☐ deliberately written to spread a market campaign

18. Q18. What part of a news article you focused on while reading news? \*

*Mark only one oval.*

☐ Headline

☐ Content

☐ Image

19. Q19. What part of a news article you think was fake? \*

*Mark only one oval.*

☐ Headline

☐ Content

☐ Image

20. Q20. For a news article, will you agree multi modalities (i.e. text and images) are more useful than single modality (i.e. text/ image)?

*Mark only one oval.*

☐ Yes

☐ No

---

This content is neither created nor endorsed by Google.

Google Forms

# Bibliography

- [1] A brief history of fake news: <https://www.cits.ucsb.edu/fake-news/brief-history>.
- [2] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multi-modal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14679–14689, 2020.
- [3] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *2012 IEEE Symposium on Security and Privacy*, pages 461–475. IEEE, 2012.
- [4] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, pages 127–138. Springer, 2017.
- [5] Syeda Zainab Akbar, Divyanshu Kukreti, Somya Sagarika, and Joyojeet Pal. Temporal patterns in covid-19 related digital misinformation in india, 2020.
- [6] Syeda Zainab Akbar, Ankur Sharma, Himani Negi, Anmol Panda, and Joyojeet Pal. Anatomy of a rumour: Social media and the suicide of sushant singh rajput. *arXiv preprint arXiv:2009.11744*, 2020.
- [7] Ahmed Al-Rawi, Derrick O [U+02BC] Keefe, Oumar Kane, and Aimé-Jules Bizimana. Twitter’s fake news discourses around climate change and global warming. 2021.
- [8] Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, et al. Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms. In *ICWSM*, pages 913–922, 2021.
- [9] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*, pages 85–90, 2018.
- [10] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.

- [11] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. Scaling up fact-checking using the wisdom of crowds. *Science advances*, 7(36):eabf4393, 2021.
- [12] Barbara G Amado, Ramón Arce, and Francisca Fariña. Undeutsch hypothesis and criteria based content analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context*, 7(1):3–12, 2015.
- [13] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018.
- [14] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [15] Oberiri Destiny Apuke and Bahiyah Omar. Fake news and covid-19: modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*, 56:101475, 2020.
- [16] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*, 2017.
- [17] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Syst.*, 16(6):345–379, November 2010.
- [18] Tobias Bartholomé and Rainer Bromme. Coherence formation when learning from text and pictures: What kind of support for whom? *Journal of Educational Psychology*, 101(2):282, 2009.
- [19] Kaustav Basu and Arunabha Sen. Epidemiological model independent misinformation source identification. 2021.
- [20] Irakli Beridze and James Butcher. When seeing is no longer believing. *Nature Machine Intelligence*, 1(8):332–334, 2019.
- [21] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.
- [22] Christina Boididou et al. Verifying multimedia use at mediaeval 2015.

- [23] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.
- [24] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.
- [25] Gary D Bond, Rebecka D Holman, Jamie-Ann L Eggert, Lassiter F Speller, Olivia N Garcia, Sasha C Mejia, Kohlby W Mcinnes, Eleny C Cenicerros, and Rebecca Rustige. ‘lyin’ted’, ‘crooked hillary’, and ‘deceptive donald’: Language of lies in the 2016 us presidential debates. *Applied Cognitive Psychology*, 31(6):668–677, 2017.
- [26] Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. Brenda: Browser extension for fake news detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’20*, page 2117–2120, New York, NY, USA, 2020. Association for Computing Machinery.
- [27] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14, 2019.
- [28] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14, 2019.
- [29] Lia Bozarth and Ceren Budak. Toward a better performance evaluation framework for fake news classification. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 60–71, 2020.
- [30] Lia Bozarth and Ceren Budak. Toward a better performance evaluation framework for fake news classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):60–71, May 2020.
- [31] Lia Bozarth, Aparajita Saraf, and Ceren Budak. Higher ground? how groundtruth labeling impacts our understanding of fake news about the 2016 us presidential nominees. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 48–59, 2020.
- [32] Petter Bae Brandtzaeg, Marika Lüders, Jochen Spangenberg, Linda Rath-Wiggins, and Asbjørn Følstad. Emerging journalistic verification practices concerning social media. *Journalism Practice*, 10(3):323–342, 2016.
- [33] Nadia M Brashier, Gordon Pennycook, Adam J Berinsky, and David G Rand. Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, 118(5):e2020043118, 2021.

- [34] Adam Breuer, Roe Eilat, and Udi Weinsberg. Friend or faux: graph-based early detection of fake accounts on social networks. In *Proceedings of The Web Conference 2020*, pages 1287–1297, 2020.
- [35] David A Broniatowski, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American journal of public health*, 108(10):1378–1384, 2018.
- [36] Clint Burfoot and Timothy Baldwin. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164, Suntec, Singapore, August 2009. Association for Computational Linguistics.
- [37] Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media*, pages 141–161, 2020.
- [38] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011.
- [39] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW '11*, pages 675–684, New York, NY, USA, 2011. ACM.
- [40] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. Misleading online content: Recognizing clickbait as "false news". In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, WMDD '15*, pages 15–19, New York, NY, USA, 2015. ACM.
- [41] Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. Causal understanding of fake news dissemination on social media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 148–157, 2021.
- [42] Mingxi Cheng, Songli Wang, Xiaofeng Yan, Tianqi Yang, Wenshuo Wang, Zehao Huang, Xiongye Xiao, Shahin Nazarian, and Paul Bogdan. A covid-19 rumor dataset. *Frontiers in Psychology*, 12:644801, 2021.
- [43] Innocent E Chiluwa and Sergei A Samoilenko. *Handbook of research on deception, fake news, and misinformation online*. Information Science Reference/IGI Global, 2019.
- [44] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

- [45] Rajdipa Chowdhury, Sriram Srinivasan, and Lise Getoor. Joint estimation of user and publisher credibility for fake news detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1993–1996, 2020.
- [46] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other- audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 439–447, New York, NY, USA, 2020. Association for Computing Machinery.
- [47] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. pages 251–263, 03 2017.
- [48] Keith Coleman. Introducing birdwatch, a community-based approach to misinformation, [https://blog.twitter.com/en\\_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation](https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation), 2021.
- [49] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4, 2015.
- [50] John Cook, Peter Ellerton, and David Kinkead. Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters*, 13(2):024018, 2018.
- [51] Jan Christian Blaise Cruz, Julianne Agatha Tan, and Charibeth Cheng. Localization of fake news detection via multitask transfer learning. *arXiv preprint arXiv:1910.09295*, 2019.
- [52] Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*, 2020.
- [53] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 492–502, 2020.
- [54] Limeng Cui, Suhang Wang, and Dongwon Lee. Same: Sentiment-aware multi-modal embedding for detecting fake news. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, page 41–48, New York, NY, USA, 2019. Association for Computing Machinery.
- [55] Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeno, and Preslav Nakov. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293, 2020.

- [56] Enyan Dai, Yiwei Sun, and Suhang Wang. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 853–862, 2020.
- [57] Anupam Das and Ralph Schroeder. Online disinformation in the run-up to the indian 2019 election. *Information, Communication & Society*, 24(12):1762–1778, 2021.
- [58] Sohan De Sarkar, Fan Yang, and Arjun Mukherjee. Attending sentences to detect satirical fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3371–3380. Association for Computational Linguistics, 2018.
- [59] Marco Del Tredici and Raquel Fernández. Words are the window to the soul: Language-based user representations for fake news detection. 2020.
- [60] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.
- [61] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada, August 2017. Association for Computational Linguistics.
- [62] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [63] Arkin Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. Drink bleach or do what now? covid-hera: A dataset for risk-informed health decision making in the presence of covid19 misinformation. *arXiv preprint arXiv:2010.08743*, 2020.
- [64] Apoorva Dhawan, Malvika Bhalla, Arora Deeksha, Kaushal Rishabh, and Ponnu-rangam Kumaraguru. Fakenewsindia: A benchmark dataset of fake news incidents in india, collection methodology and impact assessment in social media. *Journal of Computer Communications*, 2022.
- [65] Ding Ding, Edward W Maibach, Xiaoquan Zhao, Connie Roser-Renouf, and Anthony Leiserowitz. Support for climate policy and societal action are linked to perceptions about scientific agreement. *Nature Climate Change*, 1(9):462–466, 2011.



- [66] Yingdong Dou, Kai Shu, Congying Xia, Philip S. Yu, and Lichao Sun. User preference-aware fake news detection. *SIGIR '21*, page 2051–2055, New York, NY, USA, 2021. Association for Computing Machinery.
- [67] Francesco Ducci, Mathias Kraus, and Stefan Feuerriegel. Cascade-1stm: A tree-structured neural classifier for detecting misinformation cascades. In *proceedings of the 26th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pages 2666–2676, 2020.
- [68] Alexander Eitel and Katharina Scheiter. Picture or text first? explaining sequence effects when learning with pictures and text. *Educational psychology review*, 27(1):153–180, 2015.
- [69] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. Covid-19-fakes: A twitter (arabic/english) dataset for detecting misleading information on covid-19. In *International Conference on Intelligent Networking and Collaborative Systems*, pages 256–268. Springer, 2020.
- [70] Ziv Epstein, Adam J Berinsky, Rocky Cole, Andrew Gully, Gordon Pennycook, and David G Rand. Developing an accuracy-prompt toolkit to reduce covid-19 misinformation online. 2021.
- [71] Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. Do explanations increase the effectiveness of ai-crowd generated fake news warnings? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 183–193, 2022.
- [72] Ziv Epstein, Gordon Pennycook, and David Rand. Will the crowd game the algorithm? using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–11, 2020.
- [73] Ziv Epstein, Nathaniel Sirlin, Antonio Arechar, Gordon Pennycook, and David Rand. The social media context interferes with truth discernment. *Science Advances*, 9(9):eabo6169, 2023.
- [74] Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 171–175, 2012.
- [75] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Image forgery localization via fine-grained analysis of cfa artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5):1566–1577, 2012.

- [76] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. ACL, 2016.
- [77] Christie M Fuller, David P Biros, and Rick L Wilson. Decision support for determining veracity via linguistic-based cues. *Decision Support Systems*, 46(3):695–703, 2009.
- [78] Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup 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)*, pages 1683–1698, 2021.
- [79] K. Gadzicki, R. Khamsehashari, and C. Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–6, 2020.
- [80] Elizabeth A Gage-Bouchard, Susan LaValley, Molli Warunek, Lynda Kwon Beupin, and Michelle Mollica. Is cancer information exchanged on social media scientifically accurate? *Journal of Cancer Education*, 33(6):1328–1332, 2018.
- [81] Anna Gausen, Wayne Luk, and Ce Guo. Can we stop fake news? using agent-based modelling to evaluate countermeasures for misinformation on social media. In *International AAAI Conference on Web and Social Media (ICWSM)*. <https://doi.org/10.36190>, 2021.
- [82] Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. Multimodal multi-image fake news detection. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 647–654. IEEE, 2020.
- [83] Miroslav Goljan, Jessica Fridrich, and Mo Chen. Defending against fingerprint-copy attack in sensor-based camera identification. *IEEE Transactions on Information Forensics and Security*, 6(1):227–236, 2010.
- [84] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [85] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. Semeval-2019 task 7: Rumoureal, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, 2019.

- [86] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029, 2009.
- [87] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019.
- [88] Gisel Bastidas Guacho, Sara Abdali, Neil Shah, and Evangelos E Papalexakis. Semi-supervised content-based detection of misinformation via tensor embeddings. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, pages 322–325. IEEE, 2018.
- [89] Brian Guay, Gordon Pennycook, David Rand, et al. Examining partisan asymmetries in fake news sharing and the efficacy of accuracy prompt interventions. 2022.
- [90] Maike Guderlei and Matthias Aßenmacher. Evaluating unsupervised representation learning for detecting stances of fake news. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6339–6349, 2020.
- [91] Hatice Gunes and Massimo Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE international conference on systems, man and cybernetics*, volume 4, pages 3437–3443. IEEE, 2005.
- [92] Hatice Gunes and Massimo Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE international conference on systems, man and cybernetics*, volume 4, pages 3437–3443. IEEE, 2005.
- [93] Chuan Guo, Juan Cao, Xueyao Zhang, Kai Shu, and Miao Yu. Exploiting emotions for fake news detection on social media. *ArXiv*, abs/1903.01728, 2019.
- [94] Han Guo, Juan Cao, Yazhi Zhang, Junbo Guo, and Jintao Li. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 943–951, 2018.
- [95] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: A real-time web-based system for assessing credibility of content on twitter. *CoRR*, abs/1405.5490, 2014.
- [96] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International conference on social informatics*, pages 228–243. Springer, 2014.
- [97] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy.

- In *Proceedings of the 22nd International Conference on World Wide Web, WWW '13 Companion*, page 729–736, New York, NY, USA, 2013. Association for Computing Machinery.
- [98] Kilem Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *The British journal of mathematical and statistical psychology*, 61:29–48, 06 2008.
  - [99] Jonathan Haidt and Jesse Graham. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116, 2007.
  - [100] Yi Han, Shanika Karunasekera, and Christopher Leckie. Graph neural networks with continual learning for fake news detection from social media. *arXiv preprint arXiv:2007.03316*, 2020.
  - [101] Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection. *arXiv preprint arXiv:2010.08768*, 2020.
  - [102] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. Claimbuster: The first-ever end-to-end fact-checking system. *Proc. VLDB Endow.*, 10(12):1945–1948, aug 2017.
  - [103] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
  - [104] Benjamin Heinzerling. Nlp’s clever hans moment has arrived. *The Gradient*, 2019.
  - [105] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
  - [106] Benjamin D Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Eleventh international AAAI conference on web and social media*, 2017.
  - [107] Seyedmehdi Hosseini-motlagh and Evangelos E Papalexakis. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In *Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*, 2018.
  - [108] Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjuan Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. Compare to the knowledge: Graph neural fake news

- detection with external knowledge. 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)*, pages 754–763, 2021.
- [109] G Huang, Z Liu, L Van Der Maaten, and KQ Weinberger. Densely connected convolutional networks in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708, 2017.
  - [110] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2310–2318, 2017.
  - [111] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34, 2021.
  - [112] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018.
  - [113] Farnaz Jahanbakhsh, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021.
  - [114] Ujun Jeong, Kaize Ding, Lu Cheng, Ruocheng Guo, Kai Shu, and Huan Liu. Nothing stands alone: Relational fake news detection with hypergraph neural networks. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 596–605. IEEE, 2022.
  - [115] Shan Jiang and Christo Wilson. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23, 2018.
  - [116] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816, 2017.
  - [117] Zhiwei Jin, Juan Cao, Jiebo Luo, and Yongdong Zhang. Image credibility analysis with effective domain transferred deep networks. *arXiv preprint arXiv:1611.05328*, 2016.
  - [118] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

- [119] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608, 2016.
- [120] Marcia K Johnson and Carol L Raye. Reality monitoring. *Psychological review*, 88(1), 1981.
- [121] Zhezhou Kang, Yanan Cao, Yanmin Shang, Tao Liang, Hengzhu Tang, and Lingling Tong. Fake news detection with heterogenous deep graph convolutional network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 408–420, 2021.
- [122] Debanjana Kar, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. No rumours please! a multi-indic-lingual approach for covid fake-tweet detection. In *2021 Grace Hopper Celebration India (GHCI)*, pages 1–5. IEEE, 2020.
- [123] Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. Multi-source multi-class fake news detection. In *Proceedings of the 27th international conference on computational linguistics*, pages 1546–1557, 2018.
- [124] Hamid Karimi and Jiliang 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)*, pages 3432–3442, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [125] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- [126] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921, 2019.
- [127] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020.
- [128] Urja Khurana and Bachelor Opleiding Kunstmatige Intelligentie. The linguistic features of fake news headlines and statements. *Diss. Master’s thesis, University of Amsterdam*, 2017.
- [129] Yoon Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.

- [130] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [131] Walter Kintsch. The role of knowledge in discourse comprehension: A construction-integration model. In G.E. Stelmach and P.A. Vroom, editors, *Text and Text Processing*, volume 79 of *Advances in Psychology*, pages 107–153. North-Holland, 1991.
- [132] Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.
- [133] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, page 32–73, 2017.
- [134] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [135] Dhruv Kuchhal, Madhur Tandon, and Suryatej Reddy Vyalla. Whatsfarzi: Analyzing fake / manipulated content / misinformation on whatsapp, <https://precog.iiit.ac.in/blog/2019/02/11/whatsfarzi-analyzing-fake-manipulated-content-misinformation-on-whatsapp/>.
- [136] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602, 2016.
- [137] Peter J Lang. A bio-informational theory of emotional imagery. *Psychophysiology*, 16(6):495–512, 1979.
- [138] Johannes Langguth, Konstantin Pogorelov, Stefan Brenner, Petra Filkuková, and Daniel Thilo Schroeder. Don’t trust your eyes: Image manipulation in the age of deepfakes. *Frontiers in Communication*, page 26, 2021.
- [139] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.
- [140] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

- [141] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.
- [142] W Howard Levie and Richard Lentz. Effects of text illustrations: A review of research. *Ectj*, 30(4):195–232, 1982.
- [143] Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *arXiv preprint arXiv:2011.04088*, 2020.
- [144] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In icu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.
- [145] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.
- [146] Yunyao Li, Tyrone Grandison, Patricia Silveyra, Ali Douraghy, Xinyu Guan, Thomas Kieselbach, Chengkai Li, and Haiqi Zhang. Jennifer for COVID-19: An NLP-powered chatbot built for the people and by the people to combat misinformation. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.
- [147] Hause Lin, Gordon Pennycook, and David G Rand. Thinking more or thinking differently? using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition*, 230:105312, 2023.
- [148] Nankai Lina, Sihui Fua, and Shengyi Jianga. Fake news detection in the urdu language using charcnn-roberta. *Health*, 100:100, 2020.
- [149] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning. *arXiv preprint*, 2018.
- [150] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning, 2018.
- [151] Peng Liu, Wenhua Qian, Dan Xu, Bingling Ren, and Jinde Cao. Multi-modal fake news detection via bridging the gap between modals. *Entropy*, 25(4):614, 2023.
- [152] Yang Liu and Yi-Fang Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [153] Antonio Lopez and Jeff Share. Fake climate news: How denying climate change is the ultimate in fake news. *Education*, 2010, 2010.



- [154] Yi-Ju Lu and Cheng-Te Li. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648*, 2020.
- [155] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. 2016.
- [156] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1751–1754, 2015.
- [157] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics, 2017.
- [158] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference, WWW '19*, page 3049–3055, New York, NY, USA, 2019. Association for Computing Machinery.
- [159] Babak Mahdian and Stanislav Saic. Using noise inconsistencies for blind image forensics. *Image and Vision Computing*, 27(10):1497–1503, 2009.
- [160] Cameron Martel, Mohsen Mosleh, and David Gertler Rand. You’re definitely wrong, maybe: Correction style has minimal effect on corrections of misinformation online. 2021.
- [161] Richard E Mayer. Multimedia learning. In *Psychology of learning and motivation*, volume 41, pages 85–139. Elsevier, 2002.
- [162] Steven A McCornack, Kelly Morrison, Jihyun Esther Paik, Amy M Wisner, and Xun Zhu. Information manipulation theory 2: A propositional theory of deceptive discourse production. *Journal of Language and Social Psychology*, 33(4):348–377, 2014.
- [163] Mary McHugh. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82, 10 2012.
- [164] Nikhil Mehta, María Leonor Pacheco, and Dan Goldwasser. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1380, 2022.
- [165] Michail Mersinias, Stergos Afantenos, and Georgios Chalkiadakis. CLFD: A novel vectorization technique and its application in fake news detection. In *Proceedings of*

- the 12th Language Resources and Evaluation Conference*, pages 3475–3483, Marseille, France, May 2020. European Language Resources Association.
- [166] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
  - [167] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.
  - [168] Rahul Mishra and Vinay Setty. Sadhan: Hierarchical attention networks to learn latent aspect embeddings for fake news detection. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR ’19, page 197–204, New York, NY, USA, 2019. Association for Computing Machinery.
  - [169] Saumitra Mishra, Bob L Sturm, and Simon Dixon. Local interpretable model-agnostic explanations for music content analysis. In *ISMIR*, volume 53, pages 537–543, 2017.
  - [170] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1359–1367, 2020.
  - [171] Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V Mancenido, and Huan Liu. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, pages 3632–3640, 2022.
  - [172] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a twitter field experiment. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.
  - [173] Taichi Murayama, Shoko Wakamiya, and Eiji Aramaki. Mitigation of diachronic bias in fake news detection dataset. *arXiv preprint arXiv:2108.12601*, 2021.
  - [174] Kai Nakamura, Sharon Levy, and William Yang Wang. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France, May 2020. European Language Resources Association.
  - [175] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 299–307, 2017.

- [176] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, BS Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, and Amit K Roy-Chowdhury. Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging*, 2019(5):532–1, 2019.
- [177] N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.
- [178] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML’11, page 689–696, Madison, WI, USA, 2011. Omnipress.
- [179] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1165–1174, 2020.
- [180] Jeppe Nørregaard, Benjamin D Horne, and Sibel Adalı. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 630–638, 2019.
- [181] Nicole O’Brien, Sophia Latessa, Georgios Evangelopoulos, and Xavier Boix. The language of fake news: Opening the black-box of deep learning based detectors. 2018.
- [182] Nelleke Oostdijk, Hans van Halteren, Erkan Başar, and Martha Larson. The connection between the text and images of news articles: New insights for multimedia analysis. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4343–4351, Marseille, France, May 2020. European Language Resources Association.
- [183] Naomi Oreskes and Erik M Conway. *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Publishing USA, 2011.
- [184] Sunday Oluwafemi Oyeyemi, Elia Gabarron, and Rolf Wynn. Ebola, twitter, and misinformation: a dangerous combination? *Bmj*, 349, 2014.
- [185] Britt Paris, Rebecca Reynolds, and Gina Marcello. Disinformation detox: teaching and learning about mis-and disinformation using socio-technical systems research perspectives. *Information and Learning Sciences*, 2022.
- [186] Jinkyung Park, Rahul Ellezhuthil, Ramanathan Arunachalam, Lauren Feldman, and Vivek Singh. Toward fairness in misinformation detection algorithms. In *Workshop*

*Proceedings of the 16th International AAAI Conference on Web and Social Media.*  
Retrieved from <https://doi.org/10.36190>, 2022.

- [187] Jinkyung Park, Rahul Dev Ellezhuthil, Joseph Isaac, Christoph Mergerson, Lauren Feldman, and Vivek Singh. Misinformation detection algorithms and fairness across political ideologies: The impact of article level labeling. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, page 107–116, New York, NY, USA, 2023. Association for Computing Machinery.
- [188] Archita Pathak and Rohini K Srihari. Breaking! presenting fake news corpus for automated fact checking. In *Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop*, pages 357–362, 2019.
- [189] Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*, pages 21–29. Springer, 2021.
- [190] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.
- [191] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [192] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management science*, 66(11):4944–4957, 2020.
- [193] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595, 2021.
- [194] Gordon Pennycook and David G Rand. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature communications*, 13(1):2333, 2022.
- [195] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.
- [196] Billy Perrigo. How whatsapp is fueling fake news ahead of india’s elections, 2019.

- [197] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Cred-eye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 155–158, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.
- [198] Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, and Jesús Jaime Moreno Escobar. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876, 2019.
- [199] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.
- [200] Piotr Przybyla. Capturing the style of fake news. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 490–497, 2020.
- [201] Vahed Qazvinian, Emily Rosengren, Dragomir Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1589–1599, 2011.
- [202] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Ly, Chenyang Guo, and Yingchao Yu. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. *MM '21*, page 1212–1220, New York, NY, USA, 2021. Association for Computing Machinery.
- [203] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining (ICDM)*, pages 518–527. IEEE, 2019.
- [204] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937, 2017.
- [205] Bhavtosh Rath, Xavier Morales, and Jaideep Srivastava. SCARLET: explainable attention based graph neural network for fake news spreader prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2021.
- [206] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, 2013.

- [207] Julio CS Reis, Philipe Melo, Kiran Garimella, Jussara M Almeida, Dean Eckles, and Fabrício Benevenuto. A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 903–908, 2020.
- [208] Julio CS Reis, Philipe Melo, Kiran Garimella, Jussara M Almeida, Dean Eckles, and Fabrício Benevenuto. A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 903–908, Atlanta, Georgia, U.S., 2020. Proceedings of the Fourteenth International AAAI Conference on Web and Social Media.
- [209] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems (NeurIPS)*, pages 91–99, 2015.
- [210] Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17, 2016.
- [211] Victoria L Rubin and Tatiana Lukoianova. Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology*, 66(5):905–917, 2015.
- [212] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.
- [213] Fatima K Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. Fa-kes: A fake news dataset around the syrian war. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 573–582, 2019.
- [214] Giovanni C Santia and Jake Ryland Williams. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Twelfth international AAAI conference on web and social media*, 2018.
- [215] Roney Santos, Gabriela Pedro, Sidney Leal, Oto Vale, Thiago Pardo, Kalina Bontcheva, and Carolina Scarton. Measuring the impact of readability features in fake news detection. In *Proceedings of the 12th language resources and evaluation conference*, pages 1404–1413, 2020.
- [216] Wolfgang Schnotz. Commentary: Towards an integrated view of learning from text and visual displays. *Educational psychology review*, 14(1):101–120, 2002.

- [217] Wolfgang Schnotz and Maria Bannert. Construction and interference in learning from multiple representation. *Learning and instruction*, 13(2):141–156, 2003.
- [218] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [219] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. Nlp-based feature extraction for the detection of covid-19 misinformation videos on youtube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.
- [220] Gautam Kishore Shahi and Durgesh Nandini. Fakecovid—a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*, 2020.
- [221] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, 96:104, 2017.
- [222] Dilip Kumar Sharma and Sonal Garg. Ifnd: a benchmark dataset for fake news detection. *Complex & Intelligent Systems*, pages 1–21, 2021.
- [223] Richa Sharma and Arti Arya. Lfwe: Linguistic feature based word embedding for hindi fake news detection. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, mar 2023. Just Accepted.
- [224] Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4543–4556, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [225] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405, 2019.
- [226] Kai Shu, Deepak Mahudeswaran, and Huan Liu. Fakenewstracker: a tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*, 25:60–71, 2019.
- [227] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fak-newsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.
- [228] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings*

- of the international AAAI conference on web and social media, volume 14, pages 626–637, 2020.
- [229] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.
  - [230] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 312–320, 2019.
  - [231] Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston, and Huan Liu. Early detection of fake news with multi-source weak social supervision. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 650–666. Springer, 2020.
  - [232] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 436–439, 2019.
  - [233] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
  - [234] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
  - [235] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh. Spotfake: A multi-modal framework for fake news detection. In *IEEE International Conference on Multimedia Big Data (BigMM)*, pages 39–47, 2019.
  - [236] Shivangi Singhal, Mudit Dhawan, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Inter-modality discordance for multimodal fake news detection. In *ACM Multimedia Asia, MMAAsia ’21*, New York, NY, USA, 2022. Association for Computing Machinery.
  - [237] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. Spotfake+: A multimodal framework for fake news detection via transfer learning. *Proceedings of the AAAI Conference*, pages 13915–13916, 2020.
  - [238] Shivangi Singhal, Rishabh Kaushal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Fake news in india: scale, diversity, solution, and opportunities. *Communications of the ACM*, 65(11):80–81, 2022.
  - [239] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Leveraging intra and inter modality relationship for multimodal fake



- news detection. In *Companion Proceedings of the Web Conference 2022, WWW '22*, page 726–734, New York, NY, USA, 2022. Association for Computing Machinery.
- [240] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion Proceedings of the Web Conference 2022, WWW '22*, page 726–734, New York, NY, USA, 2022. Association for Computing Machinery.
- [241] Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Sachbolopls: A realtime identification of fake covid-19 content on online social media. *ACM W India Celebrations of Women in Computing*, 2020.
- [242] Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Factdrill: A data repository of fact-checked social media content to study fake news incidents in india. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1322–1331, 2022.
- [243] Laura C Smith, Kenya J Lucas, and Carl Latkin. Rumor and gossip: Social discourse on hiv and aids. *Anthropology & Medicine*, 6(1):121–131, 1999.
- [244] Leslie N. Smith. No more pesky learning rate guessing games. *CoRR*, abs/1506.01186, 2015.
- [245] Jacob Soll. The long and brutal history of fake news. politico magazine, <https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/>, 2016.
- [246] Chenguang Song, Kai Shu, and Bin Wu. Temporally evolving graph neural network for fake news detection. *Information Processing & Management*, 58(6):102712, 2021.
- [247] Ting Su, Craig Macdonald, and Iadh Ounis. Ensembles of recurrent networks for classifying the relationship of fake news titles. *SIGIR'19*, page 893–896, New York, NY, USA, 2019. Association for Computing Machinery.
- [248] Victor Suarez-Lledo, Javier Alvarez-Galvez, et al. Prevalence of health misinformation on social media: systematic review. *Journal of medical Internet research*, 23(1):e17187, 2021.
- [249] Shengyun Sun, Hongyan Liu, Jun He, and Xiaoyong Du. Detecting event rumors on sina weibo automatically. In *Asia-Pacific web conference*, pages 120–131. Springer, 2013.
- [250] Cass R Sunstein. *On rumors: How falsehoods spread, why we believe them, and what can be done*. Princeton University Press, 2014.

- [251] Wen-Ying Sylvia Chou, Anna Gaysynsky, and Joseph N Cappella. Where we go from here: health misinformation on social media, 2020.
- [252] Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. New explainability method for bert-based model in fake news detection. *Scientific reports*, 11(1):23705, 2021.
- [253] C Szegedy, W Liu, Y Jia, P Sermanet, S Reed, D Anguelov, D Erhan, V Vanhoucke, and A Rabinovich. Going deeper with convolutions. in *proceedings of the ieee computer society conference on computer vision and pattern recognition*, 2015.
- [254] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca De Alfaro. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*, 2017.
- [255] Nguyen Thanh Tam, Matthias Weidlich, Bolong Zheng, Hongzhi Yin, Nguyen Quoc Viet Hung, and Bela Stantic. From anomaly detection to rumour detection using data streams of social platforms. *Proceedings of the VLDB Endowment*, 12(9):1016–1029, 2019.
- [256] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. Defining “fake news” a typology of scholarly definitions. *Digital journalism*, 6(2):137–153, 2018.
- [257] Yori Thijssen. Breaking the news: the effects of fake news on political attitudes. Master’s thesis, University of Twente, 2017.
- [258] Margaret Topf. Three estimates of interrater reliability for nominal data. *pages Nurs Res.* 35(4):253–5. doi: 10.1097/00006199–198607000–00020. PMID: 3636827, 1986.
- [259] Fatemeh Torabi Asr and Maite Taboada. Big data and quality data for fake news and misinformation detection. *Big Data & Society*, 6(1):2053951719843310, 2019.
- [260] Kathie M d’I Treen, Hywel TP Williams, and Saffron J O’Neill. Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 11(5):e665, 2020.
- [261] Maria Tsirintani. Fake news and disinformation in health care-challenges and technology tools. In *Public Health and Informatics*, pages 318–321. IOS Press, 2021.
- [262] Udo Undeutsch. Beurteilung der glaubhaftigkeit von aussagen. *Handbuch der psychologie*, 11:26–181, 1967.
- [263] Sander Van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. Inoculating the public against misinformation about climate change. *Global Challenges*, 1(2):1600008, 2017.

- [264] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.
- [265] Anand Venkatraman, Dhruvika Mukhija, Nilay Kumar, and SJ Nagpal. Zika virus misinformation on the internet. *Travel medicine and infectious disease*, 14(4):421–422, 2016.
- [266] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.
- [267] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
- [268] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22, 2014.
- [269] George-Alexandru Vlad, Mircea-Adrian Tanase, Cristian Onose, and Dumitru-Clementin Cercel. Sentence-level propaganda detection in news articles with transfer learning and BERT-BiLSTM-capsule model. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 148–154, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [270] Nguyen Vo and Kyumin Lee. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. *arXiv preprint arXiv:2010.03159*, 2020.
- [271] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 647–653, 2017.
- [272] Christian von der Weth, Jithin Vachery, and Mohan Kankanhalli. Nudging users to slow down the spread of fake news in social media. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2020.
- [273] Soroush Vosoughi, Mostafa ‘Neo’ Mohsenvand, and Deb Roy. Rumor gauge: Predicting the veracity of rumors on twitter. *ACM transactions on knowledge discovery from data (TKDD)*, 11(4):1–36, 2017.
- [274] Shuhui Wang, Yangyu Chen, Junbao Zhuo, Qingming Huang, and Qi Tian. Joint global and co-attentive representation learning for image-sentence retrieval. In *ACM International Conference on Multimedia (ACMMM)*, page 1398–1406, 2018.

- [275] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.
- [276] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857, 2018.
- [277] Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 516–523, 2020.
- [278] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 540–547, 2020.
- [279] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240:112552, 2019.
- [280] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policymaking, 2017.
- [281] Maxwell Weinzierl and Sanda Harabagiu. Identifying the adoption or rejection of misinformation targeting covid-19 vaccines in twitter discourse. In *Proceedings of the ACM Web Conference 2022*, pages 3196–3205, 2022.
- [282] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2362–2365, 2010.
- [283] Junfei Wu, Qiang Liu, Weizhi Xu, and Shu Wu. Bias mitigation for evidence-aware fake news detection by causal intervention. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2308–2313, 2022.
- [284] Ke Wu, Song Yang, and Kenny Q Zhu. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pages 651–662. IEEE, 2015.
- [285] Kun Wu, Xu Yuan, and Yue Ning. Incorporating relational knowledge in explainable fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2021.

- [286] Liang Wu and Huan Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the eleventh ACM international conference on Web Search and Data Mining*, pages 637–645, 2018.
- [287] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2560–2569, 2021.
- [288] Chen Yang, Xinyi Zhou, and Reza Zafarani. Checked: Chinese covid-19 fake news dataset. *Social Network Analysis and Mining*, 11(1):1–8, 2021.
- [289] Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD workshop on mining data semantics*, pages 1–7, 2012.
- [290] Fan Yang, Arjun Mukherjee, and Eduard Dragut. Satirical news detection and analysis using attention mechanism and linguistic features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1979–1989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [291] Fan Yang, Shiva K Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D Ragan, Shuiwang Ji, and Xia Hu. Xfake: Explainable fake news detector with visualizations. In *The World Wide Web Conference*, pages 3600–3604, 2019.
- [292] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5644–5651, 2019.
- [293] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.
- [294] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. Ti-cnn: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*, 2018.
- [295] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.
- [296] Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. Interpretable propaganda detection in news articles. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1597–1605, Held Online, September 2021. INCOMA Ltd.

- [297] Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5444–5454, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.
- [298] Markos Zampoglou, Symeon Papadopoulos, Yiannis Kompatsiaris, Ruben Bouwmeester, and Jochen Spangenberg. Web and social media image forensics for news professionals. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 159–166, 2016.
- [299] Daniel Yue Zhang, Lanyu Shang, Biao Geng, Shuyue Lai, Ke Li, Hongmin Zhu, Md Tanvir Amin, and Dong Wang. Fauxbuster: A content-free fauxtography detector using social media comments. In *2018 IEEE international conference on big data (big data)*, pages 891–900. IEEE, 2018.
- [300] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, pages 3465–3476, 2021.
- [301] Lina Zhou, Judee K Burgoon, Jay F Nunamaker, and Doug Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation*, 13(1):81–106, 2004.
- [302] Xiang Zhou, Heba Elfardy, Christos Christodoulopoulos, Thomas Butler, and Mohit Bansal. Hidden biases in unreliable news detection datasets. *arXiv preprint arXiv:2104.10130*, 2021.
- [303] Xing Zhou, Juan Cao, Zhiwei Jin, Fei Xie, Yu Su, Dafeng Chu, Xuehui Cao, and Junqiang Zhang. Real-time news certification system on sina weibo. In *Proceedings of the 24th international conference on world wide web*, pages 983–988, 2015.
- [304] Xinyi Zhou, Atishay Jain, Vir V Phoha, and Reza Zafarani. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2):1–25, 2020.
- [305] Xinyi Zhou, Kai Shu, Vir V Phoha, Huan Liu, and Reza Zafarani. “this is fake! shared it by mistake”: Assessing the intent of fake news spreaders. In *Proceedings of the ACM Web Conference 2022*, pages 3685–3694, 2022.
- [306] Xinyi Zhou, Kai Shu, Vir V Phoha, Huan Liu, and Reza Zafarani. “this is fake! shared it by mistake”: Assessing the intent of fake news spreaders. In *Proceedings of the ACM Web Conference 2022*, pages 3685–3694, 2022.
- [307] Xinyi Zhou, Jindi Wu, and Reza Zafarani. Safe: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2020.

- [308] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.
- [309] Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2120–2125, 2022.
- [310] Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. Fact-checking meets faux-tography: Verifying claims about images. 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)*, pages 2099–2108, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [311] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Exploiting context for rumour detection in social media. In *International conference on social informatics*, pages 109–123. Springer, 2017.
- [312] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989, 2016.
- [313] Miron Zuckerman, Bella M DePaulo, and Robert Rosenthal. Verbal and nonverbal communication of deception. In *Advances in experimental social psychology*, volume 14, pages 1–59. Elsevier, 1981.
- [314] RA Zwaan and GA Radvansky. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162—185, March 1998.