

ANALYSIS OF INFORMATION POLLUTION ON WEB AND SOCIAL MEDIA

A thesis Submitted to

DELHI TECHNOLOGICAL UNIVERSITY

For the Award of degree of

DOCTOR OF PHILOSOPHY

In

DEPARTMENT OF INFORMATION TECHNOLOGY

by

PRIYANKA MEEL

(2K18/PhD/IT/08)

Under the Supervision of

Prof. Dinesh Kumar Vishwakarma



Department of Information Technology
Delhi Technological University
Bawana Road, New Delhi-110042, India
(Formerly Delhi College of Engineering)
NOVEMBER 2021



© DELHI TECHNOLOGICAL UNIVERSITY-2021
ALL RIGHTS RESERVED

DECLARATION

I declare that the research work reported in the thesis entitled “**ANALYSIS OF INFORMATION POLLUTION ON WEB AND SOCIAL MEDIA**” for the award of the degree of *Doctor of Philosophy* in the *Department of Information Technology* has been carried out by me under the supervision of Prof. *Dinesh Kumar Vishwakarma*, Professor in Department of Information Technology, Delhi Technological University, Delhi, India.

The research work embodied in this thesis, except where otherwise indicated, is my original research. This thesis has not been submitted by earlier in part or full to any other University or Institute for the award of any degree or diploma. This thesis does not contain other person’s data, graphs or other information unless specifically acknowledged.

Date:



Priyanka Meel

2K18/Ph.D./IT/08

CERTIFICATE

This is to certify that the work contained in the thesis entitled “**ANALYSIS OF INFORMATION POLLUTION ON WEB AND SOCIAL MEDIA**” submitted by Ms. Priyanka Meel (**Reg. No.: 2K18/Ph.D./IT/08**) for the award of degree of Doctor of Philosophy to the Delhi Technological University is based on the original research work carried out by her. She has worked under my supervision and has fulfilled the requirements as per the requisite standard for the submission of the thesis. It is further certified that the work embodied in this thesis has neither partially nor fully submitted to any other university or institution for the award of any degree or diploma.



Prof. Dinesh Kumar Vishwakarma
Supervisor, Professor
Department of Information Technology
Delhi Technological University, Delhi

ACKNOWLEDGMENT

I am indebted to God under whose power I pursued this Ph. D. Thanks to Almighty for granting me wisdom, health, and strength to undertake this research task and enabling me to its completion.

I am grateful to my supervisor, **Prof. Dinesh Kumar Vishwakarma**, without whom this achievement would not have been realized. It was his valuable guidance and consistent encouragement all through my research period, which helped me to overcome the challenges that came in the way. This feat was possible only because of the unconditional support provided by him. A person with an amicable and positive temperament, I consider it a great opportunity to do my doctoral programme under his guidance and learn from his research experience.

My sincere regards to **Prof. J. P. Saini, Vice-Chancellor, Delhi Technological University** and **Prof. Yogesh Singh, Former Vice-Chancellor, Delhi Technological University** for providing me with a platform for pursuing my Ph.D. work. I express my gratitude to **Prof. Kapil Sharma, Head, Department of IT**, for his kind support and for providing the necessary facilities to undertake this research.

I do not have words to thank my fellow labmates Dr **Tej Singh**, Dr **Chhavi Dhiman**, Dr **Ashima Yadav**, **Deepika** and **Ankit Yadav** who offered me their time, help and support whenever needed.

Heartfelt and endless thanks to my parents, sister, and husband, who always supported me at every stage of my life. They strengthen me by providing moral and emotional support. I am confident to be energized with their care, affection, and brilliant advice for the rest of my life.



Priyanka Meel

ABSTRACT

In the era of information overload, restiveness, uncertainty and implausible content all around; information credibility or web credibility refers to the trustworthiness, reliability, fairness and accuracy of the information. Information credibility is the extent up to which a person believes in the content provided on the internet. Every second of time passes by millions of people interacting on social media, creating vast volumes of data, which has many unseen patterns and trends inside. The data disseminating on the web, social media and discussion forums have become a massive topic of interest for analytics as well as critics as it reflects social behaviour, choices, perceptions and mindset of people. A considerable amount of unverified and unauthenticated information travels through these networks, misleading a large population. Thus, to increase the trustworthiness of online social networks and mitigate the devastating effects of information pollution; timely detection and containment of false content circulating on the web are highly required.

To analyse and address the issue of information pollution on web and social media, we have initially reviewed the most popular and prominent state-of-the-art solutions, compared them and presented. Based on the literature survey, these solutions are categorized and analysed. The prevalent approaches in each modality are studied and highlighted in detail, which helped identify the research gaps in this area. To overcome the issue, our proposed solutions are focused on two categories: **semi-supervised textual fake news classification frameworks and supervised multimodal veracity analysis frameworks.**

The first model developed for semi-supervised textual fake news classification frameworks proposes an innovative Convolutional Neural Network built on the self-ensembling concept to take leverage of the linguistic and stylometric information of annotated news articles, at the same time explore the hidden patterns in unlabelled data as well. Next, we aim to design a semi-supervised fake news detection technique based on GCN (Graph Convolutional Networks). The recommended architecture comprises of three basic components: collecting word embeddings from the news articles in datasets utilising

GloVe, building similarity graph using Word Mover's Distance (WMD) and finally applying Graph Convolution Network (GCN) for binary classification of news articles in semi-supervised paradigm.

In the category of supervised multimodal veracity analysis frameworks, the first model consists of four independent parallel streams capable enough to detect specific forgery formats. All four streams are applied to each input instance. Hierarchical Attention Network deals with headline and body part; Image captioning and headline matching module require all the three parts headline, body and image. Noise Variance Inconsistency and Error Level Analysis focuses only on images accompanied with news text. These independent predictions are finally combined using the max voting ensemble method. The second model aims Inception-ResNet-v2 to extract visual features. The models BERT and ALBERT have been used to elicit textual attributes. Diverse text input forms, like English articles, Chinese articles and Tweets, have been used to make our model robust and usable across multiple platforms. The architecture of Multimodal Early Fusion and Late Fusion has also been experimented with and analysed in detail by applying it on different datasets.

Finally, this thesis work is concluded with significant findings and future research aspects in veracity analysis of web and social media information.

List of Publications

1. **Priyanka Meel**, Dinesh Kumar Vishwakarma. “Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities.” *Expert Systems with Applications*, 152 (2020): 112986 (**Impact Factor: 6.954**). (Pub: Elsevier).
2. **Priyanka Meel**, Dinesh Kumar Vishwakarma. “HAN, image captioning, and forensics ensemble multimodal fake news detection.” *Information Sciences*, 567 (2021): 23-41 (**Impact Factor: 6.795**). (Pub: Elsevier).
3. **Priyanka Meel**, Dinesh Kumar Vishwakarma. “A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles.” *Expert Systems with Applications*, 177 (2021): 115002 (**Impact Factor: 6.954**). (Pub: Elsevier).
4. **Priyanka Meel**, Dinesh Kumar Vishwakarma. “Multi-modal Fusion using Fine-tuned Self-attention and Transfer Learning for Veracity Analysis of Web Information.” *arXiv preprint arXiv:2109.12547 (2021)*. (**Communicated**)
5. **Priyanka Meel**, Dinesh Kumar Vishwakarma. “Fake News Detection using Semi-Supervised Graph Convolutional Network.” *arXiv preprint arXiv:2109.13476 (2021)*. (**Communicated**)
6. **Priyanka Meel**, Dinesh Kumar Vishwakarma. “Deep Neural Architecture for Veracity Analysis of Multimodal Online Information.” at *IEEE 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, 2021.
7. **Priyanka Meel**, Dinesh Kumar Vishwakarma. “Machine Learned Classifiers for Trustworthiness Assessment of Web Information Contents.” at **IEEE International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)**, Noida, 2021.

List of Figures

<i>Figure 1.1: Lifecycle of False Information</i>	3
<i>Figure 1.2: Venn Diagram of False Information on social media and Web</i>	5
<i>Figure 1.3: Number of active users/month of popular social networking platforms (data source [8])</i>	10
<i>Figure 1.4: (a) Social media as a news source according to age group (b) Awareness of people towards news truthfulness (data source [8])</i>	10
<i>Figure 1.5: Images Spreading Fake News on different Social and News Media Platforms ((a)- [24], (b)- [25],(c)- [28],(d)- [29],(e)- [32])</i>	11
<i>Figure 2.1: Steps of source identification of false information</i>	24
<i>Figure 2.2: Different Network structures used in credibility assessment methods</i>	28
<i>Figure 2.3: Classification of rumor containment strategies</i>	36
<i>Figure 3.1: ConvNet Semi-Supervised Framework</i>	43
<i>Figure 3.2: ConvNet Layered Architecture</i>	44
<i>Figure 3.3: Unsupervised Weight Function</i>	46
<i>Figure 3.4: Performance of Jruvika Dataset (50% labelled) in terms of (a) Training and Validation accuracy (b) Training and Validation Loss (c) ROC Curve (d) Confusion Matrix</i>	52
<i>Figure 3.5: Performance of Fake News Data Dataset (50% labelled) in terms of (a) Training and Validation accuracy (b) Training and Validation Loss (c) ROC Curve (d) Confusion Matrix</i>	53
<i>Figure 3.6: Performance of Fake News Sample Dataset (50% labelled) in terms of (a) Training and Validation accuracy (b) Training and Validation Loss (c) ROC Curve (d) Confusion Matrix</i>	54

<i>Figure 3.7: Comparative analysis of all the three datasets for (a) Accuracy (b) Precision (c) Recall (d) F1-Score with variation in % of labelled data</i>	<i>55</i>
<i>Figure 3.8: Proposed Semi-Supervised Fake News Classification Framework.....</i>	<i>61</i>
<i>Figure 3.9: An example co-occurrence matrix for the sentence “the cat sat on the mat”</i>	<i>62</i>
<i>Figure 3.10: Design pipeline for GCN Model.....</i>	<i>65</i>
<i>Figure 3.11: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for Fake News Data Dataset</i>	<i>70</i>
<i>Figure 3.12: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for Real or Fake dataset.....</i>	<i>71</i>
<i>Figure 3.13: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for Fake News Detection dataset.....</i>	<i>71</i>
<i>Figure 4.1: Sample visual and textual Fake news on web platforms</i>	<i>76</i>
<i>Figure 4.2: Architecture of Proposed Model.....</i>	<i>78</i>
<i>Figure 4.3: Pre-processing Text and Multimedia data</i>	<i>79</i>
<i>Figure 4.4: Few Examples of generated Image Captions</i>	<i>83</i>
<i>Figure 4.5: (a) Original photograph (b) Error Level Analysis of Original photograph (c) Modified and resaved photograph (d) Error Level Analysis of modified and resaved photograph [192].....</i>	<i>85</i>
<i>Figure 4.6: (a) Performance Comparison of individual and ensemble model (b) Confusion Matrix (c) ROC curve for Fake news detection dataset</i>	<i>89</i>
<i>Figure 4.7: (a) Performance Comparison of individual and ensemble model (b) Confusion Matrix (c) ROC curve for All Data dataset</i>	<i>90</i>
<i>Figure 4.8: (a) Performance Comparison of individual and ensemble model (b) Confusion Matrix (c) ROC curve for Fake News Samples dataset</i>	<i>91</i>

<i>Figure 4.9: Overall Performance Comparison of three datasets</i>	91
<i>Figure 4.10: Architecture of Multimodal Early fusion</i>	101
<i>Figure 4.11: Architecture of Multimodal Late fusion</i>	101
<i>Figure 4.12: BERT Input Representation</i>	104
<i>Figure 4.13: (a) Scaled Dot-Product Attention (b) Multi-Head Attention</i>	105
<i>Figure 4.14: Schema for Inception-ResNet-v2 Network</i>	107
<i>Figure 4.15: Schema for (a) Stem (b) Inception-Resnet-A (c) Reduction-A blocks of Inception-ResNet-v2 Network</i>	108
<i>Figure 4.16: Schema for (a) Inception-Resnet-B (b) Reduction-B (c)Inception-Resnet-C blocks of Inception-ResNet-v2 Network</i>	109
<i>Figure 4.17: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for All Data dataset</i>	114
<i>Figure 4.18: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for Weibo dataset</i>	115
<i>Figure 4.19: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for MediaEval dataset</i>	116

List of Tables

<i>Table 1.1: Categorization of False Information</i>	5
<i>Table 1.2: Motivation Behind Information Pollution</i>	6
<i>Table 1.3: Facts about social networking platforms</i>	8
<i>Table 1.4: Few Examples of false information</i>	10
<i>Table 1.5: Public and Commercial Social Media Analytics Tools</i>	12
<i>Table 1.6: List of Fact-Checking Platforms</i>	13
<i>Table 2.1: Scalable synthetic social network graph generators</i>	28
<i>Table 2.2: Feature-Based Methods for Fake News Detection</i>	32
<i>Table 3.1: Dimensions of ConvNet Semi-Supervised Architecture</i>	48
<i>Table 3.2: Parameter Tuning in ConvNet Semi-Supervised Architecture</i>	49
<i>Table 3.3: Dataset Details</i>	50
<i>Table 3.4: Result Analysis on Fake News Detection (Jruvika) Dataset</i>	51
<i>Table 3.5 : Result Analysis on Fake News Data Dataset</i>	53
<i>Table 3.6: Result Analysis on Fake News Sample (Pontes) Dataset</i>	54
<i>Table 3.7: Performance comparison on Fake News Detection (Jruvika) Dataset</i>	57
<i>Table 3.8: Performance comparison on Fake News Data Dataset</i>	58
<i>Table 3.9: Performance comparison on Fake News Sample (Pontes) Dataset</i>	58
<i>Table 3.10: Dataset Details</i>	68
<i>Table 3.11: Result Analysis on Fake News Data Dataset</i>	70
<i>Table 3.12: Result Analysis on Real or Fake Dataset</i>	70

<i>Table 3.13: Result analysis on Fake News Detection dataset</i>	<i>71</i>
<i>Table 3.14: Comparative Performance Analysis on Fake News Data Dataset.....</i>	<i>73</i>
<i>Table 3.15: Comparative Performance Analysis on Real or Fake Dataset</i>	<i>73</i>
<i>Table 3.16: Comparative Performance Analysis on Fake News Detection Dataset</i>	<i>73</i>
<i>Table 4.1: Dataset Details</i>	<i>87</i>
<i>Table 4.2: Result Analysis on Fake News Detection Dataset.....</i>	<i>88</i>
<i>Table 4.3: Result Analysis on All Data Dataset</i>	<i>89</i>
<i>Table 4.4: Result Analysis on Fake News Samples Dataset.....</i>	<i>90</i>
<i>Table 4.5: Ablation Study of proposed Ensemble Multimodal Framework.....</i>	<i>93</i>
<i>Table 4.6: Parameter Analysis for setting thresholds in Algorithm 1</i>	<i>95</i>
<i>Table 4.7: Parameter Analysis for setting thresholds in Algorithm 2.....</i>	<i>95</i>
<i>Table 4.8: Parameter Analysis for setting thresholds in Algorithm 3.....</i>	<i>96</i>
<i>Table 4.9: Performance comparison with state-of-the-arts on Fake News Detection dataset</i>	<i>97</i>
<i>Table 4.10: Performance comparison with state-of-the-arts on All Data Dataset.....</i>	<i>98</i>
<i>Table 4.11: Performance comparison with state-of-the-arts on Fake News Samples Dataset</i>	<i>98</i>
<i>Table 4.12: Dataset Details</i>	<i>111</i>
<i>Table 4.13: Result Analysis on All Data Dataset</i>	<i>113</i>
<i>Table 4.14: Result analysis on Weibo Dataset</i>	<i>114</i>
<i>Table 4.15: Result Analysis on MediaEval Dataset.....</i>	<i>115</i>
<i>Table 4.16: State-of-the-art Comparison on All Data Dataset.....</i>	<i>117</i>

Table 4.17: State-of-the-art comparison on weibo Dataset..... 117

Table 4.18: State-of-the-art comparison on MediaEval Dataset..... 118

Table of Contents

DECLARATION	ii
CERTIFICATE	iii
ACKNOWLEDGMENT	iv
ABSTRACT	v
List of Publications	vii
List of Figures	viii
List of Tables	xi
Chapter 1 Introduction	1
1.1 Background	1
1.2 False Information Ecosystem	3
1.3 Categorization of False Information	4
1.4 Factors/Motivation for Spreading	6
1.5 Social Impact.....	7
1.6 User Perception	11
1.7 Current State of Fact Checking	13
1.8 Motivation behind the work	15
1.9 Problem Statement	17
1.10 Major Contributions of Thesis	17
1.11 Thesis overview.....	19
Chapter 2 Literature Review	22
2.1 Introduction	22
2.2 Veracity Analysis Approaches	23
2.2.1 Source Identification.....	24

2.2.2 Propagation Pattern Analysis.....	25
2.2.3 Network Structure Analysis.....	27
2.2.4 Text Based Approaches	29
2.2.5 Visual Approaches.....	30
2.2.6 Multi-modal Approaches	30
2.2.7 Feature Based Approaches	31
2.2.8 Supervised Methods	33
2.2.9 Semi-Supervised Methods	34
2.2.10 Unsupervised Methods	35
2.2.11 Containment and Intervention Methods	36
2.2.12 Other Approaches	37
2.3 Gaps Identified in the Present Study.....	39
2.4 Research Objectives	40
Chapter 3 Fake News Detection Using Semi-Supervised Textual Frameworks	41
3.1 Introduction	41
3.2 A Temporal Ensembling based Semi-supervised ConvNet for the Detection of Fake News Articles	42
3.2.1 Methodology.....	43
3.2.2 Model Parameter Description.....	48
3.2.3 Datasets.....	50
3.2.4 Result Analysis.....	51
3.2.5 Baseline Comparison.....	56
3.2.6 Significant Outcomes	59
3.3 Fake News Detection Using Semi-Supervised Graph Convolutional Network.....	60
3.3.1 Article Embedding in Euclidean Space	61

3.3.2 Similarity Graph Construction.....	62
3.3.3 GCN for Graph Classification	65
3.3.4 Implementation Details.....	67
3.3.5 Datasets.....	68
3.3.6 Result Analysis	69
3.3.7 State-of-the-art Comparison	72
3.3.8 Significant Outcomes	74
Chapter 4 Fake News Detection Using Supervised Multimodal Frameworks.....	75
4.1 Introduction	75
4.2 HAN, Image Captioning and Forensics Ensemble Multimodal Fake News Detection	77
4.2.1 Pre-processing and Word Embedding	79
4.2.2 Hierarchical Attention Network (HAN).....	80
4.2.3 Image Caption and Headline Matching with News Text (CHM).....	82
4.2.4 Noise Variance Inconsistency (NVI).....	84
4.2.5 Error Level Analysis (ELA)	85
4.2.6 Ensemble with Max Voting.....	86
4.2.7 Implementation Details.....	86
4.2.8 Result Analysis	88
4.2.9 Ablation Study	93
4.2.10 Parameter Analysis for Setting Thresholds	94
4.2.11 State-of-the-art Comparison	96
4.2.12 Significant Outcomes	99
4.3 Multi-modal Fusion Using Fine-tuned Self-attention and Transfer Learning for Veracity Analysis of Web Information	100
4.3.1 Pre-processing	102

4.3.2 BERT	103
4.3.3 ALBERT.....	106
4.3.4 Inception-ResNet-v2.....	106
4.3.5 Multi-modal Fusion	109
4.3.6 Implementation Details.....	110
4.3.7 Model Parameter Description.....	112
4.3.8 Result Analysis	113
4.3.9 State-of-the-art Comparison	116
4.3.10 Significant Outcomes	118
Chapter 5 Conclusion and Future Scope	120
5.1 Conclusions	120
5.2 Future Scope.....	122
References	125
Author Biography	139

Chapter 1

Introduction

Internet and social media have become a widespread, large scale and easy to use platform for real-time information dissemination. Outlets of social podiums have encouraged the spread of fake news with considerable impact because it is easier and faster to produce content online as barricades to join the online broadcasting industry have dropped. This chapter introduced the background of the information pollution, different flavors of false information, factors motivating its spread, social impact, user perception and current state of fact-checking. Furthermore, research problem statements, significant contributions, motivations for the research, significance of the study, and thesis organization are discussed.

1.1 Background

Social media and web platforms have become an open stage for discussion, ideology expression, knowledge dissemination, emotions and sentiment sharing. These platforms are gaining tremendous attraction and a huge user base from all sections and age groups of society. The matter of concern is that up to what extent the contents that are circulating among all these platforms every second changing the mindset, perceptions and lives of billions of people are verified, authenticated and up to the standards. Internet-based information circulation has given rise to the proliferation of fake and misleading content, which has extremely hostile effects on individuals and humanity. Inaccurate or misleading information generate because of honest reporting errors, incorrect interpretations or sometimes deliberately spreading a biased agenda to deceive readers for financial or political profit. In all their forms, internet-based social and professional interactions could be exploited to spread fake news, which has substantial adverse effects on singular clients and broader society. Thus, to increase the trustworthiness of online social networks and mitigate the

devastating impact of information pollution, timely detection and containment of false content circulating on the web are highly required [1].

The section of the data on which we are focusing is information pollution i.e. how the contents on the web are being contaminated intentionally or sometimes unintentionally. The false information may be in any format fake review, fake news, satire, hoax, etc.; affects the human community in a negative way. Approximately 65% of the US adult population is dependent on social media for daily news [2]. If we grab the information without showing severe concern about its truthfulness, we have to pay in the long run. Social networks information diffusion has robust temporal features: Bursting updates, flooding all platforms with the carnival of information within no time (of course without fact-checking) and finally, fast dying feature. Official news media is also losing trust and confidence; in the rush of securing readership, they are releasing eye-catching and sensational headlines with images. The readers do not have the time to read the actual news content; they trust the appealing headline and the image. Thus, appealing headlines give birth to a misunderstood, falsified piece of information.

Earlier rumors used to spread at a slow pace, but the advent of internet technologies and the popularity of retweeting activities on social networks have fueled the dissemination of a piece of rumor around the globe at an alarming rate. In the 2016, US presidential elections, because of some flaws in algorithmic architecture, Facebook has become a key distributor of fake news [3], which has affected people's choice of the vote and had a tremendous impact on the election result. It is a remarkable example of how fake news accounts had outperformed real news. The main lineage of work done by researchers in web and social media mining is in tweeting behavior analysis, feature extraction, trends and pattern analysis, information diffusion, visualization, anomaly detection, predictive analysis, recommender systems, and situation awareness [4] , [5], [6], [7]. Fake news detection algorithms focus on figuring out deep systematic patterns embedded inside the content of news. Another primary feature of detection is transmission behavior that strengthens the diffusion of information, which is of questionable integrity and value.

1.2 False Information Ecosystem

According to the Global digital report 2019 [8], out of the world's total population of 7.676 billion, there are 4.388 billion internet users and 3.484 billion social media users. Almost half of the world's total population depends upon the internet for their knowledge. However, how much or up to what extent the circulated facts are verified is still a big question. How much we can rely on the information content that we are browsing every day. False information is created and initiated by a small number of people. People, relations, content and time are four critical dimensions of networked data analyzed multi-dimensionally by proposing an iOLAP framework based on polyadic factorization approach [9]. This framework handles all types of networked data such as microblogs, social bookmarking, user comments and discussion platforms with an arbitrary number of dimensions. Origination, propagation, detection and Intervention are the four main facets of information pollution, which are diagrammatically represented in Figure 1.1.



Figure 1.1: Lifecycle of False Information

Origination deals with the creation of fake content by a single person, account or multiple accounts. Propagation analyses the reason behind the fast and large-scale spread of fraudulent content online. The analysis is done by [10], [11] sheds new light on fake news writing style, linguistic features and fraudulent content propagation trends; concludes that falsehood disseminates significantly faster, deeper, farther and more broadly than the truth in all the categories. False news is 70% more likely to be retweeted by many unique users, as fake stories are more novel, surprising and eye-catching; attracts human attention hence encourages information sharing. Identification of the misinformation and disinformation from the massive volume of social media data using different Artificial Intelligence technologies comes under detection. Finally, intervention methods concentrate on restricting the outspread of false information by spreading the truth.

Fake product review is an emerging field of forgery in online social networks, specifically in the field of e-commerce, as more and more people share their shopping experiences online through reviews [12]. The customer reviews are directly related to the reputation of a product in the E-commerce era. People consider ratings, feedback reviews, and comments by previous buyers to make an opinion on whether to purchase a particular item or not. The algorithms suggested in [13], [14], [15] for detecting fake movie reviews are based on sentiment analysis, temporal, statistical features and text classification. Ahmed et al. [16] use six supervised machine learning classifiers SVM, LSVM, KNN, DT, SGD, LR to detect fake reviews of hotels and fake news articles on the web using text classification. Their experiments achieve a significant accuracy of 90% and 92%, respectively. Different content-based, features-based, behavior-based and graph-based approaches [17] can be used to detect opinion spams present in various formats of fake reviews, fake comments, social network posting and fake messages. In addition to the mainstream news media, there is also a concept of alternative media [18] that aims to just present the facts and let readers use their critical thinking to explore reality by means of discussions.

1.3 Categorization of False Information

False information, which is present in the form of images, blogs, messages, stories, break-

ing news; generally termed as information pollution, has many formats that are not mutually exclusive but at the same time also have some heterogeneity that brings them under a specific category. The categorization of different information pollution formats is represented by means of a Venn diagram in Figure 1.2. Table 1.1 summarizes different categories and the impact of fraudulent content on the internet. Although each category has some salient characteristics, we have used the terms interchangeably at many places to provide a complete synergy of information pollution on the digital communication platform.

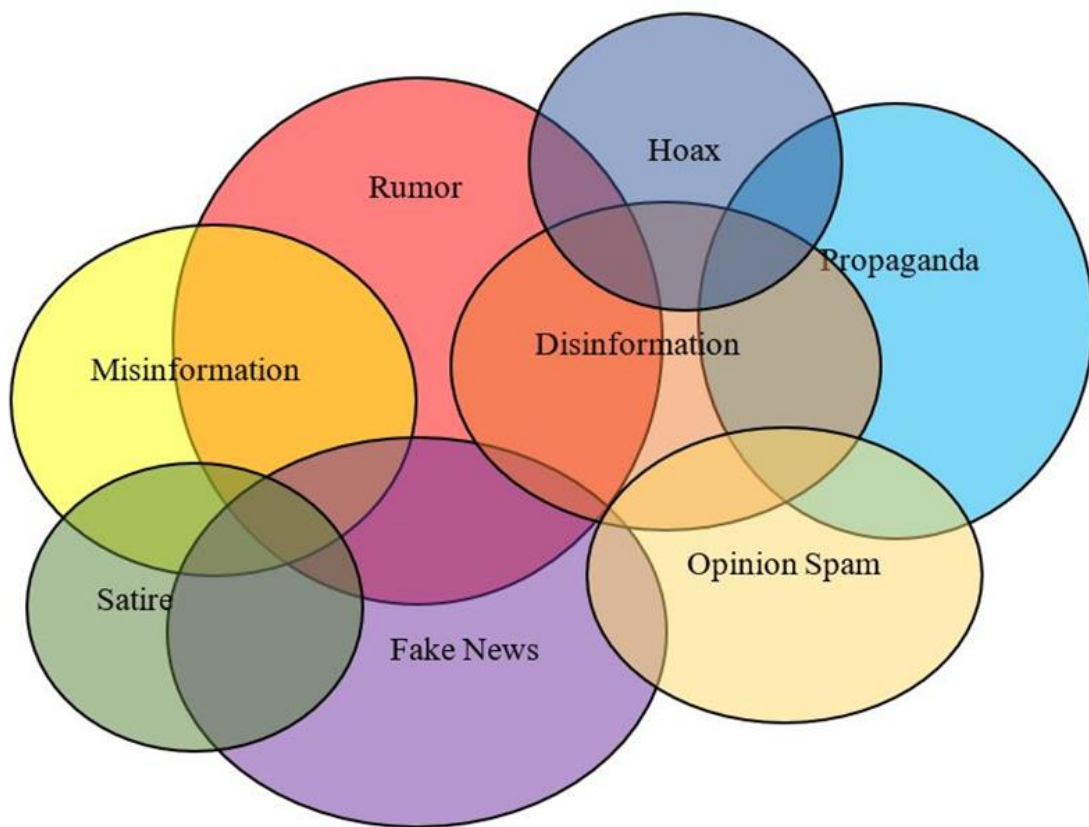


Figure 1.2: Venn Diagram of False Information on social media and Web

Table 1.1: Categorization of False Information

Category	Definition	Impact
Rumor	Unverified piece of information which is not necessarily false; may turn out to be true also	Uncertainty and confusion about facts
Fake News	False information spread under the guise of being authentic news usually distributed through news outlets or internet to gain politically or financially, increase readership, biased public opinion	to damage an agency, entity, or person or gain financial/political profit

Misinformation	Circulating information that becomes false inadvertently as a consequence of an honest mistake, carelessness or cognitive bias	Less harmful but wrong interpretation of facts can lead to significant damage
Disinformation	Deliberately deceptive information with a predefined intention	To promote a belief, idea, financial gain or tarnish an opponent's image
Clickbait	The deliberate use of misleading headlines to encourage visitors to click on a particular webpage	To earn advertising revenue, to trigger phishing attacks
Hoax	The false story, primarily through Joke, prank, humor or malicious deception, used to masquerade the truth	Falsehood is perceived as truth and reality
Satire/parody	Articles that primarily contain humor and irony, no harmful intention but have the potential to fool. The Onion [19] and Satire Wire [20] are sources of satirical news articles.	The motive is fun but sometimes exert adverse effects also
Opinion Spam	Fake or intentionally biased reviews or comments about products and services	untruthful customer opinion
propaganda	Unfairly prejudiced and deceptive information spread in targeted communities according to a predefined strategy to promote a particular viewpoint or political agenda	Political/financial profit
Conspiracy theories	an explanation of an event that invokes a conspiracy by sinister and powerful actors, often political in motivation-based entirely on prejudice or insufficient evidence	Extremely harmful to people and society

1.4 Factors/Motivation for Spreading

Interactions of people on social media give rise to a lot of information content which turns out to be false, sometimes intentionally with a predefined motive or unintentionally by mistake. The following Table 1.2 details the key reasons behind the increasing spread of misleading content on online platforms:

Table 1.2: Motivation Behind Information Pollution

















Motive	Description
Political Intent	to tarnish the public image of the opponent or promote a person or party
Financial Profit	False-positive information triggers the motivation for large-scale investments and affects stock prices. Fake ratings and reviews of products are intentionally written to increase sales.
Passion for promoting an ideology	A considerable number of people are impassioned about a particular organization, ideology, person or philosophy and they want to spread it by any means.
Fun	For amusement and fun, satirical sites write humorous content that is often mistaken for real news. This is the least severe motive, which does not have many harmful effects because intentions are not usually wrong.

Increase customer base	In the era of Internet-based journalism, online news media is rushing to secure readership and increase customer base. Thus, publishing the stories of questionable integrity and content in the process to lure readers to their websites and platforms
Rush to cover the latest news	In a competition to be the first to cover the story, journalists often publish articles without fact-checking and get millions and millions of views. Truth and integrity become liabilities in the current online journalism with aims to” Publish first, correct if necessary”
Generate advertising revenues	Fake news creators have earned a sizable profit from automated advertising engines like AppNexus, Facebook Ads and Google AdSense [21] during the 2016 US presidential elections. Earning capital through false advertising news is a significant driving force that an entire cottage industry of practitioners has indulged in this controversial endeavor.
Technological Reasons	Algorithms are structured to endorse things based on popularity, not accuracy[21]. Echo chambers and filter bubbles in search engines [22] are some of the algorithmic flows accounts for biased information circulation. Therefore, they are agnostically promoting the spread of disinformation as fake news is intentionally designed to gain more user attention.
Manipulate public opinion	In a consumer-based economy, public opinion regarding a firm, service, product or people holds significant importance as customers are going to decide the fate of stocks, sales, election results, all types of businesses and many more.

1.5 Social Impact

Social networking platforms launched in the past two decades play a role in social interactions by providing easy-to-use features to exchange information in many formats. Table 1.3 summarizes popular social networking platforms along with their customer base and salient features (data source [23] and Wikipedia). Figure 1.3 shows the popularity statistics on major social platforms (data source [8]). Figure 1.4 (a) and Figure 1.4 (b) explain some statistics based on age and country about social media users (data source [8]). Around the globe, 54% of people express strong concern about “what is real or fake” when thinking about online news. The younger section of the population is under more influence of Internet-based knowledge, and as the age grows, according to statistics, this ratio decreases. Table 1.4 supported with Figure 1.5 (a-e) explains some of the prominent havocs created in society in recent years due to information pollution and classifies them according to the taxonomy provided in section 1.3.

Table 1.3: Facts about social networking platforms

Name	Logo	Year	No. of active users/month	Salient features
Facebook		2004	2.32 billion	Supports text, images, videos, live videos, and stories; requires a valid email with an age of being 13 and older.
Twitter		2006	330 million	Registered users can post, like, and retweet but unregistered users can only read them; multilingual platform, freeware
WhatsApp		2009	1.6 billion	Voice-over IP (VoIP) and messaging service owned by Facebook; supports text, audio, video, images; freeware
Skype		2003	300 million	telecommunications application supports video chat, voice calls, instant messaging, text, audio, video, images and video conferencing
Facebook Messenger		2011	1.3 billion	messaging app and platform; exchange messages, photos, videos, stickers, audio, files, voice and video calling
Snapchat		2011	287 million	Photo and short video sharing platform; messages and pictures are accessible for a short time only after that, they become unavailable to their recipients.
You tube		2005	1.9 billion	video-sharing platform owned by Google; allows users to view, upload, add to playlists, rate, report, share, subscribe to other users and comment on videos
Tumblr		2007	642 million	Supports blogs containing multimedia and short messages.
Instagram		2010	1 billion	video and photo-sharing service owned by Facebook
Viber		2010	260 million	cross-platform, voice over IP and instant messaging service operated by a Japanese company; available in more than 30 languages
Sina Weibo		2009	462 million	biggest social media platforms in China; huge financial success with high revenue, surging stocks, total earnings per quarter and lucrative advertising sales; hybrid mix of Twitter's and Facebook's features.
Pinterest		2010	291 million	photo sharing and visual bookmarking platform; enables users to find new ideas for projects and save them; used as a "catalogue of ideas"
LinkedIn		2003	303 million	business and employment-oriented service used for professional networking that operates via websites and mobile apps
Quora		2010	190 million	a place where people can gain knowledge and share by asking and answering questions
WeChat		2011	1.098 billion	Chinese all-in-one communications app for multi-purpose messaging and calling developed by Tencent
QZone		2005	532 million	platform supports photo sharing, music, videos, writing blogs and maintaining diaries; based in China, created by Tencent

Baidu Tieba		2003	300 million	largest Chinese communication platform developed by Chinese search engine company Baidu, available in 3 languages Chinese, Vietnamese, Japanese
QQ		1999	807 million	supports microblogging, shopping, games, movies, music, and voice chat; instant messaging service
TikTok		2016	500 million	known as Douyin in China; supports customizable music videos of up to 60 seconds of length with user-generated special effects and music; Declared world's most downloaded app in the first quarter of 2018.
Reddit		2005	330 million	American social news aggregation, discussion and web content rating website; Links, text posts, images and other contents submitted by registered users can be voted up and down by other members.

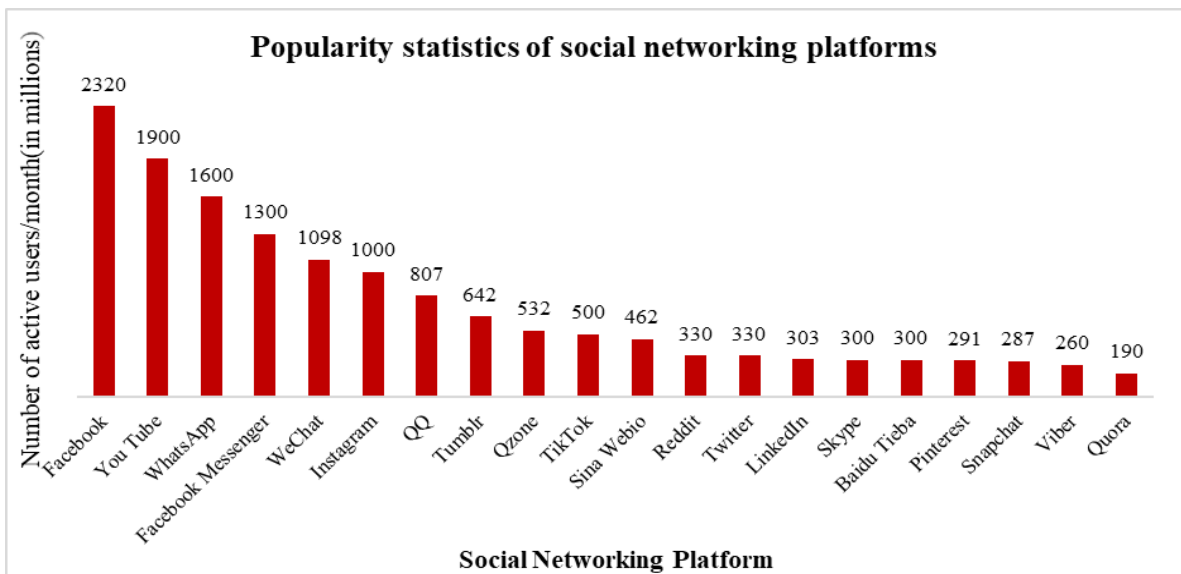


Figure 1.3: Number of active users/month of popular social networking platforms (data source [8])

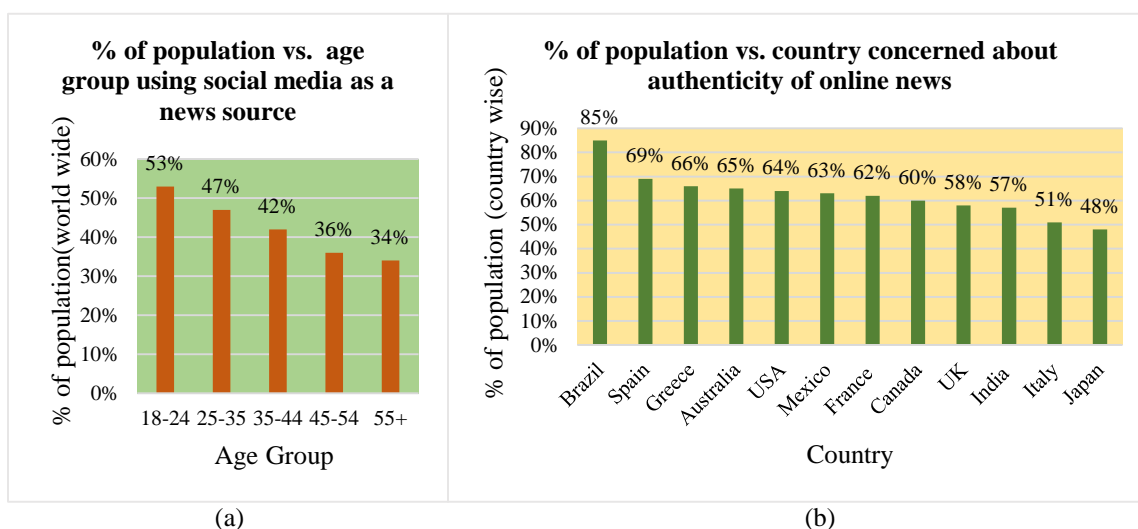


Figure 1.4: (a) Social media as a news source according to age group (b) Awareness of people towards news truthfulness (data source [8])

Table 1.4: Few Examples of false information

Fake Information/News	Classification	Truth and Impact	Year	Ref.
Radiation leakage in Japan could pollute seawater and sea salt, so additionally iodized salt could help to protect people from nuclear radiation	Rumour Fig. 1.5 (a)	Caused salt-buying frenzy in China; shopkeepers charged 10 times higher than average prices; Beijing supermarkets run out of salt	2011	[24]
Breaking: Two Explosions in the White House and Barack Obama is injured	Fake news/ Disinformation Fig. 1.5 (b)	The news was announced from the hacked Twitter account of the Associated Press; before the news was clarified costs 10 billion USD losses	2013	[25]
shootouts and kidnapping by drug gangs happening near schools in Veracruz	Rumour	Rumour triggers severe chaos in the city resulting in 26 car crashes; spread through Facebook and Twitter, as people left their cars in the middle of a street and rushed to pick up their children from school	2015	[26]
Six hundred murders take place in Chicago during the second weekend of August 2018.	Disinformation/ Fake news	The actual number of murders was one; Created fear and anxiety in society	2018	[27]
Donald Trump ends school shootings by banning schools	Satire Fig. 1.5 (c)	An article published by a satire website spread as a breaking news	2018	[28]
Newly appointed Madhya Pradesh Chief Minister Kamal Nath was former Prime Minister Rajiv Gandhi's driver	Misinformation Fig. 1.5 (d)	Kamal Nath had shared an image on Rajiv Gandhi's birth anniversary, from his official Twitter handle in	2018	[29]

		which he is driving the car, and Rajiv Gandhi is sitting by his side		
North Korea Opening its doors to Christians	Rumour	A bogus story published in a notorious fake news web site claimed without evidence. The Magazine aims at spreading the good news to devout Christian readers	2018	[30]
Don't have Paracetamol tablets, it contains the 'Machupo' virus!	Hoax	The Machupo virus, which spreads through direct contact with infected rodents, is only known to be found in South America; no cases have been reported in India so far.	2019	[31]
“Recall these fantastic, mind-boggling photographs of how Bin Laden was hosted in the White House,” Russia’s Foreign Ministry spokeswoman Maria Zakharova has commented on the photograph showing Osama Bin Laden was hosted in the White House.	Propaganda Fig. 1.5 (e)	Osama Bin Laden’s photograph has been superimposed on a photo of Mrs. Clinton meeting musician Shubhashish Mukherjee at an event in 2004. This fake image is shared on social media in Russia.	2017	[32]



(a) (b) (c) (d) (e)

Figure 1.5: Images Spreading Fake News on different Social and News Media Platforms ((a)- [24], (b)- [25],(c)- [28],(d)- [29],(e)- [32])

1.6 User Perception

Users perceive the data from social networks based on their intelligence and consciousness about the facts. According to their interests and insight, they can either forward the data assuming as true, discard it assuming as false or become neutral to the news [33]. A survey study supported by questionnaires done in 2017 by Afira et al. [34] suggests that users judge the credibility of information available online on certain factors such as a link to other sources, interest in the topic, embedded videos, embedded photos, source of information, writing style, logical explanation, peer comments, similarity with different contents and media, etc. Social media analytics tools are the principal source of monitoring,

analyzing and managing information floating on social networks in the public domain. They statistically, behaviourally and semantically analyze the data from different aspects to generate reports. Table 1.5 lists some public and commercial social media analytics tools that play a crucial role in providing suggestions and developing mass opinions.

Table 1.5: Public and Commercial Social Media Analytics Tools

Analytics Tool	Category/Function	Data source	Salient Features	Year	Ref.
Crowdboost	Analytics, Marketing, Management	Twitter, Facebook, LinkedIn	Trip adviser/shopping/online city cabs/can Schedule Unlimited Tweets and Posts, Follower Evaluation	2010	[35]
Vox Civitas	Analytics, Automatic content analysis	Twitter	Journalistic Inquiry to study public opinions after an event	2010	[36]
Whisper	Visualization, Tracing information diffusion process	Social Networks (Twitter etc.)	Visualize social-spatial extent, temporal trends, and community response to a topic	2012	[37]
Talkwalker	Analytics, marketing	Social networks, blogs, news websites	Analyze real-time conversations across social network blogs, news websites, and forums in 187 languages. It provides a wide range of data statistics related to mentions, sentiment, distribution of conversations, etc.	2009	[38]
Google analytics	Web analytics service	all social networks	tracks and reports website traffic, users activities such as session duration, pages per session, bounce rate, etc, have a real-time insight of visitors currently on the website;	2005	[39]
Hootsuite	social media management, listening, publishing, and analytics	Twitter, Facebook, Instagram, LinkedIn, Google + and YouTube.	Improve the effectiveness of ads and broadens the reach of posts; Customize reports in multiple matrices and formats; Track brand mentions better by integrating with specialized tools like Brandwatch and Talkwalker.	2008	[40]
Analytics	Optimize story-based content	Snapchat and Instagram	Create and manage stories with feature-rich publishing; provides different matrices of story popularity and reading	2015	[41]

Internet is a major hub for knowledge seekers, but out of the available information which is credible for learners is a question that needs careful attention. A recommendation framework is proposed [42] for online learning communities by merging user credibility network, domain experts group and user rating matrix, which is based on expertise, influence, longevity and centrality of individuals. This framework provides three categories of

recommendations: learning peer recommendations, domain expert’s recommendations, and learning resource recommendations. Vox Civitas [36] is a social media visual analytics web-based tool developed in 2010 for journalistic inquiry of public sentiments and opinions based on vast message exchange on Twitter. The tool exhibits temporal behaviour by collecting the contents of social media over a specific time window to perform their content analysis based on four factors: relevance, uniqueness, sentiment (positive, negative, controversial and neutral) and keywords (ranked by their TF-IDF scores) to cover the follow-up story angles of certain key events. Whisper [37] is a real-time tool that tracks the information diffusion process in social media and answers when, where and how an idea is propagated. To trace multiple pathways of community response, information propagation, social-spatial extent and temporal trends, an efficient flux line-drawing method is used.

1.7 Current State of Fact-Checking

Compromised social network accounts can be used for spreading misinformation, tarnish the reputation of opponents or they can cause multi-billion-dollar monetary losses in financial markets. Table 1.6 lists popular credibility analysis tools that are used to check the authenticity of online content. Credfinder [43] is a chrome extension developed and launched in 2016 for assessing real-time credibility of tweeter messages based on content and user-specific features. This extension has two major components: A chrome extension (client) that captures the real-time data from the tweeter timeline and a web-based backend (server) that analyses the collected tweets and calculates their credibility. Response time of credfinder is very less and it was extensively tested during 2016 US presidential elections but not as popular as it has no provision to check the images for forgery.

Table 1.6: List of Fact-Checking Platforms

Name	Year	Salient Features	Ref.
TwitterTrails	2014	An interactive online tool for investigating the propagation characteristics, refutation of stories shared on Twitter, origin, and trustworthiness	[44], [45]
TweetCred	2012	A real-time web-based system with a rating between ‘1 to 7’ to assess the credibility of each tweet in the twitter timeline.	[46]
Hoaxy	2016	A platform for collection, detection and analysis of online misinformation and its related fact-checking efforts.	[47]

Emergent	2015	Web-based automatic real-time rumor tracker; tracks social media mentions of URLs associated rumors.	[48]
CredFinder	2016	Analyses user and content features to find out the credibility of tweets. Works in real-time as an extension of the Chrome Browser.	[43]
RumorLens	2014	A tool to aid journalists in segregating posts that spread a specific rumor on Twitter, by traversing the size and distribution of the audience.	[49], [50]
COMPA	2017	System to detect compromised social network accounts. Message characteristics and behavioral user profiles are used for misinformation detection.	[51]
FluxFlow	2014	Interactive visual analysis system to detect, explore and interpret anomalous conversational threads in twitter	[52]
REVEAL	2014	Verification of social media content mainly concentrates on image authenticity from a journalistic and enterprise outlook.	[53]
InVID	2017	The platform supports authentication, fraud detection, reliability and accuracy checking of newsworthy video content and files spread via social media	[54]
ClaimBuster	2015	Allows users to perform live fact-checking with the help of finding out factual claims	[55]
TruthOrFiction	1999	Covers Politics, religion, nature, aviation, food, medical, etc., Email rumors are classified in truth and Fiction	[56]
Snopes	1994	Covers all domains of the news; label videos and News articles in 12 categories, True; Mostly true; Mixture; Mostly false; False; Unproven; Outdated; Misp captioned; Correct attribution; Misattributed; Scam; Legend	[57]
FactCheck	2003	Intends to reduce the level of confusion and deception in U.S. politics. Analyses TV ads, debates, speeches, interviews and news and labels them as True; No evidence; False	[58]
PolitiFact	2007	Covers American politics; After fact-checking labels articles as True, Mostly True, Half True, Mostly False, False and Pants on fire	[59]
Fake News Tracker	2019	Predicting fake news from data collected automatically from social context and news, also provides effective visualization facilities using NLP and deep neural networks	[60]

Hoaxy [47] is a platform for collection, detection and analysis of fraudulent online content from various viewpoints and its related fact-checking efforts. The collected contents from news websites and social media are fed into a database that is updated on a regular basis and analyzed to extract different hidden patterns. The fact-checking activities initiate on social media almost 10-12 hours after the spread of misinformation. Hoaxy is tested by collecting approximately 1,442,295 tweets and articles from 249,659 different users. Because of the limited character length of tweets, URLs of web pages are commonly shared.

COMPA [51] works by building a behavioral profile for every social network account based on message characteristics and stable habits that a user develops over time. Every new message is compared against the already built profile; if it profoundly deviates

from the learned behavior, it is triggered as a possible compromise. However, if the attacker is well aware of the capabilities of COMPA the fake message can be designed in such a way that its behavior resembles the actual one, so it can't be detected. The Flux flow [52] is an interactive visual analysis system designed for detecting, exploring and interpreting anomalous conversational threads on Twitter. It incorporates three major components: (a) data pre-processing and storage module (b) backend data analysis module (c) anomaly detection module. Flux flow represents different dimensions of information propagation such as content, topics, temporal dynamics of the spreading, sentiment, relationship and connections among different threads as well as authors.

1.8 Motivation behind the Work

Social media is a high-speed data generating and disseminating platform. Every second, millions of users are interacting on web platforms creating huge volumes of data. But contrary to traditional news sources as news channels and newspapers, the credibility of content circulating on social media platforms is questionable. Whereas the percentage of people getting dependent on social and web platforms for news and knowledge increases day by day. Social media content governs people's choices of preferences. The term "Fake news" has become widespread after the "2016 US presidential elections" where it is assumed that the fraudulent contents circulated during the elections exert considerable effects on the election results.

We got motivated to explore the current scenario of information pollution on web in terms of ecosystem, different data sharing and generating platforms, data analytics and fact-checking tools. Our literature review focuses on the four different stages of information pollution: origin, propagation, detection and intervention. We tried to highlight the technological solutions provided by various researchers that are currently available to cope up with this burning issue.

As an evolving challenge in the arena of Artificial Intelligence, fake news detection requires huge labelled data to build supervised learning-based detection models. The methods used for annotating scraped web articles include online crowdsourcing platforms such as Amazon Mechanical Turk, Kaggle etc. However, labelling the news articles scraped

from the web is too expensive and necessitates enormous human labour due to the massive volume of data. Furthermore, crowdsourcing annotations are impossible to accomplish in near real-time as well as creates lots of annotation inconsistency for extensive data labeling. With the increasing data size, the data labelling inconsistencies will be worse. Therefore, it is obligatory to design well-organized and practical frameworks for fake news classification, that is supposed to be trained on a small amount of labelled data yet have the capability to explore the hidden patterns of recent unlabelled data as well. We can use semi-supervised learning techniques when we have both labelled and unlabelled data. In that setting, unlabelled data can be used to improve model performance and generalization. Labelled data is a scarce resource. The whole labelling process is costly and needs active monitoring to avoid assessment flows.

Traditionally news articles have a standard format of describing the incident in text and only a few front-page top stories are supported with images. This scenario has been changed with the advent of online versions of news websites and social media handles as visual data attracts viewers more quickly than words do. The human brain rapidly captures and analyses visual signals just by a glance retaining a long-lasting impact in memory compared to text data. So, most of the news stories on social media are supported with visual signals and it's comparatively easy to manipulate them with the help of available photo editing tools.

It is a common practice that instead of going through long and seemingly boring textual contents in detail, users just grasp the supporting image and have a notion about the event which may be wrong if the image is misleading, out of context or manipulated. Given the above facts, it has become imperative to consider visual signals for fake news detection along with textual patterns and styles of writing. A considerable amount of research is being performed on classification techniques for textual fake news detection [61], [62] while frameworks dedicated to visual fake news detection are very few [63]. If textual and visual factors are taken collectively, fake news detection methods have proved to provide higher accuracies than unimodal detection methods.

1.9 Problem Statement

Based on the challenges faced and the motivations encountered during the literature survey, we formulated the following problem statements to be solved in our work:

- To analyse the complete ecosystem of information pollution on web platforms from different social and technological perspectives to figure out the state-of-the-art, current challenges and possible future research areas.
- To design an efficient fake news detection framework that can save time, labour, cost and inconsistencies involved during data annotation process of supervised techniques and to connect with the benefits of both labelled as well as unlabelled up-to-the-minute untapped data.
- To identify and address the problem of fraudulent content in text and image multi-modal data format for veracity analysis of web information contents.
- To design a holistic system capable enough to address wide-ranging formats of fake text and multimedia information content prevalent on web platforms.
- To validate the authenticity of our proposed framework's outcomes by testing and comparing them against the state-of-the-art in the domain on multiple publicly available broadly accepted datasets.

1.10 Major Contributions of Thesis

Major contributions of the thesis are highlighted as under:

- The work puts forward a serious concern towards the burning issue of trustworthiness and reliability of web content on social media platforms. It also establishes the significance of fact-checking and credibility analysis in the current scenario of internet-based information broadcasting.
- The fraudulent content of all varieties scattered online is categorized, and the fake information ecosystem is analyzed right from creation to disposition.
- A thorough analysis of different veracity analysis methods including source identification, propagation pattern analysis, network structure analysis, unimodal and multi-modal approaches, feature-based methods, supervised, semi-supervised and

unsupervised procedures of fake news and rumor detection are presented, which includes their merits and demerits.

- We structure a novel semi-supervised temporal ensembling based convolutional neural network architecture being trained with a limited amount of annotated corpus which leverages the concept of self ensembling for fake news classification in text news articles.
- ConvNet filters are separately applied on headline and body part of the news articles and then extracted feature vectors are concatenated to take advantage of both the slices. The training of proposed neural network using varied proportions of unlabelled and labelled samples for three different datasets delivers the holistic performance analysis and accuracy trends under different circumstances.
- The ConvNet semi-supervised framework via self ensembling for fake news classification, utilizes the semantics of labelled data, at the same time learn new hidden patterns of unlabelled data as well. Self-ensembling via temporal ensembling is a cost-effective but powerful way to squeeze more performance out of a convolutional neural network, irrespective of whether the samples are annotated or not.
- The second semi-supervised fake news detection framework is based on GCN (Graph Convolutional Networks). Graph Convolutional Network can harness the best advantage of convolutions as well as data structuring capabilities of graphs, capable of representing structural and contextual dependencies between its nodes (documents) to draw meaningful insights out of complex data and associated parameters.
- A novel multimodal framework is proposed to incorporate holistic fake news detection of all the modalities (text and images) and different forgery formats (Fake writing style of the text, images with wrong context and doctored images).
- The aim of the Hierarchical Attention Network deep model is to learn implicit fake patterns of writing style in news text. Along with this, different visual manipulations in the form of photoshopped and digitally altered images as well as clickbait are also intended to detect. The final goal is to aggregate the multi-stream detection architecture into a single binary classification system.

- Transformer-based approaches BERT & ALBERT architecture are solely based on self-attention also called intra attention mechanism dispensing with recurrence and convolutions completely. Attention mechanism without recurrence and convolutions allows to draw global dependencies between input and output which removes all restrictions of sequential processing and motivates more of exploring the benefits of parallelization.
- The main contribution of transformer models is that they are non-sequential, meaning that they don't require that the input sequence be processed in the order. The transformer supports multiple folds of parallelization and can achieve a new benchmark in terms of quality and performance.
- Inception-ResNet-v2 pre-trained on ImageNet is being further fine-tuned for our veracity analysis task to harness the advantages of transfer learning which is more suitable for smaller datasets and saves a lot of time for training the model from scratch.
- Contribution of using transfer learning is that the pre-trained models can be scaled for a variety of application specific tasks just by adding few final layers and fine tuning the weights as well as adjusting the hyperparameters on a comparatively smaller labelled dataset.
- Our research is also crucial in this respect as it helps researchers open up with potential future scope as well as different machine learning, deep learning technologies and publicly available datasets.
- This work is expected to prove a landmark within the province of veracity analysis and encourage other researchers to propose further advancements to counter multiple visual and textual forgery formats that would take the scope of the solution to new heights.

1.11 Thesis Overview

This thesis is organized into five chapters. The brief outlines are given below:

- **Chapter 1:** This chapter introduces the background of the information pollution, different flavours of false information, factors motivating its spread, social impact,

user perception and current state of fact checking. Furthermore, research problem statements, significant contributions, motivations and significance of the study are discussed.

- **Chapter 2:** This chapter explains the merits and demerits of existing state-of-the-art methods. The section is dedicated to the literature review, where the existing state-of-the-arts of fake news detection and veracity analysis are reviewed. A thorough analysis of different veracity analysis methods including source identification, propagation pattern analysis, network structure analysis, unimodal and multi-modal approaches, feature-based methods, supervised, semi-supervised and unsupervised procedures of fake news and rumor detection are presented, which includes their merits and demerits. The prevalent approaches in each categories are studied and highlighted in detail, which helps identify the research gaps in this area. Finally, the research objectives are also briefly addressed.
- **Chapter 3:** This chapter details the models developed for semi-supervised textual fake news classification frameworks. The initial model proposes an innovative Convolutional Neural Network semi-supervised framework built on the self-ensembling concept to take leverage of the linguistic and stylometric information of annotated news articles, at the same time explore the hidden patterns in unlabelled data as well. Next, we aim to design a semi-supervised fake news detection technique based on GCN (Graph Convolutional Networks). The recommended architecture comprises of three basic components: collecting word embeddings from the news articles in datasets utilising GloVe, building similarity graph using Word Mover's Distance (WMD) and finally applying Graph Convolution Network (GCN) for binary classification of news articles in semi-supervised paradigm.
- **Chapter 4:** This chapter presents supervised multimodal veracity analysis frameworks. The first model consists of four independent parallel streams that are capable enough in detecting specific forgery formats. All four streams are applied to each input instance. Hierarchical Attention Network deals with headline and body part; Image captioning and headline matching module require all the three parts headline, body and image; Noise Variance Inconsistency and Error Level Analysis

focuses only on images accompanied with news text. These independent predictions are finally combined using the max voting ensemble method. The second model aims Inception-ResNet-v2 to extract visual features. The techniques BERT and ALBERT have been used to elicit textual attributes. Diverse forms of text input, like English articles, Chinese articles and Tweets have been used to make our model robust and usable across multiple platforms. The architecture of Multimodal Early Fusion and Late Fusion has also been experimented and analyzed in detail by applying on different datasets.

- **Chapter 5:** This chapter provides a summary of proposed works, significant findings, contributions and limitations. In this chapter, we also suggest some future directions.

Chapter 2

Literature Review

This chapter highlights the merits and demerits of existing state-of-the-art methods. We have reviewed various fake news detection approaches ranging from rule-based methods to machine learning and deep learning-based solutions. It helps us to discover the research gaps in existing solutions in the relevant area. We also highlighted containment and intervention methods available in the literature. Further, the research objectives are formulated based on these research gaps, which are addressed in this thesis.

2.1 Introduction

Nowadays, news publication, propagation, and consumption have diverted to online social media networks and web portals, which has given rise to falsified and fabricated news articles containing both textual and visual information formats. Social media's growing popularity accelerates the spread of fake news. The transition of mainstream print media to electronic media and now to social media has posed a considerable challenge in front of the Journalists because the “*WhatsApp University*” is producing huge volumes of unverified and unauthenticated news content based on perceptions rather than facts. The credibility and trustworthiness of media as the *fourth pillar of democracy* is also under crisis. To attract and deceive readers for rapid distribution of falsified information content, multimedia technology specifically tampered or sometimes irrelevant and out of context attracting images are extensively used. Two key reasons are proposed [64] to describe the escalation of misinformation websites: 1) a pecuniary one, substantial advertising revenue generation by viral news articles, and 2) a more political one, influence of public opinion by fake news articles on specific topics.

The power of social media relies on when it comes to reporting breaking news in real-time. It was on 15 Jan 2009, [65] when first-time social media was used to solve the

crisis. The plane U.S. Airways Flight 1549 crashed and had an emergency landing in the frigid waters of the Hudson River in New York. Janis Kurma who was on a ferry at that time near the incident site, tweeted with a photo of the crashed plane to his 170 followers **“There’s a plane in the Hudson. I’m on the ferry going to pick up the people. Crazy.”** This tweet helped in saving the lives of onboard 155 passengers and crew members within minutes. At that time Twitter, for many, was just a funny word in the dictionary but this incident proved as a miraculous turning point, helped this platform to become the powerhouse of social media and now it’s as ubiquitous as a bird in the sky. This event became so popular that the movie *“Sully: Miracle on the Hudson”* was also picturized on it in 2016 [66] as it was the first attempt of utilizing social media for crisis management situations. Authors in [67] and [68] analysed the importance of social media posts in managing disasters and emergencies.

2.2 Veracity Analysis Approaches

Information circulating online on web platforms has various attributes like headline/title, text/body, author, associated image, URL, etc. Any changes in these features incline news to behave abnormally, making it a piece of fake news or a rumour. Researchers have tried to verify different attributes, webpage URL, author, text, associated image, time series analysis and propagation statistics to design models for veracity analysis. False contents, also popularly termed as information pollution or infodemic, are amplifying at an alarming rate. Efforts have already started for providing solutions to detect and mitigate fraudulent content using up-to-date artificial intelligence technologies. A considerable number of methodologies embrace fake news uncovering textual data. With the rise of multimedia data on users' posts and news, research incorporating the identification of visual fake news has increased and improved the preciseness of algorithms.

During the COVID-19 pandemic, the circulation of fake news or rumors is being highly observed. In an article titled **"Fake News, Real Arrests"** [69], the author categorically emphasized the increase in the number of arrests by police due to the spread of fake news across different states of India. It has also been quoted that **"Virus of fake news spreads faster than Corona Virus itself"**. Social media is the most accessible tool to spread fake

news rapidly and bots are being programmed to make the task easier. Existing literature enumerates several criteria for classifying the veracity analysis approaches. We categorize various methods available in literature for fake news detection and veracity analysis in the following twelve sections.

2.2.1 Source Identification

Source detection refers to find out a person or location from where the fraudulent information in the social network or web started spreading. Along with other containment methods, identifying the original source of information pollution plays a vital role in reducing online misinformation. In various application domains, origin identification is very important such as Medicine (to find the source of the epidemic), Security (to detect the source of the virus), social network (to identify the origin of the wrong information), financial networks (for finding the reasons of cascade failures), etc. The following Figure 2.1 summarizes the steps involved in the source detection process.

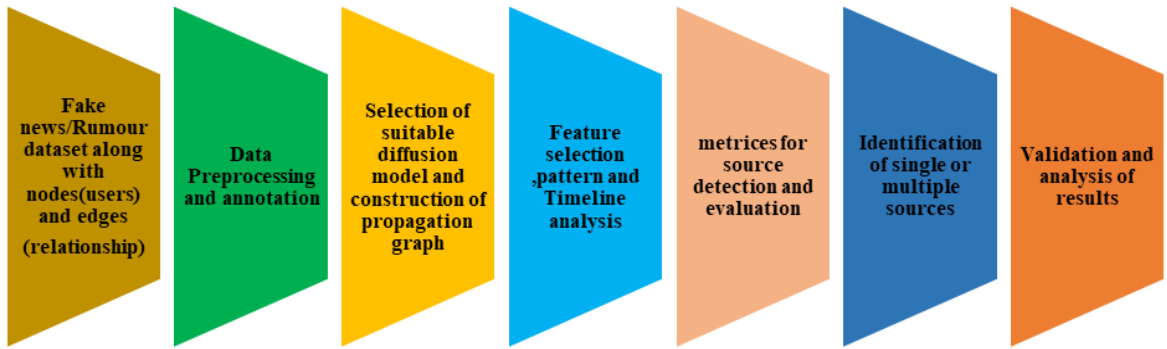


Figure 2.1: Steps of source identification of false information

A bio-inspired method which solely depends upon the infected time of observers was developed in [70], proposes a Physarum-inspired Mathematical Model of misinformation source detection under the constraint of limited observers and SI model of the diffusion process. The model gives higher locating accuracy and less error rate when compared to experimental results of four benchmark networks with traditional Gaussian and GaussianSI model. Shelke and Attar [6] provides a state-of-the-art survey of different source detection methodologies in case of single and multiple misinformation source along with different available datasets and experimental setups. A two-stage source localization algorithm for probabilistic weighted graph [71] is designed which models the heterogeneity

of social relationships by using probabilistically varying weights for the edges. In the first stage of the algorithm, the most likely candidate cluster to contain the source of the rumor is identified. In the second stage, the source is estimated from the set of nodes inside the most likely candidate cluster. To minimize the estimation error of source and analyze the rumor centrality maximum likelihood estimator [72] is used that examines the asymptotic behavior of infected nodes in detail for regular trees, general trees and general graphs. Along with the infection source the infection region i.e. a subset of nodes infected by each source in a network is identified considering SI propagation model with homogeneous spreading rates based on approximations of the infection sequence count. Choi Jaeyoung et al. [73], [74] identifies rumor source using different approaches such as batch queries, interactive queries, Maximum-A-Posteriori-Estimator (MAPE). Zhu and Ying [75] tries to identify source using a path-based approach and [76] estimates spreading source in network based on observer nodes.

Researchers have formulated various approaches to identify fake accounts on online platforms. In [77], authors have suggested a solution to discriminate fake Twitter accounts from genuine ones using dimensionality reduction and feature selection techniques. The authors in [78] use basic machine learning algorithms to classify fake accounts on the social media platform Instagram. It is not always viable to stop the propagation of fake news by recognizing and blocking the fake accounts because a user can create multiple accounts on an online platform or sometimes real accounts could also be a source of fake news. Louni and Subbalakshmi [79] addressed the problem of finding the origin of rumor in large scale social network. The salient feature of their proposed algorithm is probabilistically varying internode relationship strengths achieved by assigning random non-homogeneous edge weights to the original network graph. Query response method [80] using simple batch querying and interactive querying with directions is also out of the box analytical approach for rumor source detection.

2.2.2 Propagation Pattern Analysis

The majority of the research in the propagation dynamics of misinformation is done in line

with epidemic models, which categorizes the people in different classes then derives equations to perform steady-state analysis. People who never heard the rumor, Ignorant are similar to **Susceptible (S)**, those who are spreading rumors, Spreaders are similar to **Infective (I)** and people who heard rumor but do not spread it, Stiflers similar to **Removed (R)**. The dynamics of rumor spreading on homogeneous network LiveJournal are studied in [81] with consideration of forgetting rate, spreading rate, stifling rate and average degree using SIR (Susceptible-Infected-Removed) epidemiological model. The same group of researchers further extended their work by adding a new category of people **Hibernators (H)**, coming from the spreaders due to forgetting mechanism and later becoming spreaders again due to remembering mechanism in SHIR(Susceptible-Infected-Hibernator-Removed) model [82]. SIDR(Spreader-Ignorant-Doubter-Stifler) model is proposed in [83].

Mean-field equations and steady-state analysis are done to study SHIR rumor diffusion model in social networks. Another model based on users forget and remember mechanism is presented by J. Gu et al. [84] in which an individual's state keeps on switching between active (with the message) and inactive (without message). A nature-inspired approach based on forest fire model is proposed by Indu and Thampi [85] to figure-out the diffusion path of rumors and find out the most influential users in rumor diffusion. The model evaluates the probability of each node to be affected by misinformation and finally identify all the rumor affected nodes to estimate the complete range of rumor spread. The study concluded that only a few users have tweeted the rumour and 90% of the messages are retweets. Marcelo et al. [86] analyzed the propagation dynamics, follower-followees relationship, number of tweets per user, the vocabulary of tweets, retweet behavior for conformed truths and rumors supported by a case study of 2010 earthquake in Chile. The research concluded that false stories are questioned much more than confirmed truths.

A rumour propagation model for emergency situations based on the interactions of seven stakeholders of population ignorant(I), Wise(W), spreader (S), unbeliever (U), indifferent (IN), Opponent(O) and reasonable immune (RI) is proposed using an active immune mechanism [87]. Experiments show that network properties profoundly affect the diffusion process. Rumour propagation analysis on online social site BlogCatalog is done by formalizing a dataset of an undirected graph $G(V,E)$ contains 10312 nodes and 3 33 983

edges using stochastic epidemic model [88]. The complex structure of social networks can be modeled using different graphical formats such as Assortative correlated scale-free networks, Uncorrelated scale-free networks, Homogeneous networks, Inhomogeneous networks and Random Graphs. Analysis of rumor diffusion in complex structures is done by using the stochastic model [89], which are further analyzed by analytical and numerical solutions of mean-field equations.

A content-based probabilistic model [90] utilized four properties of rumor propagation temporal, structural, linguistic and social tie for identification of unverified tweets in the aftermath of a disaster in an early stage. The salient feature of the approach is a tweet that has at least one rumor propagation feature is being extracted, and its probability of being a rumor is analyzed. Another key finding of the method is that rumours contain high sentiments, generally dominated by words related to social ties and actions like hearsay.

The way rumour and fake news diffuse in a network is also a prominent feature of analysis and mitigation of misinformation. Non-sequential propagation structure of microblog posts is being identified by Ma et al. [91] to acquire discriminative attributes for generating powerful top-down and bottom-up representations using Recursive Neural Networks for rumour identification. They also proposed a kernel-based method of rumour detection termed as propagation tree kernel [92] to capture higher-order features for segregating various types of rumours based on the structure of their dissemination trees. Vu and Jung [93] proposed a propagation graph embedding method for rumour detection based on Graph Convolutional Neural Network. Experimental results illustrate the efficacy of their suggested method by reducing the detection error up to 10% as equated with state-of-the-arts. Other studies of combining attention mechanism with propagation graph structure [94], an ensemble of user representation learning with news propagation dynamics [95], implications of searchability on rumour self-correction [96] are also worth considering.

2.2.3 Network Structure Analysis

Network structures are innovative methods of credibility assessment of a target article [97], [98]. A model is being constructed in Dynamic Relational Networks [99] by using related news articles that are mutually evaluating each other's credibility based on the facts of who,

what, where, when, why and how. Each article unit contains one article node and many fact nodes. Nodes of one article unit are mutually evaluated by consistency among their fact nodes with another available article. For fairness of evaluation, each user can build his network by using a bottom-up approach. Structure of small world peer-to-peer social networks [100] and large web-based social networks spanning large geographical areas [101] are analyzed through various modeling techniques to deduce some important characteristics of propagation and area related properties. In the case of small-world network, the connectivity between users is scale-free in the form of undirected, directed and weighted graphs. Figure 2.2 represents some of the network structures being constructed for credibility assessment.

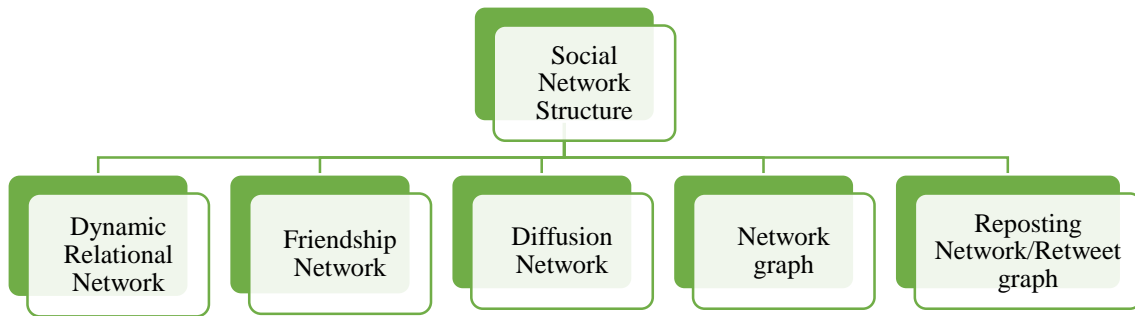


Figure 2.2: Different Network structures used in credibility assessment methods

To model network structures and user connectivity of online social networks, scalable synthetic graph generators are used. They provide a wide variety of generative graph models that can be used by researchers to generate graphs based on different extracted features such as propagation, temporal, connectivity, follower-followee, etc. Some of the tools and their characteristics are summarized in Table 2.1.

Table 2.1: Scalable synthetic social network graph generators

Synthetic Graph Generator	Salient Features	Ref.
Darwini	Can be used efficiently to study propagation and detection of false contents by means of generating different social connections in the form of a graph for which darwini can produce local clustering coefficient, degree distributions, node page rank, eigenvalues and many other matrices.	[102]
DataSynth	Scalable synthetic graph generator with customizable schemas and properties. Introduces novel features of representing the correlation between the structure of a graph and properties.	[103]
BTER	Capture clustering coefficient and degree distribution, useful in reproducing graphs with massive community structure network.	[104]

Myriad	A toolkit for expressive data generator programs can generate nodes and edges data for visualizing and experimenting online social network connections. The naive feature is that can be executed in a massively parallel manner.	[105]
R-MAT	“recursive matrix “a simple, parsimonious graph model that can quickly generate realistically weighted, directed and bipartite graphs. Diverse real social network and web connectivity graphs can be well approximated by an R-MAT model with appropriate choice of parameters.	[106]
LFR	Graph generator used to evaluate community detection algorithms. Capable of clustering large graphs that exceed main memory using external memory.	[107]
gMark	schema-driven, domain-independent, highly configurable and extensible graph instance, and query workload generator. Practical usability has increased many folds with its customizable schemas for varied application domains.	[108]
Apache Spark framework	Basic properties of power-law distribution of the number of user communities, the dense intersections of social networks, and others are used to generate a graph similar in structure to existing social networks. A very small amount of resources and faster execution speed in comparison with other similar generators.	[109]
Attributes Synthetic Generator (ASG)	Consider feature similarity and label homophily among individuals when forming links in the network. To tune the social network parameters exactly to the generated network particle, swarm optimization is used. shared similarity among individuals to form the links in the network. Statistics taken from real OSNs are used to form the nodes attributes. Time efficient and require only limited parameter optimization.	[110]
Multi-Link Generator (MLG)	Follows the preferential attachment model for handling multiple networks that contain different link types. The model starts with few nodes and as networks grow, more nodes and links are added to the model. MLG is scalable and efficient in time and parameter optimization.	[110]

2.2.4 Text Based Approaches

Many researchers have exploited features from textual contents like textual frequency-based features, semantic features, sentiment analysis, writing style, text pattern, the polarity of contents etc. Pan et al. [111] have studied how the style of writing is responsible for shaping the views of the people. They used a unique structure called knowledge graphs to categorize given news text as fake or real. Gautam and Jerripothula [112] put forward a novel framework for text fake news detection using Random Forest classifier that leverages the capabilities of paraphrasing tool Spinbot, grammar checking tool Grammarly and Glove word-embedding for feature extraction. Li and Zhou [113] employ BERT language model for binary text fake news detection by connecting the dots between the claim and pieces of evidence. Lengthy text news articles are first summarized to extract the claims as well as keywords that are subsequently searched on web to extract the related articles to treat them as evidences or facts to be matched against the claim. Finally, the scores of similarities are

fed into a fact verification model to classify a news article as real or fake. To detect ambiguous information on COVID 19 Elhadad et al. [114] collected factual data from WHO, UNICEF, United Nations and epidemiological data from different fact check websites to consolidate into a repository. This accumulated factual data is used to build a fake news classification system based on ensemble of ten different machine learning classifiers using seven feature extraction techniques implemented and tested against twelve performance metrics.

2.2.5 Visual Approaches

Images are an integral part of print, electronic and digital media news. Visual snapshots alone or mostly combined with text overwritten when manipulated spread forgery to a great extent. A novel method is proposed [115] to detect photographic splicing detection via illumination inconsistencies and deep learning. This approach is capable enough of locating forged regions, eliminates the laborious feature engineering process and provides outperforming accuracies on the same datasets paralleled to state-of-the-arts. Elkasrawi et al. [116] devised a semi-automatic approach of verifying the credibility of images in online articles by means of a two phased process embraces clustering and hierarchical image feature analysis, accounts to highest 88% accuracy for a dataset of 50 images. Fake images generated by Generative Adversarial Networks (GANs) by changing context and semantics of original image via. Image-to-image translation looks very realistic. The study proposed by Marra et al. [117] performed on a dataset of 36302 images describes the performance of multiple image forgery detectors both in standard conditions and under in the influence of compression for GAN generated fake images. Jin et al. [118] designed a novel method based on visual and statistical image features implemented on real-world Sina Weibo dataset for microblog news verification.

2.2.6 Multi-modal Approaches

The modalities of text and image, when used in conjunction, mislead people by spreading fake news enormously. Although there are not many standard open-source datasets available for multi-modal fake news classification, having text and images, but few researchers have utilized the rarely available datasets for the task. Yang et al. [119] used explicit and

implicit features from text and images using Convolutional Neural Networks (CNNs) for counterfeit news detection. Wang et al. [120] have created an end-to-end model that uses an event discriminator to classify fake news exploiting modalities of text as well as image. Their model uses CNN for text and VGG-19 to elicit photographic features and concatenate them to discriminate events and classify false news. Inspired by EANN, Khattar et al. [121] proposed a comparable framework using bi-LSTM instead of Text-CNN and formulated an architecture of auto-encoder and decoder. Same latent vectors were used for both encoder and decoder. Along with textual sentiment analysis and image segmentation process, Shah and Kobti [122] implemented cultural algorithm with situational and normative knowledge for multi-modal feature extraction. The final feature vector is passed through SVM classifier layer for veracity analysis experimented on Weibo and Twitter datasets. A neural network based multi-modal fake news detection system is proposed by Giachanou et al. [123] that combines textual, visual and semantic information being implemented on MediaEval , Politifact and GossipCop datasets. Vishwakarma et al. [124] processed multi-modal data in a different format by extracting the text written on images and then web scrapping this text claim in order to a reality parameter based on top 15 google search results. The value of the calculated reality parameter is compared against a threshold, which if exceeds a set point, the information is categorized as real otherwise fake. Meel and Vishwakarma [125] researched a framework for multi-modal fake news detection based on ensemble of Hierarchical Attention Network, Image Captioning and Forensics.

2.2.7 Feature Based Approaches

Extensive studies [126] , [127] , [16] , [128] have been done for counterfeit news detection using propagation pattern-based features, social features, content-based features and user profile features. Propagation pattern-based features signify the diffusion patterns of fake news and rumours which could be effectively modelled by constructing propagation graphs with nodes and edges. Social features such as the number of likes, shares, comments, retweets etc. signifies the response of the people and the way society has perceived a particular event. Content-based features are extracted from the choice and frequency of particular words in the language used to describe the news and user profile features are specifically

related to the user’s personality, social interactions and account information such as creation time, authenticity, verification etc. The following Table 2.2 details some of the literature work related to feature-based methods and their limitations in the area of fake news detection.

Table 2.2: Feature-Based Methods for Fake News Detection

Reference	Features	Method	Limitation
Propagation Pattern-based Features			
[126]	Propagation Graph is Constructed and analyzed in terms of size, depth, maximum breadth and structural virality	Considered the diffusion patterns of true and fake news stories on Twitter using statistical methods such as Histogram, Quarterly counts, Complementary Cumulative Distribution Function (CCDF)	Most of the work done so far is based on Supervised learning which is unable to extract the latest real-time spreading patterns. Malicious accounts such as trolls, social bots and cyborg users intensify the spread of fake news providing false propagation statistics.
[129]	Time series analysis of trending news, # nodes, # edges, Network density, network associativity, shortest path of propagation	K-Nearest Neighbor with Dynamic Time Warping (KNN-DTW) algorithm is used for classification of promoted campaigns from real news	
[130]	Propagation initial tweet, Propagation max subtree, max level, max and average degree, max and average depth	Decision Tree is built for classifying tweets as credible or non-credible based on propagation pattern-based features	
Social Features			
[131]	# likes, comments, shares, negative comments, perception, captivation, controversy, min/max/mean/std response distance	Identification of potential misinformation targets using Linear and Logistic Regression, SVM, KNN and Neural Network	People who are part of the same social or cultural group exhibits homogeneous interests, orientation and perceptions. So, its relatively easier to spread propaganda or biased agenda in such communities which are inclined towards their interest. Such social media groups are also vulnerable to echo chamber effects.
[132]	Number and time of reply, retweet, comment, shares	Detection and classification of fake news and fake tweets related to six different events using WEKA tool	
[127]	# likes, retweets, replies, comments	SVM, Logistic Regression, Naïve Bayes and CNN methods are used to classify events and tweets into real or fake	
Content-based Features			
[133]	# first/second/third order pronoun, # positive/negative words, # question/exclamation marks, # words/characters, # URL/@/hashtag	RNN with attention mechanism is used to fuse multiple features for effective rumour detection	Explicit features related to the frequency of occurrence of particular words in writing style are considered

[16]	Uni-gram, bi-gram, tri-gram and four-gram analysis	SVM, LSVM, KNN, DT, SGD, LR supervised machine learning methods with TF and TF-IDF feature extraction are used for fake news detection	but the patterns can extensively be analyzed beforehand to design fraudulent or false information manually or algorithmically that can satisfy all frequency-related explicit features
[134]	# characters / words/uppercase letters / Exclamation marks /user mention/ Hashtag /positive words/Negative words	Random forest, naïve Bayes, decision tree and feature-rank Naïve Bayes machine learning methods are used for credibility assessment	
User Profile Features			
[130]	Registration Age, Follower count, Friend count, Is verified, Has URL, Has Biodata	Decision Tree is built for classifying tweets as credible or non-credible based on user profile features	Huge volumes of information is being shared on online platforms every day that even genuine and intellectual users are unknowingly trusting and sharing false news as it is almost impossible to cross-check the credibility of every news that is highly persuasive in nature.
[128]	Account authenticity, Account registration time, Follower count, Following count, Account age, Tweets count, Retweet count	Forest Fire Nature Inspired method based on user profile features is used for modelling rumour spreading in social networks	
[134]	Follower count, Friend count, User credibility in terms of # replies and retweets, Age, Gender, political orientation, User preferences	Random forest, naïve Bayes, decision tree and feature-rank Naïve Bayes machine learning methods are used for credibility assessment	

Although feature-based methods prove good enough up to a certain limit in identifying fake news; they are primarily based on explicit patterns of diffusion, social behaviour, text writing style and user profiles; so, can easily be analysed in advance and circumvented. Bearing in mind the limitations of the feature-based methods, we tried to strengthen our proposed framework by overcoming the restrictions of supervised technologies and utilizing implicit hidden patterns from the latest up-to-the-minute news articles.

2.2.8 Supervised Methods

Most of the literature is concentrated on building supervised learning models using machine or deep learning algorithms. Ajao et al. [135] proposed a sentiment scoring function “emoratio” using psychological and linguistic capabilities of the Linguistic Inquiry Word Count (LIWC) tool to train supervised models on PHEME dataset for correctly classifying misinformation with text sentiment analysis. A systematic comparative analysis of multiple deep learning methods incorporating long short-term memories (LSTM), convolutional neural networks (CNN), ensemble methods and attention mechanism for text fake news

classification is presented by Kumar et al. [136]. It was concluded that convolutional neural network ensembled with Bidirectional LSTM using attention mechanism gives highest 88.75% accuracy. Circulation of artificially generated fake images and altered images on social platforms is another challenging subfield of fake news classification. The research conducted by Marra et al. [137] on a dataset of 36,302 images provides a solution by using both conventional and deep image forgery methods for detecting fake images generated using image-to-image conversion, built on Generative adversarial networks (GAN) model. Veracity classification using time and attack [138], deep neural network for early fake news detection [139], style analysis of hyperpartisan news [140], Multimodal Fusion with Attention-based Recurrent Neural Network (Att. RNN) [133] are worth mentioning pioneering research in the arena of credibility analysis on social networks. Language model based on Transformers such as Bidirectional Encoder Representations from Transformers (BERT) [141], BAKE and exBAKE [142] which are improvements over BERT and Spot-fake-a multimodal supervised framework based on BERT and VGG19 [143] utilizes the capabilities of encoders and decoders for efficient article classification into real and fake.

2.2.9 Semi-Supervised Methods

To harness the benefits of linguistic, phonological, semantic, stylometric, syntactic capabilities of labelled data and at the same time decipher the patterns hidden in unlabelled data, several semi-supervised techniques are available in the literature to utilize in different applications. Engelen and Hoos [144] provide a detailed systematic study of available semi-supervised technologies including Clustering, Self-training, Co-training, Boosting, Graph-based semi-supervised methods, Perturbation methods, Maximum margin, Generative models etc. which conceptually are situated between supervised and unsupervised learning. Π -model and Temporal Ensembling are two standard semi-supervised benchmarks [145]. These are based on the principle of self-ensembling for setting ensemble targets for unsupervised data and are successfully experimented on SVHN and CIFAR-10 image classification task. Laine and Aila also figured out the fact that Π -model and Temporal Ensembling both provides tolerance against incorrect labels and improves the classification accuracy in the fully labelled case as well. The research conducted by Shi et al. [146] demonstrates Graph Temporal Ensembling (GTE) framework, a slight variant of

temporal ensembling encapsulated in AlexNet backbone architecture for histopathology image analysis with noisy labels successfully validated on lung and breast cancer datasets. A novel semi-supervised method for detecting deceptive and fake opinion reviews is proposed in [147] which is primarily based on co-training and expectation maximization. The feature vector extracted from reviews incorporates four features: Part-of-speech, Linguistic, Word count and Sentiment polarity. To access the credibility of online blogs written in the Arabic language a deep co-learning end-to-end semi-supervised solution is proposed [148]. Another content based semi-supervised misinformation detection method is suggested by Guacho et al. [149] using tensor embeddings. This framework is based on three main steps: representing the article text as tensor-based embeddings, constructing the embeddings graph based on the K-Nearest Neighbour method and finally propagating beliefs using Fast Belief Propagation (FaBP) Network.

2.2.10 Unsupervised Methods

Annotated samples cannot characterize the authenticity of the news on a recently developed event as they become obsolete very quickly due to the dynamic nature of news. To obtain high-quality up-to-the-minute labelled sample is the foremost challenge in supervised and semi-supervised learning methodologies. The motive behind development of unsupervised technologies is to completely eliminate the need of annotation and real time solution to the problem of information trustworthiness. A novel unsupervised method of detecting fake news circulating on WhatsApp is implemented by Gaglani et al. [150] leveraging Transfer Learning techniques. The framework utilizes semantic similarity between claims circulated on WhatsApp and associated articles scrapped from the web to classify the claims as true or false. Hosseinimotlagh and Papalexakis [151] recommended an unsupervised ensemble method that consolidates results from different tensor decompositions into coherent, comprehensible and high precision groups of articles that belong to different categories of false news. Unsupervised microblog rumour detection framework is proposed [152] based on recurrent neural network and autoencoders by analysing the temporal dynamics and crowd wisdom. Yang et al. [153] recommended a generative approach of unsupervised fake news classification on social media by exploiting user's engagement, opinion and their credibility factors. A three phased graph-based approach utilizing biclique identification, graph-

based feature vector learning and label spreading for fake news detection in the absence of annotated data is successfully designed by Gangireddy et al. [154].

2.2.11 Containment and Intervention Methods

Twitter data is extensively used to analyze the rumor spread during and after the Great Japan Earthquake of March 11, 2011 [155], performing a comparative study of disaster and normal situation tweets and spreading patterns. The work concluded with establishing the fact that rumor tweets spread easily, but rumor disaffirmation tweets do not spread more than a few nodes in the network. Anti-rumour news and campaigns are used to alleviate the spreading of rumor. Software developers and technology firms have begun developing human-driven mechanisms as well as tools to identify and quarantine fake news. Mainstream news organizations also constitute teams of fact-checkers and investigating units. Figure 2.3 classifies some of the prominent technologies used to intervene in the spread of malicious content online.

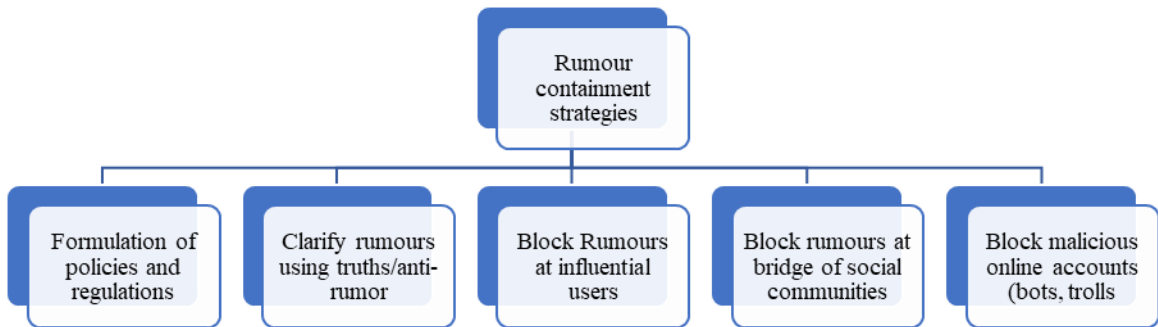


Figure 2.3: Classification of rumor containment strategies

Significant efforts for the mitigation of fraudulent content are done in [156] by identifying a set of highly influential nodes, which are decontaminated first and in turn diffuse the confirmed news in their complete friend network. In a small size, social network GVS (Greedy viral stopper) algorithm is used to find out the set of most influential nodes. If the network structure is very vast, then the community-based heuristic algorithm is used. The highest disadvantage of this method is that it has assumed that facts and misinformation spread with the same rate in a network which proves out to be false in many research studies.

The authors in [157] found that the crowd has the potential to self-correct. Corrections to the misinformation emerge in the social networks themselves but are muted and not widely propagated. In order to mitigate the rumors in Vehicular Social Networks [158], a specially authorized node is introduced in each network which has the responsibility of spreading anti-rumor messages to spread correct information. Blocking rumors at highly influential users and at the community bridges are two main strategies of proactive measures along with the remedial method of spreading truths to mitigate information pollution. A mathematical model [159] based on the categorization of the population into susceptible, defended, recovered (active, immunized), infected (contagious, misled) is introduced to investigate the methods of rumor containment with parameters of degree, betweenness, core, overlapped and separated. By predicting the possible future path of rumor propagation, can try to block it at influential users and bridges of social communities.

Formulation of policies and regulations for contents posted on social media and legal laws for wrongdoers will motivate the users to think rationally before resharing or posting. Social bots, which are social media accounts operated by computer algorithms can give a wrong impression pertaining to the popularity of information and are endorsed by many people that enable the echo chamber effect for the propagation of fake news. Apart from social bots, cyborg users and trolls are also malicious accounts that amplify the spread of fake news must be blocked [160]. Community signals, user's flags and expert opinions [161]leverage the detection as well as minimize the spread of fraudulent information by stopping the propagation paths. "Fake news game" [162] is an educational game that provides key containment strategies to inoculate the public against the risk of fake news.

2.2.12 Other Approaches

Cognitive psychology is a method of analyzing human perceptions. The cognitive process examines four main ingredients coherency of the message, credibility of the source, consistency of message, general acceptability of message using collaborative filtering property of social networks to detect misinformation, disinformation, and propaganda [163]. The proposed genetic framework measures the credibility of the source of information as well as the quality of new ideas on twitter dataset with 90% accuracy. A system Rumour Gauge

[164] is designed to accurately predict the veracity of real-world rumors on Twitter before verification by trusted channels using Hidden Markov Model. However, the system is incapable of differentiating between malicious and accidental misinformation. The stance is the overall position held by a person towards an idea, object or belief. Review of different methods of rumor identification using stance classification in four categories of supporting, denying, Querying and Commenting is presented in [7]. The work is done in various areas of knowledge-based, style-based, propagation-based, user-based and credibility based fake news detection including manual as well as automatic fact-checking in homogeneous, heterogeneous and hierarchical networks are summarized by Zafrani et al. [5]. Kumar and Shah [4] focused on three types of false information opinions based on fake reviews, Fact-based hoaxes and intent-based disinformation.

O'Brien et al. [165] proposed an iterative Graph-based method of credibility classification. Tri-relationship among publisher, news piece and user [166] explores the role of social context for trustworthiness analysis. Shu et al. [167] try to improve fake news detection accuracy by exploring different characteristics of social media user profiles based on experienced and naïve users. Hawkes process [168] is a probabilistic framework of fake news detection. Investigative journalism and wisdom of crowd [169], unsupervised Bayesian network [170], filter out misleading and false websites [171] are some of the other prominent methods of content analysis. Veracity analysis of fake news by scrapping and authenticating the web search is proposed in [172].

Authors in [173] propose a browser extension “BRENDA” implemented with Google Chrome browser for automated credibility assessment of claims using deep neural network. This automatic process is fast, real time, produces the evidence along with the classification and does not need to leave the web page. A Clickbait Video Detector (CVD) is designed by Varshney and Vishwakarma [174] to detect the clickbait videos circulating on web based on cognitive evidences extracted out of user profile, video content and human consensus. To extract the hidden patterns from unlabeled data a semi-supervised approach based on the temporal ensembling method is proposed by Meel and Vishwakarma [175]. Analysis of misinformation detection under varying time constraints under the

consideration of three of different types of attacks: Evasion attacks, Poison attacks, Blocking attacks is thoroughly studied by Horne et al. [138].

2.3 Gaps Identified in the Present Study

An extensive literature survey leads us to figure out the following problems of current research in the field of credibility analysis of online information:

- One of the most challenging parts in providing a solution to fake news malice is labelling massive volumes of data to train supervised artificial intelligence models.
- Annotating vast volumes of data is too expensive in terms of time and cost as well as requires enormous human resources.
- For extensive data labelling, crowdsourcing or human expert annotations creates lots of annotation inconsistencies.
- Supervised technologies cannot achieve feature engineering of the latest online information and near real-time performance.
- Most of the research, to date, is primarily centred on text fake news detection using the traditional machine-learned classifiers. Little work has been reported in the literature using deep learning techniques and image forensics analysis.
- The disadvantage of machine learning is that it requires manual feature extraction and also fails to detect implicit hidden patterns in fake-news-writing style text. Along with this, traditional classifiers fail to give good results in the presence of huge volumes of data.
- A considerable amount of research is being performed on classification techniques for textual fake news detection while frameworks dedicated to visual fake news detection are very few. If textual and visual factors are taken collectively, fake news detection methods have proved to provide higher accuracies than unimodal detection methods.
- Less work is reported in literature to address multi-modal information authenticity issue that will take the benefits of latest language modelling and transfer learning technologies.

2.4 Research Objectives

A robust fake news detection system must be generic, compact, efficient, and straightforward. To overcome the limitations and research gaps discussed earlier, the main objectives of this thesis work are as follows:

- To review the existing state-of-the-art veracity analysis approaches along with their advantages and disadvantages to figure out current challenges and future perspectives in the area of information pollution on web and social platforms.
- To design a semi-supervised Fake News Detection framework that can learn from the semantics of labelled data and intrinsic patterns of unlabelled data to optimize in terms of time, cost, labour and annotation inconsistencies.
- To address the highly circulated forgery format of a news story coupled with headline text along with partially labelled training samples.
- To take advantage of self-learning cognition of designed framework which facilitates the architecture to learn from up-to-the minute latest forgery formats fed into the process in the form of unlabelled articles.
- To design framework aimed to avoid cumbersome feature engineering of supervised learning methods, crowdsourcing or human expert annotation inconsistencies, reduces cost and workforce requirements.
- To design a holistic system that would be capable enough to address wide-ranging diverse combinations of forgery formats of fake text and multimedia information content prevalent on web platforms.
- To propose an algorithm for the ensemble of different methods for higher accuracy in fake news detection.
- To evaluate the performance of the proposed fake news analysis frameworks on publicly available news datasets.
- To compare the effectiveness of our proposed method with state-of-the-art available methodologies.

Chapter 3

Fake News Detection Using Semi-Supervised Textual Frameworks

This chapter presents two semi-supervised fake news detection frameworks applied to textual news articles. The first approach is constructed on temporal ensembling-based ConvNet architecture where the missing labels are calculated as proxies using the self ensembling process. In the second approach Graph Convolutional Network is used for binary classification of articles into fake and real. Graph Convolutional Network can harness the best advantages of convolutions as well as data structuring capabilities of graphs to draw meaningful insights out of complex data and associated parameters. Further, the classification results of both semi-supervised approaches are validated on standard datasets and compared with existing state-of-the-art methods.

3.1 Introduction

Supervised artificial intelligence techniques are being well tested to provide promising performance in multiple real-life applications. As an evolving challenge in the arena of Artificial Intelligence, fake news detection requires huge labelled data to build supervised learning-based detection models. The methods used for annotating scraped web articles include online crowdsourcing platforms such as Amazon Mechanical Turk, Kaggle etc. [16]. However, labelling the news articles scraped from the web is too expensive and necessitates enormous workforce due to the massive volume of data. Furthermore, crowdsourcing annotations are impossible to accomplish in near real-time as well as create lots of annotation inconsistency for extensive data labelling. With the increasing data size, the data labelling inconsistencies will be worse. Therefore, it is obligatory to design well-organized and practical frameworks for fake news classification, that is supposed to be trained on a small amount of labelled data yet have the capability to explore the hidden

patterns of recent unlabelled data as well. We can use semi-supervised learning techniques when we have both labelled and unlabelled data. In that setting, unlabelled data can be used to improve model performance and generalization. Labelled data is a scarce resource. The whole labelling process is costly and needs active monitoring to avoid assessment flows.

In a nutshell, the research problem is to design an efficient fake news detection framework that can save time, labour, cost and inconsistencies involved during data annotation process of supervised techniques and to connect with the benefits of both labelled as well as unlabelled up-to-the-minute untapped data. To counter all the above-stated problems, we proposed two semi-supervised fake news detection architectures described in the following sections.

3.2 A Temporal Ensembling based Semi-supervised ConvNet for the Detection of Fake News Articles

Supervised learning is a pricey, time-consuming, tedious and painstaking approach as we need to label huge data capacities beforehand. Unsupervised learning's usages spectrum is limited. Hence, the paradigm of semi-supervised learning to harness the best out of limited labelled as well as a large amount of unlabelled data is well suited. In semi-supervised learning methods, the unlabelled portion of data is huge and more recent but it lacks the true labels needed for error calculation and updating the network parameters through optimizers using backward error propagation. To fulfil this necessity, the network needs to have a proxy for the true labels of unlabelled data.

The concept of Temporal Ensembling is introduced by Laine and Aila [145]. This framework is based on the principle of self-ensembling which is an effective method of getting label proxy that can be used as a substitute for the missing labels. Self-ensembling combines the outputs of all previous epochs of a neural network in the form of a weighted sum to serve as an unsupervised target to compare against the current epoch output. During the training phase, each input is fed into the network only once and its corresponding output is memorized to update the ensemble of previous predictions. In iterative epochs, ensemble predictions approach towards the true labels and guide the network in the right direction.

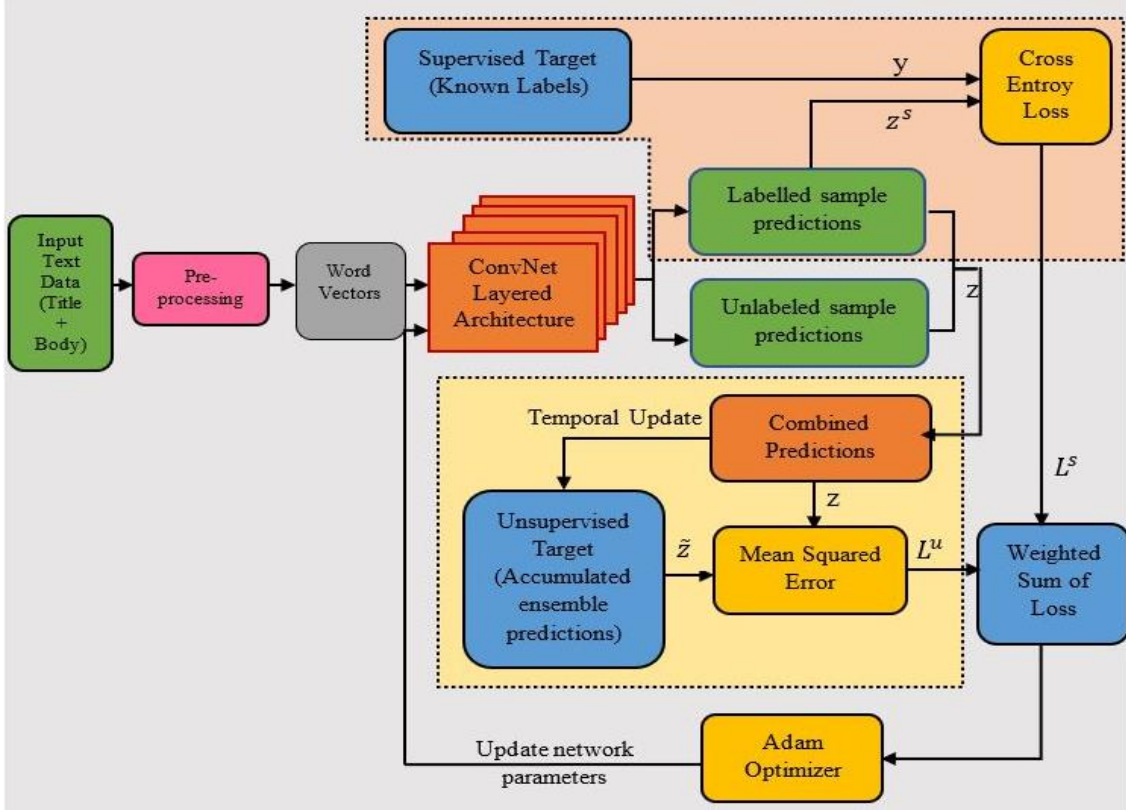


Figure 3.1: ConvNet Semi-Supervised Framework

3.2.1 Methodology

The framework of the proposed model is demonstrated in Figure 3.1 and descriptive layered architecture of Convolutional neural network is presented in Figure 3.2. Algorithm 1 describes the corresponding training procedure of ConvNet semi-supervised architecture. The training data consists of total N input samples out of which $M = |S|$ is the number of labelled samples used (S is a set of labelled samples). For every $i \in S$, we have an identified accurate label $y_i \in C$ where C denotes the number of different classes. Here we have $C = \{Real, Fake\} \cup \{1, 0\}$ as we are considering a binary classification problem. Labels y_i are available for annotated input data training samples. B is the set of minibatch indices. In each minibatch, we calculate the model output and the loss based on the previous outputs and the available labels. Supervised loss component (cross-entropy) is calculated only for labelled inputs ($M = |S|$) and unsupervised loss component (mean squared error) is calculated for all the input training samples (N).

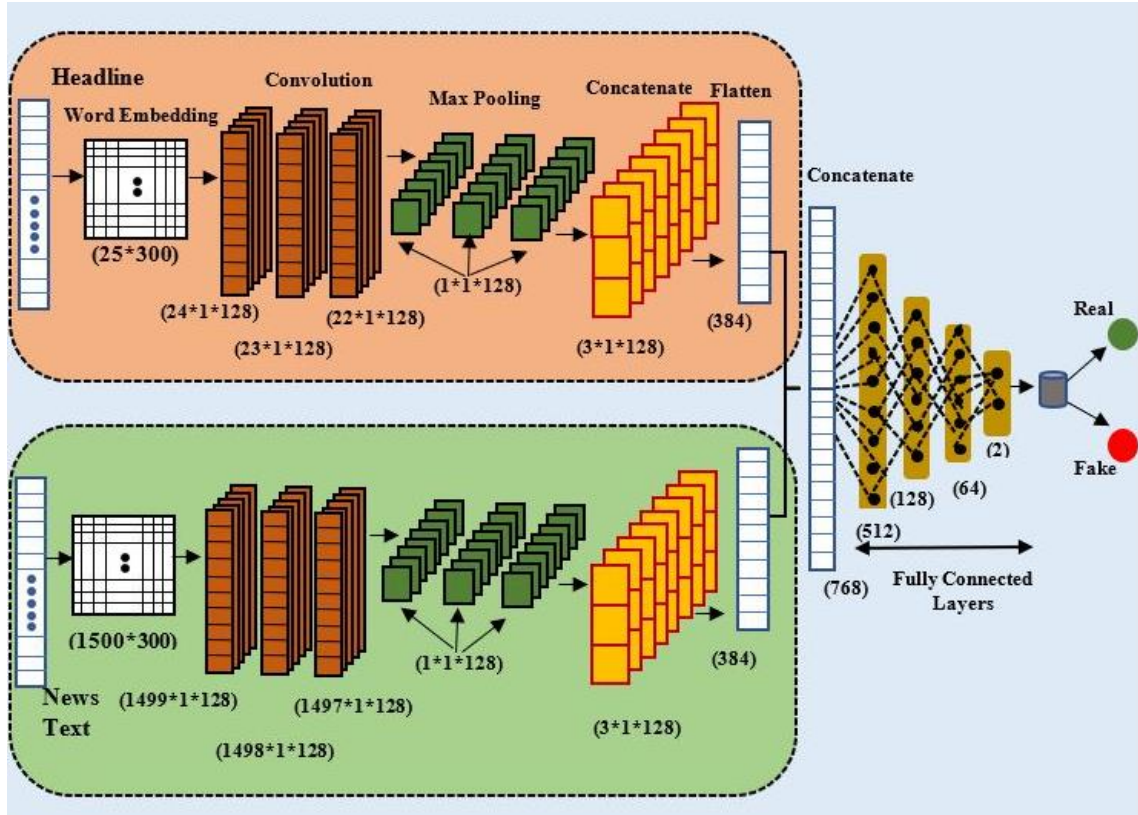


Figure 3.2: ConvNet Layered Architecture

Input samples are news instances with two attributes Headline/Title and News Text/Body. Word2vec word embedding is used to represent each input word to a 300-dimensional vector space. Word embeddings are a class of natural language processing technique used to represent words in the form of a numeric vector in geometric space out of their semantic meaning, where the distance between two vectors would relate to semantic relation between two words. Each row in input tensor represents one word and the total number of rows represent the maximum length of the input tensor. Maximum sequence length for a title is taken as 25 and text is taken as 1500 words which cover almost all instances completely. Padding and truncation methods are used to handle the out of size input headline and text pairs.

Algorithm 3.1: Learning in ConvNet Semi-supervised Framework

Parameter Initialization

- x_i =training sample
 - S =set of training sample with known labels
 - y_i =label for labelled input x_i , $i \in S$
 - α =ensemble momentum, $0 \leq \alpha \leq 1$
-

•	$w(t)$ =unsupervised weight ramp-up function	
•	$f_{embedding}(x)$ =word embedding	
•	$f_{\theta}(x)$ =neural network with trainable parameter θ	
•	$\mathbf{Z} \leftarrow \mathbf{0}_{[NXC]}$	▷ initialize ensemble predictions
•	$\tilde{\mathbf{z}} \leftarrow \mathbf{0}_{[NXC]}$	▷ initialize unsupervised target vectors
1:	for t in $[1, \text{num epochs}]$ do	
2:	for each minibatch B do	
3:	$x'_{i \in B} \leftarrow f_{embedding}(x_{i \in B})$	▷ convert words into word embeddings
4:	$z_{i \in B} \leftarrow f_{\theta}(x'_{i \in B}, t)$	▷ output of neural network
5:	$L^s_{i \in B} \leftarrow -\frac{1}{ B } \sum_{i \in (B \cap S)} \log z_i^s [y_i]$	▷ supervised loss component
6:	$L^u_{i \in B} \leftarrow \frac{1}{ B } \sum_{i \in B} \ z_i - \tilde{z}_i\ ^2$	▷ unsupervised loss component
7:	Loss $\leftarrow L^s_{i \in B} + w(t)L^u_{i \in B}$	▷ total loss
8:	update θ	▷ optimize trainable network parameters
9:	end for	
10:	$\mathbf{Z} \leftarrow \alpha \mathbf{Z} + (1-\alpha)z$	▷ temporal update ensemble predictions
11:	$\tilde{\mathbf{z}} \leftarrow \frac{\mathbf{Z}}{(1-\alpha^t)}$	▷ bias correction for unsupervised target vector
12:	end for	
13:	return θ	

At the end of each training epoch, the current predictions (z) are accumulated and the temporal ensemble outputs (previous predictions) \mathbf{Z} are updated according to Eq. (3.1), where α is a momentum characteristic that determines how far the ensemble ranges into training history. On the first training epoch since no data from previous epoch is available, \mathbf{Z} and $\tilde{\mathbf{z}}$ both are initialised as zero tensors. After the first epoch, we have $\mathbf{Z} = (\mathbf{1} - \alpha)z$, this start-up bias is fixed according to Eq. (3.2). In a nutshell, \mathbf{Z} encompasses an ensemble of the weighted average of network outputs from earlier epochs, which gives greater weights to the recent epoch's outputs than distant epochs.

$$\mathbf{Z} = \alpha \mathbf{Z} + (1 - \alpha)z \quad (3.1)$$

$$\tilde{\mathbf{z}} = \frac{\mathbf{Z}}{(1-\alpha^t)} \quad (3.2)$$

L^s is the standard cross-entropy loss calculated only for labelled input samples whereas Mean Squared Error L^u is calculated for overall (labelled and unlabeled) samples between current epoch outputs (z_i) and the temporal outputs (\tilde{z}_i). The overall loss function Eq. (3.5) is a linear combination of the supervised loss component Cross-Entropy Eq. (3.3) and unsupervised loss component Mean Squared Error Eq. (3.4). Overall Loss is optimized

using Adam optimizer which is a stochastic gradient descent technique based on the adaptive estimation of first and second-order moments. In the initial epochs of training, supervised loss factor dominated the learning gradients and the total loss whereas at the later stages more contribution is being done by unlabelled data.

$$\mathbf{Supervised\ Loss} = \mathit{Cross\ Entropy}(z^s, y) = L_{i \in B}^s = -\frac{1}{|B|} \sum_{i \in (B \cap S)} \log z_i^s [y_i] \quad (3.3)$$

$$\mathbf{Unsupervised\ Loss} = \mathit{MSE}(z, \tilde{z}) = L_{i \in B}^u = \frac{1}{c|B|} \sum_{i \in B} \|z_i - \tilde{z}_i\|^2 \quad (3.4)$$

$$\mathbf{Overall\ Loss} = \mathit{Cross\ Entropy}(z^s, y) + w(t) \mathit{Mean\ Squared\ Error}(z, \tilde{z}) \quad (3.5)$$

The unsupervised component of the loss function is weighted by a time-dependent function w_T that slowly ramps up along a Gaussian curve, starting from zero and expressed according to Eq. (3.6). T is the max epoch value and t is the current epoch. The behaviour of unsupervised weight function with respect to the epochs (the network is trained for 100 epochs) in our experiments is plotted in Figure 3.3.

$$\mathbf{Unsupervised\ weight\ ramp\ up\ function} = w_T(t) = \exp(-5 \left(1 - \frac{t}{T}\right)^2) \quad (3.6)$$

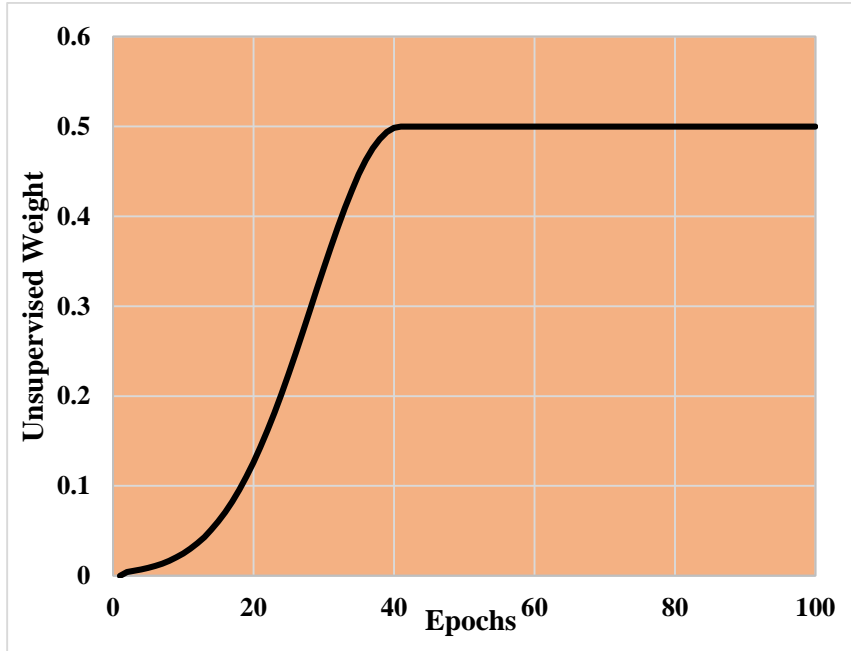


Figure 3.3: Unsupervised Weight Function

Convolution kernels are applied parallel on two input tensors title and text. To produce non-linearity in the convolution output ReLU activation function is used. Three convolution kernels of sizes [2,3,4]*Embedding dimension(300) with 128 filters of each size are applied on title and text tensors. Size of the filter signifies that how many words we are processing together at one time as the number of columns in the 2D matrix are same for all words. The output of 128 filters of the same size is passed on to max-pooling layer. This layer combines all outputs produced by different filters of the same size and merge them into a single feature by applying the max operation. The max operation ensures that we captured the most relevant feature of the sentence or document. There are two main advantages of pooling: (a) The number of parameters or weights are considerably reduced, thus reducing the computation cost. (b) It reduces overfitting. Table 3.1 represents all input and output dimensions of proposed CNN architecture along with the number of trainable parameters in each layer.

Ensembling momentum α is taken as 0.6, which is also called accumulation decay constant of temporal ensembling. α is a term that determines the influence of past predictions on the temporal ensemble. The network is being trained up to 100 epochs with a batch size of 64. L2 is the kernel regularizer in 2D convolution layers with a parameter value 0.01. Adam optimizer with learning rate 0.0002 is used, momentum parameters β_1 and β_2 for optimizer are taken as 0.9 and 0.980, respectively. Feature vectors extracted from title and text tensors separately are being concatenated together to get the final feature map of most important salient features. We have used four dense layers after final feature vector extraction for softer dimensionality reduction. The first and second dense layer contains 512 and 128 neurons, respectively followed by a third dense layer of 64 neurons. To support binary classification final dense layer contains two neurons. Dropout probability is 0.5 in all the dense layers. The initial three dense layers are supported with ReLU activation function whereas the final dense layer is supported with softmax function for binary classification of articles into Real or Fake.

Table 3.1: Dimensions of ConvNet Semi-Supervised Architecture

Layer	Input Dimension	Output Dimension	Description	Parameter #
Title-Conv2d-1	(25,300,1)	(24,1,128)	128 filters of kernel size 2*300	60100
Title-Conv2d-2	(25,300,1)	(23,1,128)	128 filters of kernel size 3*300	90100
Title-Conv2d-3	(25,300,1)	(22,1,128)	128 filters of kernel size 4*300	120100
Title-maxpool2d-1	(24,1,128)	(1,1,128)	Max pooling operation	0
Title-maxpool2d-2	(23,1,128)	(1,1,128)	Max pooling operation	0
Title-maxpool2d-3	(22,1,128)	(1,1,128)	Max pooling operation	0
Title -Concatenate	(1,1,128); (1,1,128); (1,1,128)	(3,1,128)	Concatenating title feature vectors of 3 different kernels after pooling	0
Title-flatten	(3,1,128)	(384)	1-Dimensional single feature vector	0
Text-Conv2d-1	(1500,300,1)	(1499,1,128)	128 filters of kernel size 2*300	60100
Text-Conv2d-2	(1500,300,1)	(1498,1,128)	128 filters of kernel size 3*300	90100
Text-Conv2d-3	(1500,300,1)	(1497,1,128)	128 filters of kernel size 4*300	120100
Text-maxpool2d-1	(1499,1,128)	(1,1,128)	Max pooling operation	0
Text-maxpool2d-2	(1498,1,128)	(1,1,128)	Max pooling operation	0
Text-maxpool2d-3	(1497,1,128)	(1,1,128)	Max pooling operation	0
Text -Concatenate	(1,1,128); (1,1,128); (1,1,128)	(3,1,128)	Concatenating text feature vectors of 3 different kernels after pooling	0
Text-flatten	(3,1,128)	(384)	1-Dimensional single feature vector	0
Concatenate (Title, Text)	(384) ;(384)	(768)	Combining feature vectors from title and text branch	0
Dense-1	(768)	(512)	Dense layer with 512 neurons	42150
Dense-2	(512)	(128)	Dense layer with 128 neurons	21550
Dense-3	(128)	(64)	Dense layer with 64 neurons	2550
Dense-4	(64)	(2)	Dense layer with 2 neurons	102

3.2.2 Model Parameter Description

We implemented the framework in section 3.2.1 using python 3.7 programming language on 64-bit windows 10 operating system with installed memory (RAM) of 8 GB and Intel Core i5 7th generation 7200U processor. The model is trained on NVIDIA TESLA P100 GPU and 13GB RAM on Kaggle Notebooks Online GPU platform. The architecture is being trained independently on each of the three datasets and then tested. Each of the three datasets is being split into 70% for training and 30% for testing purpose. Out of the training data samples, 20% is being used for validation and the remaining are used for training iteratively by changing the proportion of labelled and unlabelled data. Five independent

training and testing iterations are being done for every dataset by altering the percentage of labelled data starting from 10%, 20%, 30%, 40% and 50 % respectively. Python Libraries such as Tensorflow, Keras, Pandas, Gensim, Scikit-learn, matplotlib etc. are used for implementation and result plotting purpose. The following Table 3.2 justifies the choice of sensitivity parameters in the convolutional neural network and highlights the reasons for specific parameter adjustments in the network architecture.

Table 3.2: Parameter Tuning in ConvNet Semi-Supervised Architecture

Parameter	Value	Justification
Title Length	25	Average title length calculated combinedly of all the instances of three datasets is 22 and maximum comes out as 27. So, 25 is being chosen as the title length by truncating the last few words of larger titles which are negligible and padding the smaller ones with zeros.
Text Length	1500	Average text length calculated combinedly of all the instances of three datasets is 1319 and maximum comes out as 1627. So, 1500 is chosen as the text length by truncating the last few words of larger news text which are very less and padding the smaller ones with zeros.
Word Embedding Dimension	300	Standard size of Word2Vec pre-trained word vector representation to capture the complete semantic meaning of a word in the language
Convolution Kernel Size	(2x300); (3x300); (4x300)	To analyze the semantic meaning of bigram, trigram and four-grams in the sentence. Words taken together as a phrase provides the accurate meaning of their usages in the sentence rather than a single word. 300 is the embedding dimension.
Dense Layers	(512); (128); (64); (2)	Four dense layers are added with 512, 128, 64 and 2 neurons respectively for smoother dimensionality reduction to retain all crucial information for correct binary classification.

Title of a news article describes it as a whole with specific keywords, so initially, filters are applied on title and text part of news articles in parallel and then feature vectors are concatenated together using early fusion as features extracted out of both the branches corresponds to the same text modality. Finally, to converge the framework into a binary classification system multiple dense layers are added with gradually reducing number of neurons for softer dimensionality reduction so that all the distinguishable features required for classification can be retained. Semi-supervised convolutional neural network coupled with temporal self-ensembling architecture performs optimum feature engineering out of both labelled and unlabelled corpus. Hence, the final configuration of the neural network has been obtained with the dimensions of layers described in Table 3.1 and specific parameter adjustments detailed in Table 3.2.

3.2.3 Datasets

Three different datasets Fake News Detection (Jruvika), Fake News Data and Fake News Sample (Pontes) hosted on Kaggle platform and fully labelled are being used for model training and testing purpose. The datasets consist of multiple attributes, so we extracted headline, body and label part of each one of them. The following Table 3.3 details in the specifics of datasets:

Table 3.3: Dataset Details

Dataset Name	Details	Attributes Used	Total entries	Fake news	Real news
Fake News Detection by Jruvika [176]	Hosted on Kaggle, contains site URL, Headline, Body and Label	Headline/Title, Body/Text, Label	3988	2121	1867
Fake news Data [177]	Hosted on Kaggle, contains id, Title, Author, Text and Label	Headline/Title, Body/Text, Label	20700	10360	10340
Fake News Sample by Guilherme Pontes [178]	Hosted on Kaggle, contains 17 attributes including site URL, Headline, Body and Label	Headline/Title, Body/Text, Label	45569	20226	25343

Fake News Detection [176] dataset is uploaded on the Kaggle website by Jruvika. It contains four attributes site URL, Headline, Body and Label (Real/Fake). The dataset originally contains 4009 news instances. After initial data cleaning such as removing the entries with missing labels, missing headline and body we have 3988 rows with 2121 Fake and 1867 Real news instances.

Fake News Data [177] was hosted on Kaggle website two years ago as a dataset for Kaggle competition, now available openly with annotations for analysis and learning purpose. It contains 20, 800 samples with five labels: Id, Title, Author, Text and Label. After removing entries with missing fields, we have 20, 700 entries in the dataset with 10360 Fake news and 10340 real news entries. We extracted only Title, Text and label part for our model training, testing and comparison purpose.

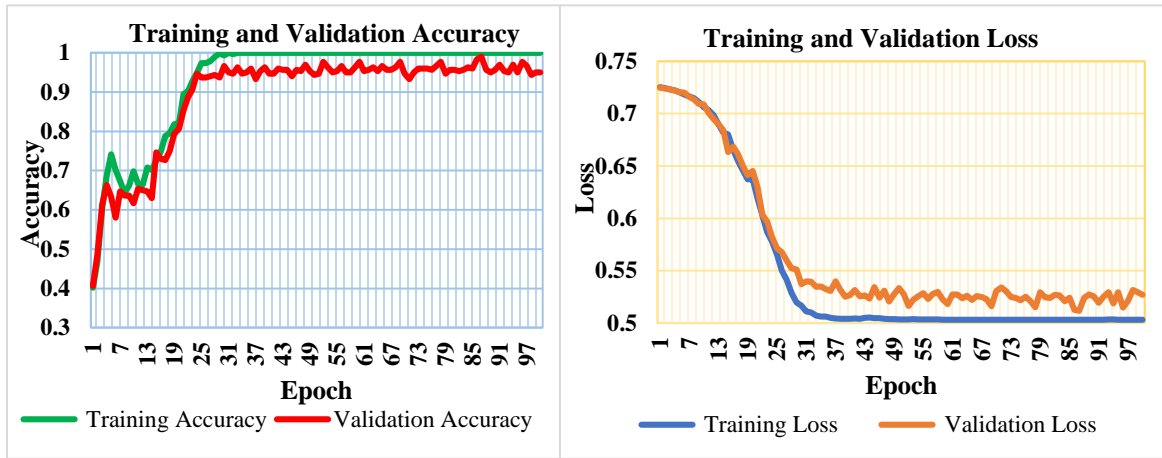
Fake News Sample [178] dataset is hosted on Kaggle by Guilherme Pontes and contains news articles labelled into different categories hate, satire, clickbait, political, conspiracy, fake, reliable, rumour, unreliable etc. and has 17 different attributes of each news. We filtered rows with fake, rumour, unreliable into Fake news category and reliable into real news category. After preliminary filtering, we have a total of 45 569 rows with three fields Headline, Body and Label (Real/Fake), out of which 25 343 are real news and 20 226 are fake news articles.

3.2.4 Result Analysis

The performance evaluation of our framework on Fake News Detection (Jruvika) dataset in terms of accuracy, precision, recall and F1 score for different percentages of labelled data is detailed in Table 3.4. Figure 3.4 (a) highlights the trends of change in training and validation accuracy with epochs. Figure 3.4 (b) focuses on model loss during training and validation phase. The area under the ROC curve is 0.98, which represents the robustness of the model as shown in Figure 3.4 (c). The confusion matrix values for the test samples are made known in Figure 3.4 (d). All the values and graphs in Figure 3.4 are plotted for the final iteration of experiments for Jruvika dataset in which portion of labelled data is 50%.

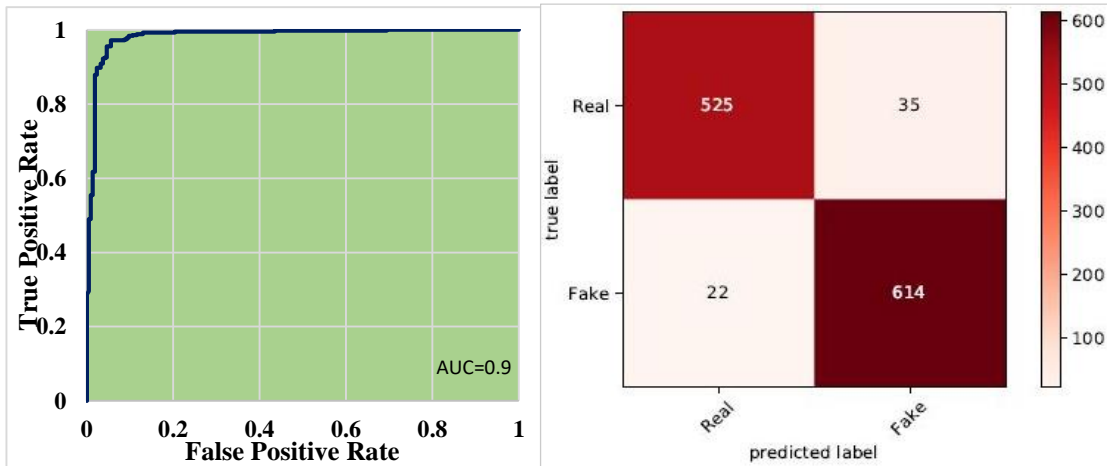
Table 3.4: Result Analysis on Fake News Detection (Jruvika) Dataset

Ratio of labelled data	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
10%	80.10	76.83	82.32	79.48
20%	89.71	88.26	90.00	89.12
30%	93.47	94.62	91.25	92.90
40%	94.31	95.38	92.32	93.82
50%	95.23	95.97	93.75	94.84



(a)

(b)



(c)

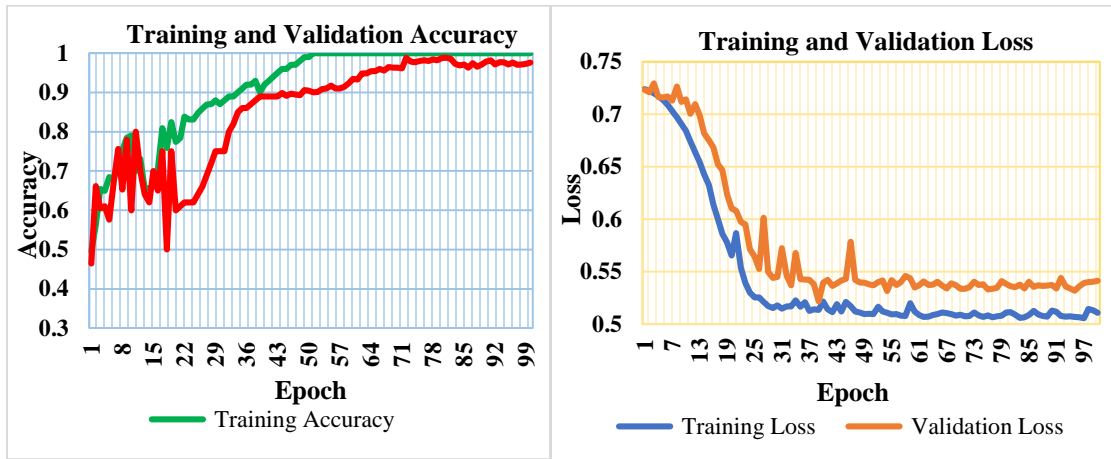
(d)

Figure 3.4: Performance of Jruvika Dataset (50% labelled) in terms of (a) Training and Validation accuracy (b) Training and Validation Loss (c) ROC Curve (d) Confusion Matrix

The highest value of accuracy, precision, recall and F1 score for Fake News Data dataset is 97.45%, 95.62%, 99.45% and 97.49%, respectively for 50% labelled data, represented in Table 3.5. Figure 3.5 (a) –(d) plots the results for the highest % of labelled training data iteration in terms of accuracy vs. epochs, loss vs. epochs, ROC-AUC curve and confusion matrix.

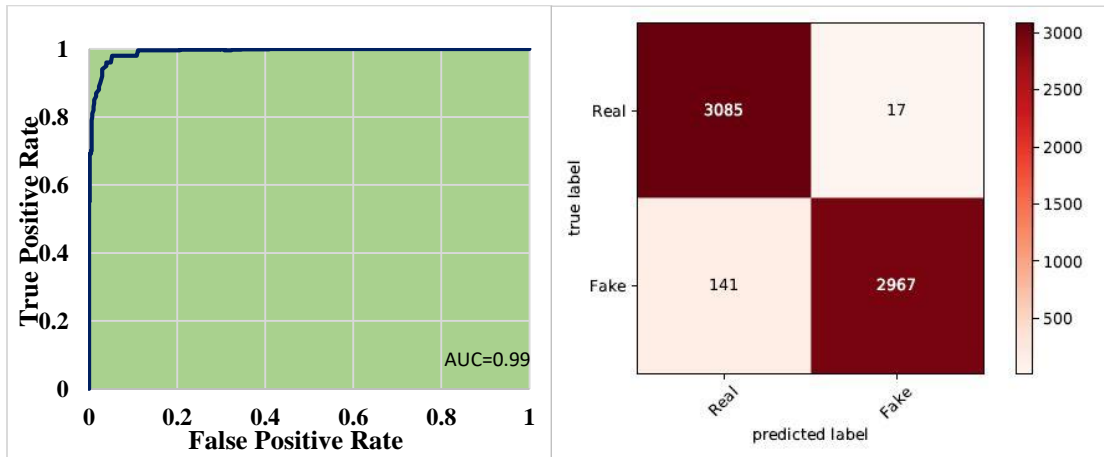
Table 3.5 : Result Analysis on Fake News Data Dataset

Ratio of labelled data	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
10%	83.55	87.96	77.77	82.55
20%	90.14	90.63	89.52	90.07
30%	91.09	92.41	89.52	90.07
40%	97.02	94.83	99.45	97.08
50%	97.45	95.62	99.45	97.49



(a)

(b)



(c)

(d)

Figure 3.5: Performance of Fake News Data Dataset (50% labelled) in terms of (a) Training and Validation accuracy (b) Training and Validation Loss (c) ROC Curve (d) Confusion Matrix

The performance matrices on Fake News Sample (Pontes) dataset are highlighted in Table 3.6. The highest classification accuracy for the dataset is 93.05 % for 50% labelled

data and Area Under the ROC curve is 0.94. Figure 3.6 (a) –(d) represents the analytical curves for training, validation and testing phase in terms of accuracy and loss trends, ROC-AUC curve and confusion matrix.

Table 3.6: Result Analysis on Fake News Sample (Pontes) Dataset

Ratio of labelled data	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
10%	82.42	85.35	82.57	83.93
20%	88.51	90.09	89.14	89.61
30%	91.21	93.26	90.75	91.98
40%	91.62	93.51	91.26	92.37
50%	93.05	95.32	92.02	93.64

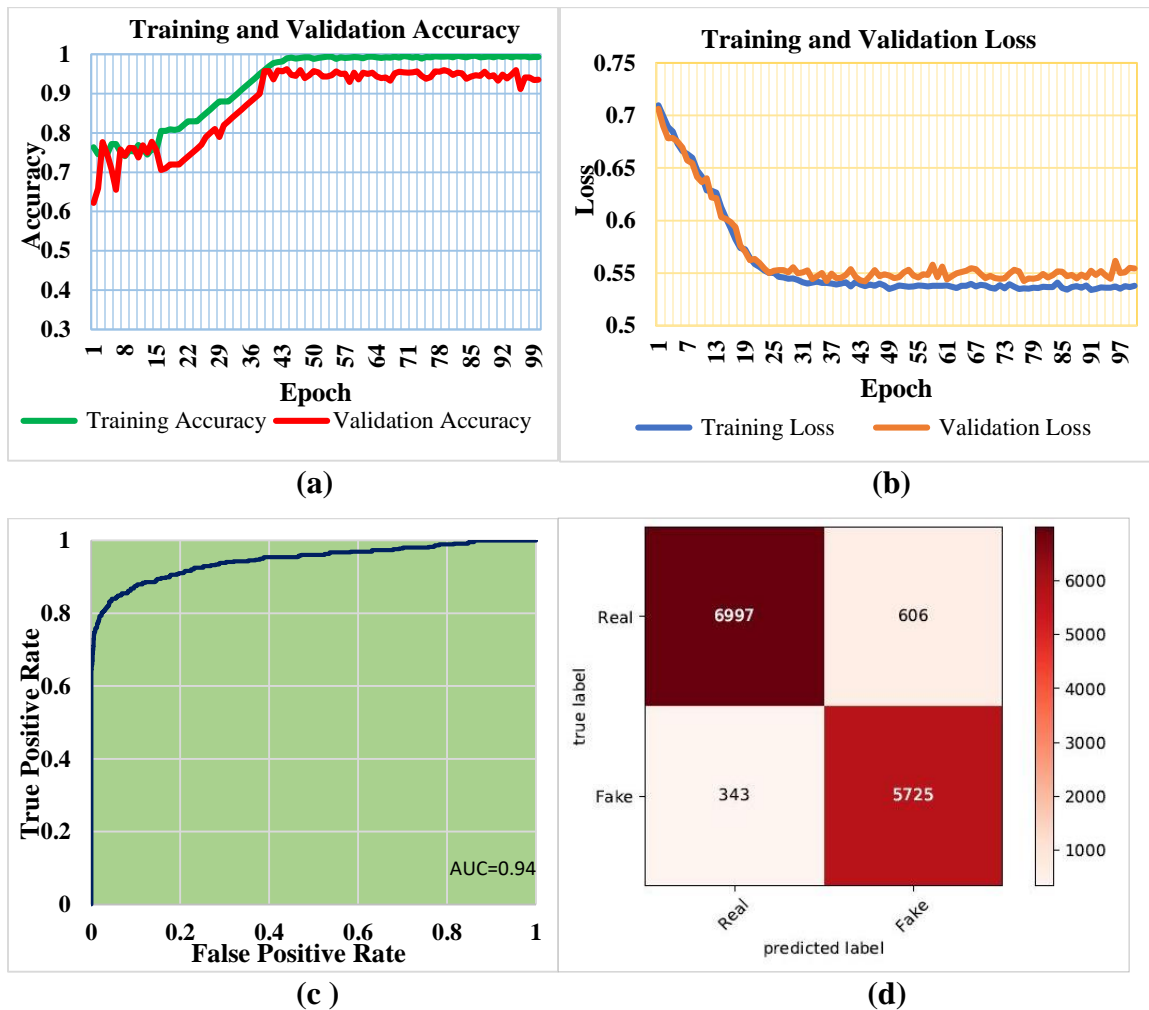


Figure 3.6: Performance of Fake News Sample Dataset (50% labelled) in terms of (a) Training and Validation accuracy (b) Training and Validation Loss (c) ROC Curve (d) Confusion Matrix

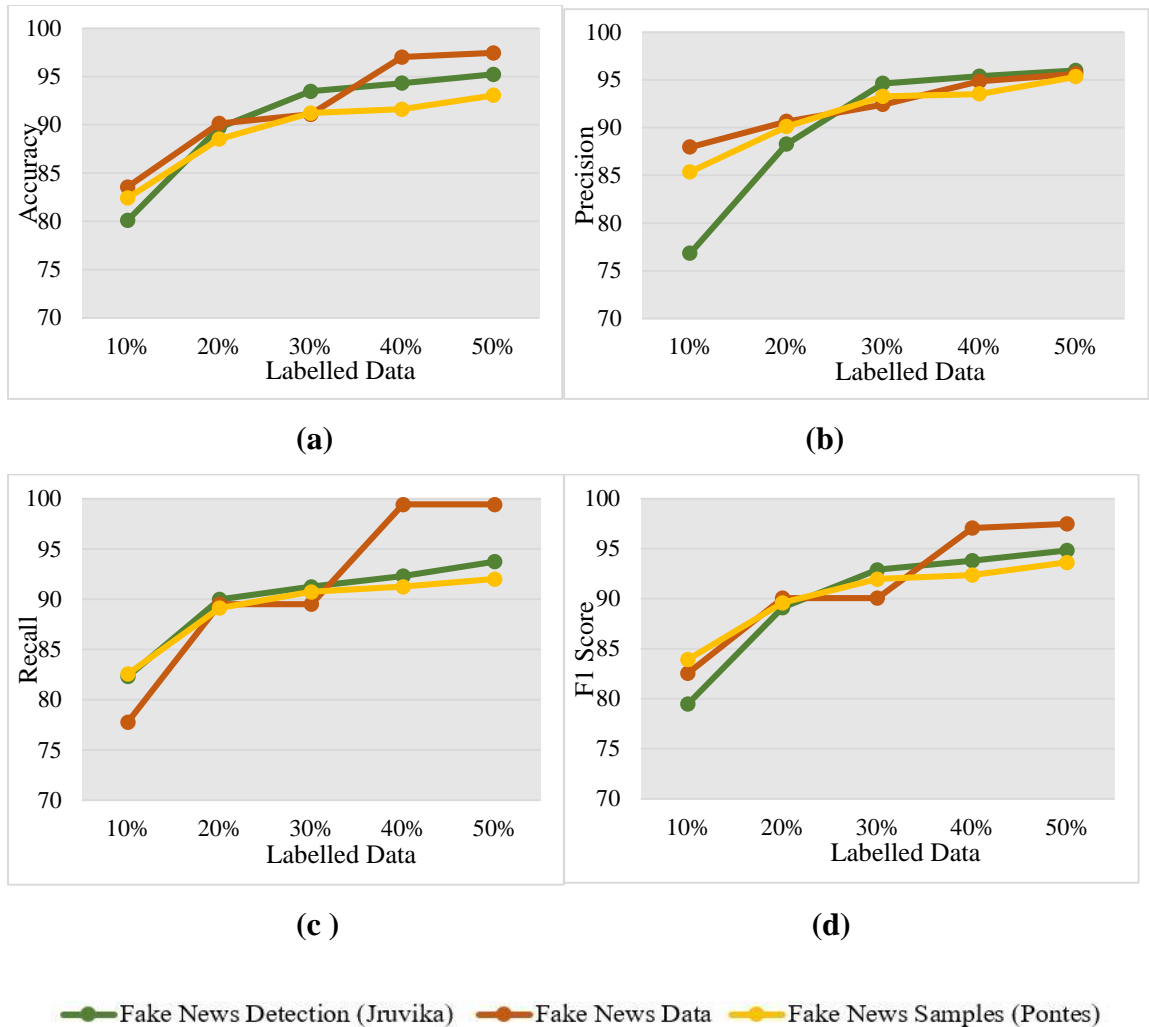


Figure 3.7: Comparative analysis of all the three datasets for (a) Accuracy (b) Precision (c) Recall (d) F1-Score with variation in % of labelled data

Figure 3.7 analysis the comparative performance of all the three datasets on performance matrices for the varying proportion of labelled training data. It can be observed from the analysis of Table 3.4, 3.5 and 3.6 that when we increase the labelling of training samples from 10% to 20% the accuracy improvement is considerable but from 40% to 50% its almost stable, the reason is that the ensemble predictions of unknown labels done by the proposed ConvNet semi-supervised architecture are as good as the real ones. So, after training the model with relatively a smaller number of labelled samples it will give good results and increasing the volumes of labelled data after that will not affect the performance of the model in much considerable way. Experimentation of proposed framework on Jruvika, Fake News Data and Pontes datasets confirm the fact that the model is skilled enough for classifying fake news and can get high performance even with limited labelled training

samples. Even though the proposed framework establishes its dominance on multiple criteria still there are some limitations or inadequacies, which are listed as follows:

- We have focused our fake news detection system for text news analysis only but multimedia in the form of images is a highly influential part of the online information system and could also be included in future research.
- The proposed system lacks source credibility or authenticity factors which may also play a crucial role in identifying fake news.
- Our system classifies online information into two categories true or false. The model can be converted into a multiclass fake news analysis system or categorization based on some rating/scale will make the solution more robust.
- The system could be extended as a stand-alone application or browser plugin to make it useful in real life.
- Timeline analysis of a news story by web scrapping different versions or reverse searching it on the internet can lead to a deeper understanding of the purpose and genesis of fraudulent content.

3.2.5 Baseline Comparison

To compare the effectiveness of our proposed framework with contemporary techniques, we implemented the baseline methods with all the parameter settings as described in the original research paper with same data split that we have used in our work. We employ all labelled data to train the baseline models as the data in the original dataset is fully labelled. Compared to the baseline we train the proposed model with partially labelled data which is a very small portion of the original dataset. Table 3.7, 3.8 and 3.9 compare the performance of our work with earlier state-of-the-arts on Fake News Detection (Jruvika), Fake News Data and Fake News Sample (Pontes) dataset. The following methods are used for baseline comparison.

- Bali et al. [179] developed Random Forest (RF), Support Vector Classifier (SVC), Naïve Bayes (NB), K-Nearest Neighbour(KNN), Multi-Layer Perceptron (MLP), AdaBoost (AB) and Gradient Boosting (XGB) classifiers by extracting sentiment

polarity, n-gram count, 50-dimensional GloVe word embedding and cosine similarity features between headline and body part of news articles.

- Agarwalla et al. [180] applied Punkt statement tokenizer from NLTK library for tokenizing headline and body part of news instances after preprocessing. They applied Support Vector Machine (SVM), Logistic Regression and Naïve Bayes with Lidstone smoothing machine learning techniques for classification.
- Karimi and Tang [181] proposed a Hierarchical Discourse Level Structure using Bi-LSTM, which identifies structure-related properties of fake news and real news articles by constructing dependency trees.
- Vishwakarma et al. [124] introduced a method of extracting keywords from captions written on images and news text, then these keywords are authenticated on the internet after web scrapping using a rule-based classifier. The classifier uses a reality parameter (Rp) which is calculated by checking the credibility of the top 15 Google search results. In our experiments, we extracted keywords from headline and news text and them apply the rule-based classifier for an impartial comparison.
- Ajao et al. [182] explored hybrid CNN and RNN deep models for veracity analysis of online information content. This method is based on identifying relevant features associated with fake news stories without previous knowledge of the domain.
- Kumar et al. [136] proposed that CNN with bidirectional LSTM ensemble network with attention mechanism provides the best fake news classification accuracy after converting text into vectors using 100-dimensional GloVe word embedding.

Table 3.7: Performance comparison on Fake News Detection (Jruvika) Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Bali et al. [179]	86.20	92.00	92.00	92.00
Agarwalla et al. [180]	83.87	81.22	82.67	81.93
Karimi and Tang [181]	83.09	80.22	82.01	81.10
Vishwakarma et al. [124]	88.30	85.20	88.40	86.77
Ajao et al. [182]	92.02	92.20	92.77	92.48
Kumar et al. [136]	93.00	94.06	87.30	90.55
Proposed Work	95.23	95.97	93.75	94.84

Table 3.8: Performance comparison on Fake News Data Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Bali et al. [179]	91.05	93.00	94.00	93.49
Agarwalla et al. [180]	86.32	85.89	81.26	83.51
Karimi and Tang [181]	85.90	88.80	80.70	84.56
Vishwakarma et al. [124]	85.00	95.04	80.00	86.87
Ajao et al. [182]	92.91	91.05	92.78	91.90
Kumar et al. [136]	94.89	93.06	98.20	95.56
Proposed Work	97.45	95.62	99.45	97.49

Table 3.9: Performance comparison on Fake News Sample (Pontes) Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Bali et al. [179]	92.03	91.00	92.14	91.56
Agarwalla et al. [180]	88.00	87.90	88.00	87.94
Karimi and Tang [181]	84.06	80.90	85.70	83.23
Vishwakarma et al. [124]	89.01	92.50	95.70	94.07
Ajao et al. [182]	90.07	91.50	90.02	90.75
Kumar et al. [136]	89.00	87.01	88.08	87.54
Proposed Work	93.05	95.32	92.02	93.64

The baseline comparisons supported with accuracy, precision, recall and F1-score metrics, discussed in Table 3.7, 3.8 and 3.9 establishes the excellence and superiority of proposed ConvNet semi-supervised framework over the contemporary techniques. Our proposed method outperformed in achieving the goal of detecting fake news using minimum possible annotated data at a high accuracy rate so that the far-reaching effect of fake news can be avoided and damage can be minimized. The model perfectly learned complex data patterns present in news articles to help it precisely classify fake and real news articles. The following salient technical features make our solution better than other contemporary frameworks to overcome the limitations and problems of existing fake news detection systems:

- The semi-supervised convolutional neural network performs optimum feature engineering out of both labeled and unlabelled corpus.
- Self-ensembling provides consensus predictions of the labels of unannotated data using previous epochs outputs of network-in-training.
- Accumulated temporal ensemble predictions are proved to be a better predictor for the unknown labels than the output of the most recent training epoch, thus suitable to be used as a proxy for the labels of unannotated data.

- Convolution kernels of three different sizes 2x300, 3x300 and 4x300 (300 is the embedding dimension) are used to capture bigrams, trigrams and four-gram features of the text.
- The title of a news article describes it as a whole with unique keywords, so initially, filters are applied on title and text parallelly and then feature vectors are concatenated together using early fusion because both the branches relate to the same text modality.
- Finally, to converge the framework into a binary classification system, multiple dense layers are added for softer dimensionality reduction so that all the distinguishable features used for classification can be retained.

3.2.6 Significant Outcomes

Self-ensembling via temporal ensembling is a cost-effective but powerful way to squeeze more performance out of a convolutional neural network, irrespective of whether the samples are annotated or not. The following claims encapsulate the significant outcomes of our work: -

- We structure a novel semi-supervised temporal ensembling based convolutional neural network architecture being trained with a limited amount of annotated corpus which leverages the concept of self ensembling for fake news classification in text news articles.
- Temporal ensembling aggregates the outputs of all previous epochs into a collaborative prediction that is expected to be closer to the accurate unknown labels of unannotated inputs. Thus, the labels inferred this way acts as an unsupervised training target to compare against for unlabelled data.
- ConvNet filters are separately applied on headline and body part of the news articles and then extracted feature vectors are concatenated to take advantage of both the slices.
- The framework is validated on three publicly available large datasets hosted on Kaggle platform namely Fake News Detection (Jruvika), Fake News Data and Fake News Sample (Pontes) for different percentage of labelled and unlabelled data.

- The training of proposed neural network using varied proportions of unlabelled and labelled samples for all the three datasets delivers the holistic performance analysis and accuracy trends under different circumstances.

3.3 Fake News Detection Using Semi-Supervised Graph Convolutional Network

The semi-supervised framework of Graph Convolutional Network combines the best features of convolutions and data modelling capabilities of graphs. The proposed architecture incorporates a three-step approach to classify the articles as fake or real. First step is to transform the given textual data from a dataset in the form of vectors. These vectors represent the linguistic features of the text and can be utilized to represent an article in Euclidean Space. Global Vector (GloVe) embedding is used to transform the articles into vectors. Each article is interpreted as the mean of vectors of the words it contains. The result of this embedding is used to construct a similarity graph between articles in the dataset. Each node represents an article and most similar nodes are connected by an edge in the graph. To calculate the similarity, i.e. distance between two articles, Word Mover's Distance is used. It uses semantically meaningful relations between words to find the similarity between two articles. Before converting text into vectors basic pre-processing steps such as: dropping unused columns, removing null and missing data, removing stop words, tokenizing the articles, lemmatizing each word, converting labels name etc. are done for all the datasets. The subsequent procedure involves word embedding and similarity graph construction. This results in the creation of graphs containing three types of nodes labelled as real, fake and unlabelled which can be input to the graph convolutional neural networks for finally predicting the output. The process is pictorially emphasized in Figure 3.8 and also comprehensively detailed in algorithm1.

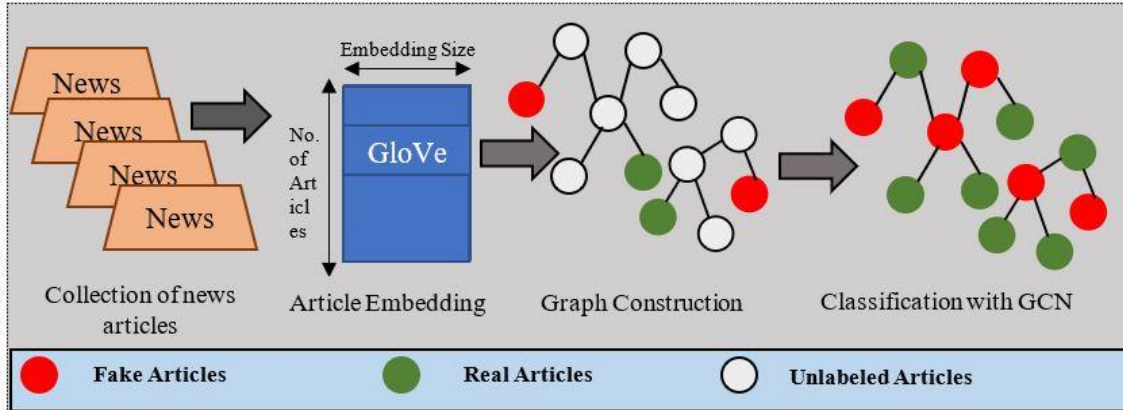


Figure 3.8: Proposed Semi-Supervised Fake News Classification Framework

Algorithm 3.2: Learning in Semi-supervised GCN Framework

- 1: Perform initial pre-processing on input samples
 - 2: Split the input samples into 80% training, 10% validation, 10% testing
 - 3: Article embedding in Euclidean space using GloVe with Embedding Dimension=300
 - 4: Represent every article as the mean vector of the corresponding GloVe embedding
 - 5: Similarity graph construction using WMD with K-nearest neighbour; K=3 and K=5
 - 6: Train GCN of 4 layers up to 120 epochs with dropout=0.5 and learning rate=0.01
 - 7: Evaluate the performance on test set
-

3.3.1 Article Embedding in Euclidean Space

Article embedding in Euclidean Space means transforming text to multidimensional vectors by using word embedding. Word embedding is the process of constructing mathematical equivalents, i.e., representation in the mathematical form of each word from some corpus. It is implied that words with similar or same sense will have close representations. In Natural Language Processing, word embedding has proved to be one of the concepts that have huge applications, as it lays the foundation for solving many real-world tough problems. The process involves constructing vectors for each word in any number of dimensions such that it can be visualized as a vector in a vector space in those number of dimensions. This conversion of words to vectors uses neural network methodology to find the values of vectors. The vector can have any number of dimensions ranging from a single digit to a few hundred. As words with similar or same sense will have close representations, it can be implied that the meaning of words has been learnt by a model, which in turn can be useful for solving challenging problems. Our method will utilize the final word vectors created by the embedding of words for each article by representing it in an euclidean space. Further this will also be converted into low dimensions using a dimensionality reduction

technique and contextual associations in news articles will be obtained by using a similarity graph representation.

To create a vector which maps to each news article, our proposed semi-supervised method uses a pre-trained GloVe model of 300 dimensional embeddings and for all the words appearing in that article we compute a mean vector which is finally mapped to that article. GloVe (Global Vectors) an unsupervised learning method was developed at Stanford University as an open-source project and was launched in 2014. Word Vectors in such an n-dimensional space, occurs in the manner as similar words appear close to each other while words with different meanings appear far from each other. One major advantage that GloVe offers is its dependency on global statistics instead of local statistics as used by Word2Vec or some other embedding techniques. GloVe is based on the basic principle of co-occurrence matrix. A simple example of a co - occurrence matrix of window size 1 is depicted in Figure 3.9, illustrates how words befall collectively that ultimately produces the relations between individual words. Co-occurrence matrix for words is calculated just by adding the mutuality of occurrence of words in an obtained text. In general use cases this co-occurrence matrix is decomposed by employing dimensional reduction techniques like PCA and SVD.

	the	cat	sat	on	mat
the	0	1	0	1	1
cat	1	0	1	0	0
sat	0	1	0	1	0
on	1	0	1	0	0
mat	1	0	0	0	0

Figure 3.9: An example co-occurrence matrix for the sentence “the cat sat on the mat”

3.3.2 Similarity Graph Construction

The classification step requires a graph input, which is done by constructing a similarity graph representing closeness between different nodes, i.e., each news article. Such a graph can be represented as a k-nearest neighbour graph. In particular, for every news article, we

look forward towards the k -nearest neighbours in the embedding space. The k neighbours can be determined by calculating the Word Mover's distance between one article to all the other news articles.

A k -NN (k -nearest neighbour) graph is one in which node p and node q have an edge between them if node p is one of the top k nearest of all its neighbours or vice-versa. Every k -nearest-neighbours of a location in the n -dimensional space will be determined to utilise a "closeness" relationship wherever closeness is usually described within words of a distance measure as Word Mover's Distance. Thus, with given articles in a vector space, a k -NN graph of points can be created by calculating the remoteness between each pair of points and connecting each point with the k most proximal ones.

The Word Mover's Distance (WMD) proposed by Kusner et al. [183] in 2015 is a distance between documents that takes benefit of semantic relations among words that are apprehended by their embeddings. It is a special case of Earth Mover's Distance [184] and also provides extraordinary performance when coupled with K -nearest neighbour in classification tasks. It basically measures the dissimilarity between two text documents as the minimum distance the word vectors of one document need to travel to reach the word vectors of another document. This distance proved to be quite effective, obtaining state-of-art error rates for classification tasks.

WMD metric is a distance function between text documents that leads to unparalleled low k -nearest neighbour document classification error rate. Precisely representing the remoteness between two documents has far reaching utilities in document retrieval, multilingual document matching, text clustering etc. To calculate a distance between two text documents the basic unit is "travel cost" between two words. Let "A" and "B" be the representations of two text documents with $|A|$ and $|B|$ are the number of unique words in both the documents respectively. Each word w_i in document A is allowed to be converted into any word in document B. $T_{ij} \geq 0$ denotes the amount of distance word w_i in A has to travel to convert into word w_j in B. T is a sparse transportation flow matrix. To transform A completely into B we have to confirm that the complete outgoing flow from i^{th} word

equals A_i and incoming flow to j^{th} word must match B_j represented mathematically as Eq. (3.7).

$$\sum_j T_{ij} = A_i \text{ and } \sum_i T_{ij} = B_j \quad (3.7)$$

The WMD distance between two documents is the minimum weighted cumulative cost required to move all words from document A to document B represented as:

$$\text{WMD}(A, B) = \sum_{ij} T_{ij} c(w_i, w_j) \quad (3.8)$$

Formally the WMD between two documents is defined as the value of the optimal solution of the following transportation problem which is a special case of Earth mover's distance.

$$\min_{T \geq 0} \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} c(w_i, w_j) T_{ij} \quad (3.9)$$

$$\sum_{j=1}^{|B|} T_{ij} = A_i \quad \forall i \in \{1, 2, 3, \dots, |A|\} \quad (3.10)$$

$$\sum_{i=1}^{|A|} T_{ij} = B_j \quad \forall j \in \{1, 2, 3, \dots, |B|\} \quad (3.11)$$

$$T_{ij} \geq 0 \quad \text{for all } i, j \quad (3.12)$$

Word Mover's Distance (WMD) is derived upon current events in embeddings of the words which determine the semantically significant description of those words of local co-occurrences within those sentences among new articles. Salient features of Word Mover's Distance can be characterized as:

- It is easy to learn, practice and free from the effect of any hyperparameters.
- It can be easily described as the distance among two text contents that can be split and described as the distances which are sparse within different words.
- It commonly includes some information represented as the word2vec or Glove and heads to huge operations accuracy.

3.3.3 GCN for Graph Classification

Graph Convolutional Network is a creative development of convolutional neural networks which functions directly on graphs. The model scales linearly in the number of graph edges by using a competent layer-wise propagation rule that is based on first-order approximation of spectral convolutions on graphs. GCN model is capable of learning hidden layer representations that encodes both node feature and graph structure to an extent useful for semi-supervised classification.

Graph Convolutional Networks (GCN) are an ideological extension to Convolutional Neural Networks (CNN) where convolution process is applied on a graph instead of pixels which constitute the image. As CNN can capture information from the images and this information then can be used to classify images, a similar approach can be built over graphs as well. A filter analogous to a CNN filter can be employed in case of GCN to capture the similarity in graphs. Graph Convolutional Networks (GCN), like Convolutional Neural Networks (CNN) have bounded no. of hyper-parameters, which leads these techniques to occupy less memory and thus, multiple levels of information can be built to give a remarkable result when compared to a traditional learning algorithm. Figure 3.10 illustrates the basic design pipeline for a GCN model.

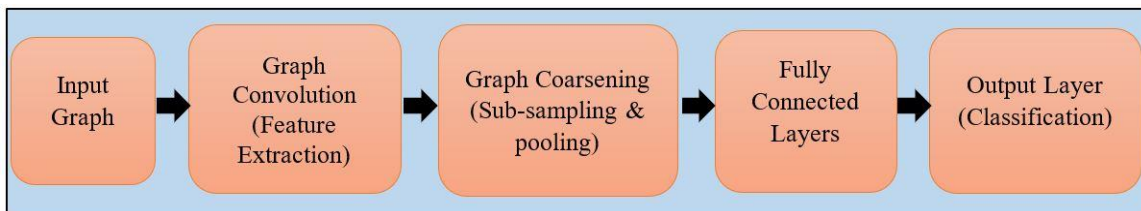


Figure 3.10: Design pipeline for GCN Model

The primary purpose of graph classification is to foretell the labels of nodes within that particular graph. A node in the graph denotes typically a real-world object i.e. in our framework it represents news articles. The graph has been extensively employed to represent objects of real-world and the connection between them. Node/graph classification results prediction of labels of nodes in the graphs, for this purpose several researches take advantage of relationships among nodes to increase the classification efficiency.

GCN is one of the prominent variants of Graph Neural Network being used for a lot of real-life applications with non-Euclidean real world data having graph structure. The major difference between CNN and GNN is that GNN is a generalized variant of CNN built to operate on irregular or non-Euclidean structured data whereas the CNN can function only in case where the underlying data is regular (Euclidean). GCN is developed by Thomas Kipf and Max Welling [185] in 2016. Convolution in GCN is almost the same operation as the convolution in CNN. It implies multiplying the input neurons with a set of weights known as kernels or filters acts as sliding window and enables the network to learn features from neighbouring cells. Different layers may contain filters of different weights but within the same layer same filter will be used known as weight sharing. The particular variant of GCN that we use in our framework for semi-supervised classification is Spectral Graph Convolutional Network proposed by Thomas Kipf and Max Welling.

Neural networks apply non-linear activation functions to represent the non-linear features in latent dimensions. Forward pass equation in neural network is represented as in Eq. (3.13) where σ represents activation function, $H^{(l+1)}$ and $H^{(l)}$ represents feature representation at l^{th} and $(l+1)^{\text{th}}$ layer, $W^{(l)}$ weight and $b^{(l)}$ bias at layer l .

$$H^{(l+1)} = \sigma(W^{(l)} H^{(l)} + b^{(l)}) \quad (3.13)$$

Forward pass equation in Graph Convolutional Network can be represented as:

$$H^{(l+1)} = \sigma(W^{(l)} H^{(l)} A) \quad (3.14)$$

A is adjacency matrix representing the connections between the nodes in graph structured real world data. Adjacency matrix A enables the model to learn the feature representations based on nodes connectivity. Bias b is omitted to make the model simpler. Eq. (3.14) is the first-order approximation of spectral graph convolution propagating the information along the neighbouring nodes within the graph.

The aim of Graph Convolutional Network framework is to learn a function of features on a graph which takes as input a feature matrix X of dimension $N \times D$ where N is the number of nodes and D is the number of input features. A is the adjacency matrix for representing the overall graph structure. A and X are the input to GCN architecture producing

the node level output Z (an $N \times F$ feature matrix, where F is the number of output features per node). A nonlinear function representing every neural network layer in GCN can be written as in Eq. (3.15) and Eq. (3.16) highlights layer-wise propagation across the network.

$$H^{(l+1)} = f(H^{(l)}, A) \quad (3.15)$$

$$f(H^{(l)}, A) = \sigma(A H^{(l)} W^{(l)}) \quad (3.16)$$

With $H^{(0)}=X$ is initial input, $H^{(L)}=Z$ is node level output at last layer (or z for graph level output), L is the number of layers, $W^{(l)}$ is a weight matrix for the l^{th} neural network layer and σ is a non-linear activation function. Instead of using matrix A for semi-supervised classification in practical scenario Kipf and Welling [185] proposed to use symmetric normalization $D^{-1/2} A D^{-1/2}$, so Eq. (3.15) can be rewritten as:

$$f(H^{(l)}, A) = \sigma(\widehat{D}^{(-1/2)} \widehat{A} \widehat{D}^{-1/2} H^{(l)} W^{(l)}) \quad (3.17)$$

With $\widehat{A} = A + I$, I is identity matrix and \widehat{D} is the diagonal node degree matrix of A . In the process of semi-supervised classification with GCN the network initially starts training on the labelled nodes, subsequently propagating the information to unlabelled nodes by updating weight matrices that are shared across all nodes. This process can be summarized in the following steps:

- Accomplish forward propagation through GCN.
- Put on the sigmoid function row-wise on the last layer in the GCN.
- Calculate the cross-entropy loss on known node labels.
- Backpropagate the loss and update the weight matrices ‘ W ’ in each layer.

3.3.4 Implementation Details

In this sub-section, we extensively detail the experimentation settings, datasets, parameter selection and insights gained from the wide range of experiments performed. The implementation is done on Google Colab which offers up to 13.53 free RAM and 12 GB NVIDIA Tesla K80 GPU. The proposed framework is built and implemented in Python 3 on top of

the Keras deep learning framework. To make our implementation better and effective the experiments are repeated for different volumes of labelled and unlabeled data samples. The performance scores and comparison of the proposed framework are listed in terms of F1-measure, accuracy, recall and precision evaluation metrics. Analysis of the results is being done in numerical as well as in graphical representation using the shapes of accuracy-loss curves with epochs and area under the curve plots individually for each dataset.

After initial pre-processing each dataset is split into 8:1:1 ratio for training, validation and testing respectively. Articles are embedded in vector space using 300-dimensional GloVe word embedding. Word Mover’s Distance is employed for constructing similarity graph utilizing K-nearest neighbours for two different values of K i.e. K=3 and K=5. Graph convolutional network involves 4 layers convolutions for feature extraction with intermediate leaky ReLU activation and dropout layers. GCN of 4 layers is trained up to 120 epochs with dropout 0.5, learning rate 0.01 , 16 hidden units and weight decay of 5×10^{-4} . After fully connected layers in output layer activation function is softmax for binary classification of news instances into real or fake.

3.3.5 Datasets

Three different fully labelled datasets Fake News Data, Real or Fake and Fake News Detection introduced on Kaggle platform are used for experimentation and validation of our work. The headline, body and label part of each one of the datasets are being utilized for model training and testing purpose. The details of the datasets are listed in following Table 3.10.

Table 3.10: Dataset Details

Dataset Name	Details	Attributes Used	Total entries	Fake news count	Real news count
Fake News Data [177]	Hosted on Kaggle, contains id, Title, Author, Text and Label	Headline/Title, Body/Text, Label	20700	10360	10340
Real or Fake [186]	Hosted on Kaggle platform, contains four fields Id, Headline, Body, Label	Headline/Title, Body/Text, Label	6000	3000	3000
Fake News Detection [176]	Hosted on Kaggle, contains site URL, Headline, Body and Label	Headline/Title, Body/Text, Label	3988	2121	1867

Fake News Data [177] was presented on Kaggle website for Kaggle competition two years ago, now available publicly with annotations for research and learning purpose. It comprises of 20, 800 instances with five attributes: Id, Title, Author, Text and Label. After initial pre-processing, we have 20, 700 entries in the dataset with 10360 Fake news and 10340 real news entries. We utilized Title, Text and label features for our model training, testing and comparison purpose.

Real or Fake [186] dataset was introduced on Kaggle platform three years ago, now is being extensively used for research purpose. It contains four attributes Id, Headline, Body and Label out of these four we have used headline, body and label part in our research. The dataset initially has 6335 entries reduced to 6000 after preliminary data pre-processing. Segregation of dataset is 3000 real and 3000 fake news instances.

Fake News Detection [176] dataset is compiled on the Kaggle website by Jruvika. It contains four attributes site URL, Headline, Body and Label (Real/Fake). The dataset initially contains 4009 news instances. After initial data cleaning such as removing the entries with misplaced labels, missing headline and body we have 3988 rows with 2121 Fake and 1867 Real news samples.

3.3.6 Result Analysis

Performance of the proposed framework on three different datasets is being evaluated and compared for accuracy, precision, recall and F1 score. The experiments for each dataset have been repeated for two different values of K (K=3, K=5) and varied proportions of labelled training data ranging from 20% to 50%. Table 3.11,3.12 and 3.13 highlights the results obtained from experiments for Fake News Data, Real or Fake and Fake News Detection datasets, respectively. Figure 3.11, 3.12 and 3.13 enlightens the Accuracy-Epoch curve, Loss-Epoch curve and ROC curves for each one of the datasets.

Table 3.11: Result Analysis on Fake News Data Dataset

% La- belled data	K=3				K=5			
	Accu- racy (%)	Preci- sion (%)	Recall (%)	F1-score (%)	Accu- racy (%)	Preci- sion (%)	Recall (%)	F1-score (%)
20%	79.29	82.52	74.65	78.39	66.03	61.00	79.27	68.94
30%	83.26	86.07	77.32	81.46	74.66	85.32	72.44	78.35
40%	87.99	75.99	88.60	81.81	82.36	69.02	87.45	77.15
50%	91.18	93.05	94.27	93.65	88.67	81.71	89.26	85.32

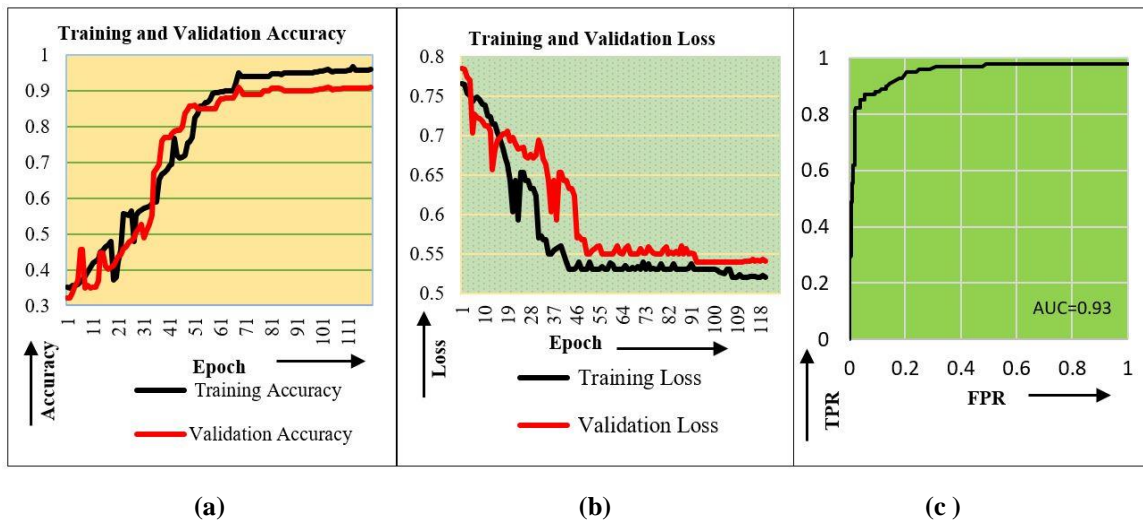
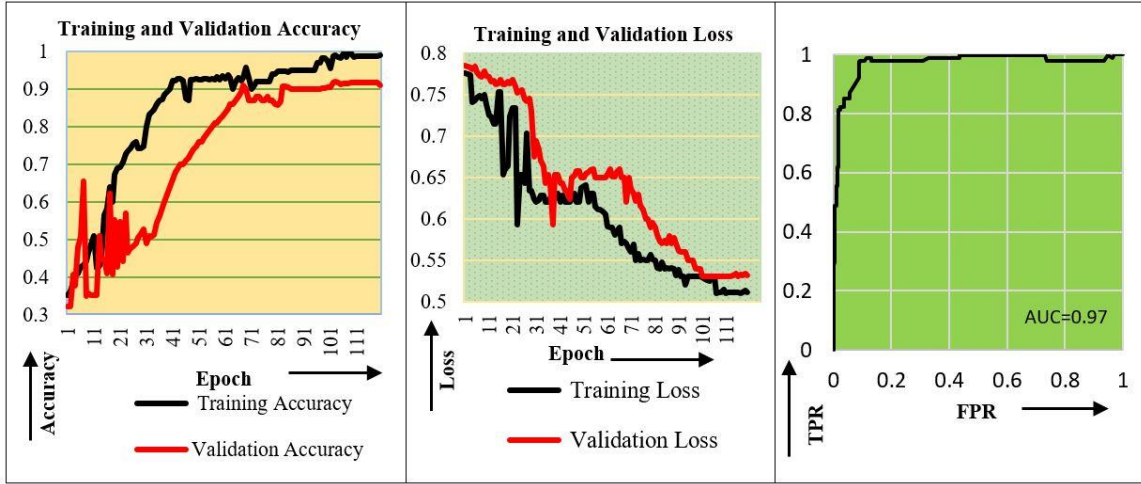


Figure 3.11: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for Fake News Data Dataset

Table 3.12: Result Analysis on Real or Fake Dataset

% La- belled data	K=3				K=5			
	Accu- racy (%)	Preci- sion (%)	Recall (%)	F1-score (%)	Accu- racy (%)	Preci- sion (%)	Recall (%)	F1-score (%)
20%	79.55	83.75	79.99	81.83	80.00	83.69	70.08	76.28
30%	86.32	75.00	86.76	80.45	83.33	75.00	79.27	77.08
40%	91.02	89.00	90.23	89.61	87.19	76.68	93.55	84.28
50%	95.27	89.47	95.99	92.61	92.34	84.29	92.52	88.21

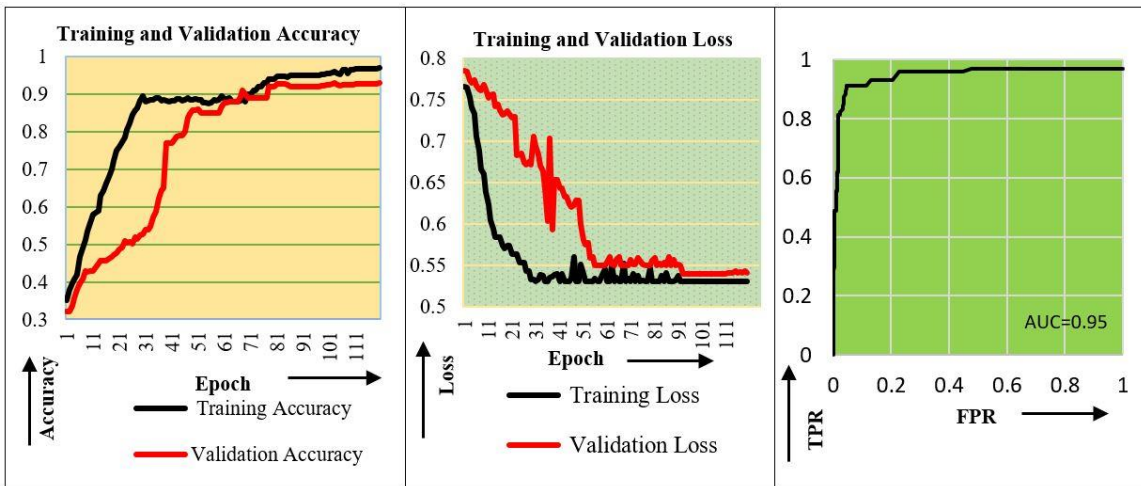


(a) (b) (c)

Figure 3.12: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for Real or Fake dataset

Table 3.13: Result analysis on Fake News Detection dataset

% La- belled data	K=3				K=5			
	Accu- racy (%)	Preci- sion (%)	Recall (%)	F1-score (%)	Accu- racy (%)	Preci- sion (%)	Recall (%)	F1-score (%)
20%	75.03	76.06	71.27	73.58	77.32	81.33	76.65	78.92
30%	81.26	75.60	86.59	80.72	86.23	87.88	78.19	82.75
40%	85.04	88.27	87.77	88.02	88.69	86.01	86.62	86.31
50%	92.03	92.07	94.87	93.45	90.37	83.33	90.17	86.61



(a) (b) (c)

Figure 3.13: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for Fake News Detection dataset

It is evident from the results analysed from Table 3.11, 3.12 and 3.13 that $K=3$ provides better accuracies than $K=5$ and as we increase the amount of labelled data in training process precision of detecting fake news also increases. The highest accuracy obtained from Fake News Data dataset is 91.18%, Real or Fake dataset is 95.27 % and Fake News Detection dataset is 92.03% respectively. Graphical representation of the experimental process in Figures 3.11, 3.12 and 3.13 advocates the overall effectiveness of the framework in terms of quality and quantity of performance.

3.3.7 State-of-the-art Comparison

To compare the efficacy of our designed architecture with contemporary methods, we calculated the performance of five different baseline methods in terms of accuracy, precision, recall and F1 score with the dataset split as 8:1:1 for training, validation and testing. State-of-the-art juxtaposition on each of the Fake News Data, Real or Fake and Fake News detection datasets are outlined in Table 3.14, Table 3.15 and Table 3.16, correspondingly. The approaches used as baselines for state-of-the-art comparison are as follows:

- Bali et al. [179] proposed Support Vector Classifier (SVC), Random Forest(RF), Naïve Bayes (NB), Multi-Layer Perceptron (MLP), K-Nearest Neighbour(KNN), AdaBoost (AB) and Gradient Boosting (XGB) methods by extracting sentiment polarity, 50-dimensional GloVe word embedding , n-gram count, and cosine similarity features between title and text part of news articles.
- Agarwalla et al. [180] applied Punkt statement tokenizer from NLTK library with Support Vector Machine (SVM), Logistic Regression and Naïve Bayes with Lidstone smoothing for classification.
- Karimi and Tang [181] developed a Hierarchical Discourse Level Structure using Bi-LSTM, which extracts structure-related properties of articles by building dependency trees.
- Vishwakarma et al. [124] suggested a framework of keyword extraction with web scrapping using a rule-based classifier. The classifier uses a reality parameter (Rp) which is considered by checking the credibility of the top 15 Google search results.

Table 3.14: Comparative Performance Analysis on Fake News Data Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Bali et al. [179]	91.05	93.00	94.00	93.49
Agarwalla et al. [180]	86.32	85.89	81.26	83.51
Karimi and Tang [181]	85.90	88.80	80.70	84.56
Vishwakarma et al. [124]	85.00	95.04	80.00	86.87
Proposed Method	91.18	93.05	94.27	93.65

Table 3.15: Comparative Performance Analysis on Real or Fake Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Bali et al. [179]	93.03	92.00	92.44	92.22
Agarwalla et al. [180]	90.00	89.90	90.00	89.94
Karimi and Tang [181]	89.06	82.90	88.70	85.70
Vishwakarma et al. [124]	91.01	90.50	95.70	93.03
Proposed Method	95.27	89.47	95.99	92.61

Table 3.16: Comparative Performance Analysis on Fake News Detection Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Bali et al. [179]	86.20	92.00	92.00	92.00
Agarwalla et al. [180]	83.87	81.22	82.67	81.93
Karimi and Tang [181]	83.09	80.22	82.01	81.10
Vishwakarma et al. [124]	88.30	85.20	88.40	86.77
Proposed Method	92.03	92.07	94.87	93.45

The above discussion concludes that our proposed model for veracity analysis of web information is reasonably promising. The precision over all the three datasets Fake News Data, Real or Fake and Fake news detection, is a decent development over the parallel methods. Graph Convolutional Network and Word Mover’s Distance for calculating distance are the prominent technologies that have facilitated refining the preciseness of our results. Finally, experimenting with different values of K and varied proportions of labelled data helps us achieve 95.27% highest fake news detection accuracy.

3.3.8 Significant Outcomes

The significant outcomes of the proposed framework are as follows:

- The architecture presents a semi-supervised fake news detection technique based on GCN (Graph Convolutional Networks) trained with limited amount of labelled data.
- Elaboratively elucidate three building blocks of the framework: collecting word embeddings from the news articles in datasets utilising GloVe, constructing similarity graph using Word Mover's Distance (WMD) and finally applying Graph Convolution Network (GCN) for binary classification of news articles in semi-supervised paradigm.
- Graph Convolution Network can harness the best advantage of convolutions as well as data structuring capabilities of graphs to draw meaningful insights out of complicated data and associated parameters.
- The implemented technique is validated on three different datasets by varying the volume of labelled data.
- Experimental results are analysed for two different graph formulations constructed by taking $k=3$ and $k=5$.
- Comparison with other contemporary techniques also reinforced the supremacy of the proposed framework.

Chapter 4

Fake News Detection Using Supervised Multimodal Frameworks

This chapter focuses on supervised multimodal veracity analysis frameworks. The first approach consists of Hierarchical Attention Network, Image Captioning and headline matching module, Noise Variance Inconsistency and Error Level Analysis. These independent predictions are finally combined using the max voting ensemble method. The second model aims Inception-ResNet-v2 to extract visual features. BERT and ALBERT architectures have been used to elicit textual attributes. Diverse text input forms, like English articles, Chinese articles and Tweets, have been used to make our model robust and usable across multiple platforms. The architecture of Multimodal Early Fusion and Late Fusion has also been experimented and analyzed in detail by applying it on different datasets. The effectiveness of the proposed architectures is validated through experiments on standard datasets and state-of-the-art comparisons.

4.1 Introduction

At present, web platforms are governing our lives; their dark side enfolds human society as they are used as a medium for spreading misleading fake content to serve extremely malicious motives. Multimedia has become an integral part of human life as it has more conclusive, convincing and long-lasting memory effects. Almost every circulating news story is strengthened with accompanying images or videos. Multiple data formats have made forgery identification in online circulated news articles quite complicated because each data format has different characteristics. As the inherent attributes of varied data formats differ considerably, so as the techniques to detect their forgeries.

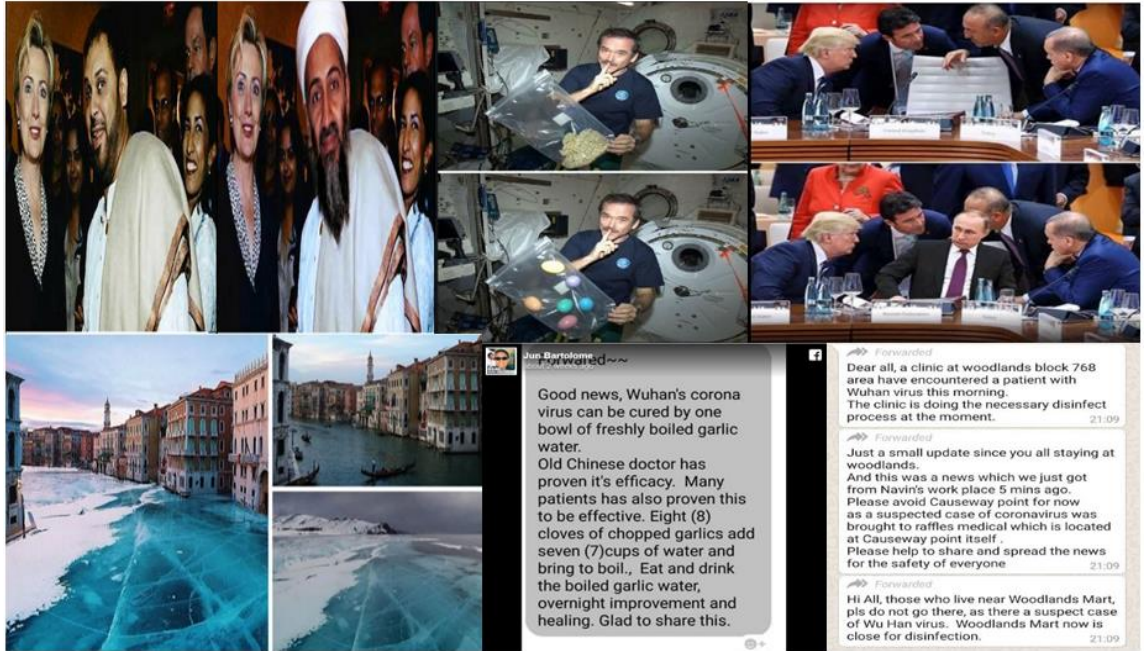


Figure 4.1: Sample visual and textual Fake news on web platforms

(a)Osama Bin Laden hosted in White House [187] (b) Astronaut Chris Hadfield Tested the Effects of Marijuana in Space [188] (c) Vladimir Putin surrounded by world leaders in G20 summit [189] (d)Frozen Venice [190](e)Boiled Garlic can cure coronavirus [191] (f)Woodlands Mart is closed for disinfecting coronavirus [192]

Some of the widely circulated fake news instances, along with doctored images are shown in Figure 4.1(a-d). Figure 4.1(e) and Figure 4.1(f) illustrate fake messages circulating on social media during the outburst of coronavirus in the early months of 2020 [191]. There are two main challenges in designing a framework for counterfeit news detection. First, false news is purposely shaped to confuse viewers and imitate traditional news sources, leading to a difficult situation in which it is tough to distinguish real news stories from fabricated ones. Second, the pace and amount at which false news is generated override the likelihood of rigorously testing and verifying all products by submitting reports for verification to human experts.

A photograph showing Osama Bin Laden meeting Hillary Clinton in the White House as in Figure 4.1(a) was shared in 2017 on social media in Russia. The reality was that Osama Bin Laden’s photograph had been superimposed on a photo of Mrs. Clinton meeting musician Shubhashish Mukherjee at an event in 2004. Figure 4.1(b) shows astronaut Chris Hadfield holding what looks like a bag of marijuana aboard the International Space

Station. This photo was posted on Facebook in November 2018 along with a caption stating that the astronaut was testing the effects of the drug in space. Actually, an image of Hadfield holding a bag of Easter Eggs was doctored to make it appear as if the astronaut was showing off a pouch of marijuana. Following the 2017 G20 summit in Germany, Russian media personalities and politicians started sharing a photoshopped image Figure 4.1(c) of Vladimir Putin being surrounded by other world leaders, including Donald Trump. The original photo shows those leaders gathered around an empty chair. Figure 4.1(d) an image circulated on Reddit.com in 2014 showing Frozen Grand Canal of Venice is actually a doctored image of Venice's Grand Canal and Russia's Frozen Lake Baikal. In all the above fake pictorial instances after manipulating the visual information the false event is very well narrated in text and it becomes extremely difficult to find out this forgery even by using advanced deep learning techniques if visual information is not being considered. Hence, we proposed two multimodal frameworks to use both textual and visual information for fake news detection on social media.

4.2 HAN, Image Captioning and Forensics Ensemble Multimodal Fake News Detection

A deep Hierarchical Attention Network is being trained with the text part, which is the concatenation of headline and body of news to extract the hidden patterns of fake news. The image accompanied with the news is being described using the automated caption generator tool and then the caption as well as the headline of news are being matched against the actual news text content. The resemblance between them indicates how much an image and headline have to do with the description of the news.

One of the easiest and widely used methods of image doctoring is photoshopping them and then deceiving human eyes by adding extra local noise in tampered parts so that the manipulations remain undetected. To counter this Noise Variance Inconsistency and Error Level Analysis, image forensic techniques are used. The Noise Variance Inconsistency method tests the presence of extra noise that remains inconsistent in visual data as compared with the original random noise that is evenly distributed in the image. Error

Level Analysis method works by inducing error in the image and then computing the difference between the two, revealing information about the doctored image. Finally, to gain the advantage of deep learning, image caption and headline matching with body, image forensics technique and to make the framework robust enough to detect all sorts of forgeries present in online news, we ensemble the independent models using the max voting technique as represented diagrammatically in Figure 4.2.

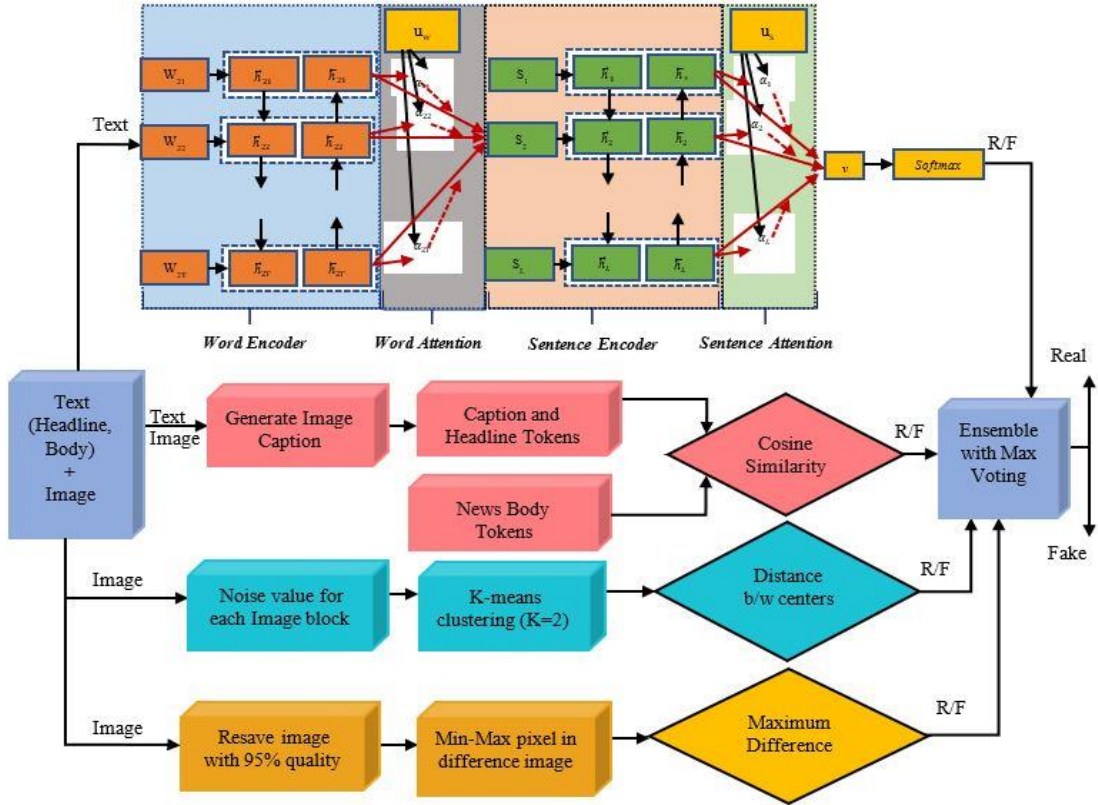


Figure 4.2: Architecture of Proposed Model

The proposed system consists of four independent parallel streams that are capable enough in detecting specific forgery formats. All four streams are applied to each input instance. Hierarchical Attention Network deals with headline and body part; Image captioning and headline matching module require all the three parts headline, body and image; Noise Variance Inconsistency and Error Level Analysis focuses only on images accompanied with news text. These independent predictions are finally combined using the max voting ensemble method. An exhaustive analysis is done on the experimental results of the

independent and combined ensemble framework. As an advantage of the ensemble architecture, the accuracy of the hybrid multimodal fake news detection framework improves considerably.

4.2.1 Pre-processing and Word Embedding

Pre-processing is the step of cleaning the data by removing excess, unnecessary and redundant parts of the text. It retains only meaningful tokens which are further converted into vectors using Word Embedding. Text data is pre-processed by using the NLTK python library with methods of Stop Word Removal, Stemming and Lemmatization, Normalization and Tokenization. Words and symbols that have no meaning are being removed, Stemming /Lemmatization transforms text to its root form, Normalization converts text to canonical form and Tokenization transforms longer strings into smaller tokens. Stemming uses a crude heuristic process to chop off the ends of words to reduce inflection and convert them into their root forms. The root word sometimes maybe the canonical form of the original word. Text normalization is extremely important for web content and social media information that contains a lot of noisy text in the form of abbreviations, misspellings and out-of-vocabulary words.

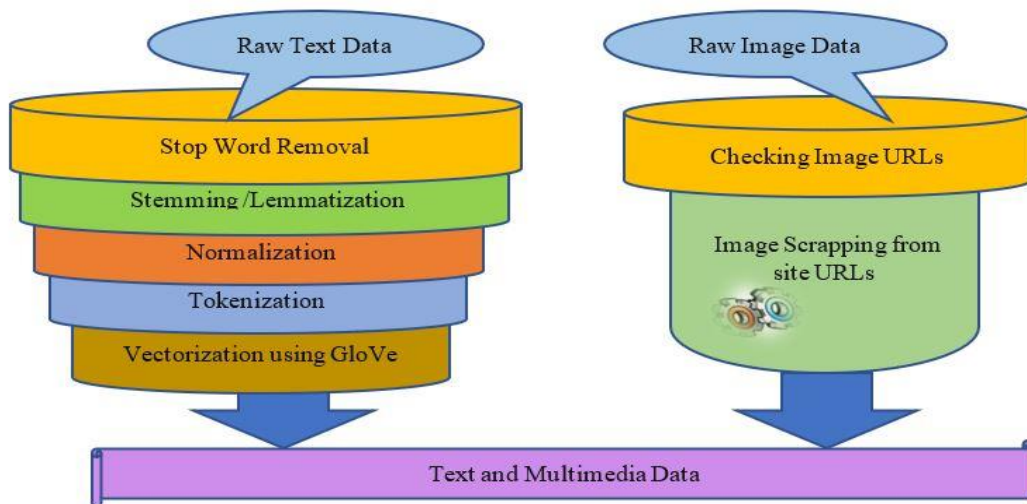


Figure 4.3: Pre-processing Text and Multimedia data

We use pre-trained GloVe word embedding to obtaining vector representation of text tokens. Glove generates word embeddings by aggregating the global word-word co-occurrence matrix from a corpus in an unsupervised learning algorithm. Compared with other word vector representations GloVe provides the advantage of capturing global statistics and semantics of words using a co-occurrence matrix. Multimedia data is pre-processed by checking that all image URLs are valid, for those instances in which images are missing or URLs are corrupted the corresponding multimedia is being scraped from the web using the BeautifulSoup Python library [193] . Figure 4.3 highlights the steps used for text and multimedia data pre-processing. Text and multimedia data are pre-processed separately and then concatenated together to make each instance complete in terms of its three parameters: headline/ title, body/text and image.

4.2.2 Hierarchical Attention Network (HAN)

Hierarchical Attention Network uses stacked recurrent neural networks that consist of four parts (a) Word Encoder (b) Word-Level Attention Layer (c) Sentence Encoder (d) Sentence-Level Attention Layer. The word encoding is followed by an attention mechanism to extract additional informative words that contributes more to sentence meaning. A sentence vector is formed by aggregating the representations of important words. The same attention procedure is then applied at the sentence level after being encoded using Bi-directional GRU to extract more useful sentences to form an article-level news vector. Final news vector v is a high-level representation of the news body and is used as a feature to classify the news as Real or Fake. w_{it} represents t^{th} word in i^{th} sentence, where T and L are the total number of words in a sentence and the total number of sentences in a document, respectively. x_{it} is the word vector corresponding to the word w_{it} and W_e is the embedding matrix. The word annotation h_{it} is calculated by concatenating the forward and backward hidden states $h_{it} = [\vec{h}_{it}, \overleftarrow{h}_{it}]$.

$$x_{it} = W_e w_{it}, t \in [1, T] \quad (4.1)$$

$$\vec{h}_{it} = \overrightarrow{GRU}(x_{it}), t \in [1, T] \quad (4.2)$$

$$\overleftarrow{h}_{it} = \overleftarrow{GRU}(x_{it}), t \in [T, 1] \quad (4.3)$$

The word annotation h_{it} is fed into tanh function with weight and bias as in Eq. (4.4) to get u_{it} as improved hidden annotation/representation of h_{it} . Improved word annotation u_{it} and word context vector u_w are utilized to get normalized importance weight α_{it} through a SoftMax function, which finally gives sentence vector s_i according to Eq. (4.6). u_w is a trainable word context weight vector that is used to measure the importance of the word. It is randomly initialized and jointly learned during the training process.

$$u_{it} = \tanh(W_w h_{it} + b_w) \quad (4.4)$$

$$\alpha_{it} = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)} \quad (4.5)$$

$$s_i = \sum_t \alpha_{it} h_{it} \quad (4.6)$$

Bidirectional GRU is utilized to encode the sentences and concatenates forward and backward hidden states to get $h_i = [\vec{h}_i, \overleftarrow{h}_i]$, which focuses considerably on sentence i , at the same time summarizes the neighbouring sentences also.

$$\vec{h}_i = \overrightarrow{GRU}(s_i), i \in [1, L] \quad (4.7)$$

$$\overleftarrow{h}_i = \overleftarrow{GRU}(s_i), i \in [L, 1] \quad (4.8)$$

The same attention mechanism is again used at the sentence level to reward sentences that are pieces of evidence to appropriately classify a fake news article. Improved sentence annotation u_i and normalized importance of weight α_i are calculated for each sentence to get the overall high-level news vector v which represents a complete news article (headline concatenated with the body). u_s is a trainable sentence context weight vector that is used to measure the importance of the sentence. It is initialized randomly and jointly learned during the training process.

$$u_i = \tanh(W_s h_i + b_s) \quad (4.9)$$

$$\alpha_i = \frac{\exp(u_i^T u_s)}{\sum_i \exp(u_i^T u_s)} \quad (4.10)$$

$$v = \sum_i \alpha_i h_i \quad (4.11)$$

Each sentence is padded or truncated to make 32 as an average word count. The average sentence count in each article is 40. Hyperparameter tuning is done using a validation set. GloVe word embedding and GRU annotations have 100 and 50 dimensions respectively which gives an overall 100-dimensional word/sentence annotation after being processed using Bidirectional GRU. The network is being trained using a batch size of 32 till 40 epochs, drop out regularization probability is set to 0.5, activation function is ‘relu’, loss is calculated using binary cross-entropy and adam is used as an optimizer. Finally, softmax is applied at the final stage to get the binary classification probabilities as real or fake.

4.2.3 Image Caption and Headline Matching with News Text (CHM)

Appealing headlines and images are posted on social media to increase user visits to the website but for fabricated articles, the text contents of the news have no relevance with the title and image contents. This type of forgery is also called “Clickbait”. Automatic caption generation for an image is known as Image Summarization. It includes a mechanism that takes the image as the input and generates a suitable caption for that image by looking at the objects and components of the image. In our work, we use Microsoft’s CaptionBot for automatic caption generation. The bot, from Microsoft’s Cognitive Services team, is the result of hefty research into how to model objects in photographs so that a computer can understand them. Their system can recognize “a broad range of visual concepts” and also performs entity extraction so that it can recognize celebrities. It incorporates three separate services to process the images: Computer Vision API, Bing Image Search API and Emotion API. The Computer Vision API explores the components of the photo, mixing them with the data from the Bing Image Search API, and runs it over any faces it spots through their Emotion API. This analyses human facial expressions to detect anger, fear, contempt, happiness, sadness, surprise or disgust.

Image caption generated using CaptionBot is pre-processed and tokenized. News Headlines and News Text are also pre-processed and converted into tokens. Glove word embedding is used to convert the text tokens into word vectors. Now caption and headline tokens are matched for semantic similarity with news content according to Algorithm 1. If

these have a similar context and are talking about the same thing, the chance is good that the news is real. A counter is maintained to count the number of semantically similar words according to calculated cosine similarity. If this count is greater than 15% of the total number of tokens in news text, the article is categorized as real. If the matching fails, it signifies that the image and the headline have no significance with the news content; hence it will be classified as fake news. Figure 4.4 highlights a few sample images from the dataset and their automatically generated captions. Cosine similarity between two different n-dimensional vectors \vec{a} and \vec{b} in n-dimensional space can be calculated by dividing the dot product of two vectors by the product of their magnitudes according to Eq. (4.12). We have represented every text token into a 100-dimensional vector space using GloVe word embedding.

$$\text{cosine similarity} = \cos \theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2} \sqrt{\sum_1^n b_i^2}} \quad (4.12)$$

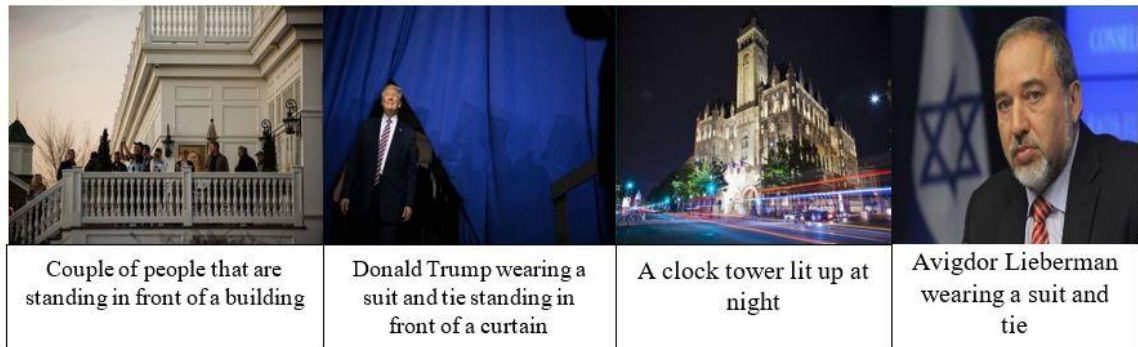


Figure 4.4: Few Examples of generated Image Captions

Algorithm 4.1 Caption and Headline Matching with News Text

```

1: Image Caption=CaptionBot(ImageURL)
2: Token1=Tokenize (Image Caption)
3: Token2=Tokenize (News Headline)
4: SearchStr=Concatenate (Token1, Token2)
5: Text=Tokenize (News Body Content)
6: counter= 0
7: For each token in SearchStr
8:     For each token in Text
9:         Calculate (Cosine_similarity)
10:        if (Cosine_Similarity>0.75)
11:            Counter+=1
12: if counter>=0.15(length(Text))
13:     return "Classified as Real News"

```

4.2.4 Noise Variance Inconsistency (NVI)

Instead of changing the complete image, the prevalent visual forgery format is to crop, alter or superimpose a portion of the image to spread the intended fake notion about the event. Parts of the original image when cropped, altered or superimposed will create noise inconsistencies contrary to the uniformly distributed random noise in the actual image. The sub-block image noise consistency analysis can reveal the traces of doctoring by using Noise Variance Inconsistency. Noise is the most common tool to hide the footprints of tampering in images. A typical practice is to disguise the hints of altering by adding arbitrary local noise to the forged image areas. Such types of forgery can be detected by identifying noise inconsistencies in the images. Image originality is negatively affected by added extra noise. The random inherent image noise is uniformly distributed throughout but intentionally added extra local noise in a specific part of the image causes inconsistencies in noise variance. So, it turns out to be an effective way to figure out the manipulated regions in images.

The image is divided into certain blocks to decide which block will move into which cluster, original or tampered to check the consistency of noise in the whole image. We have chosen 32x32 as the block size. A mask is applied to every block to calculate the value of noise in that particular block. K-means clustering ($K=2$) is applied to the calculated noise values to cluster them into two groups. If the centres of the two clusters are almost near, we consider no image forgery as this is the random noise of the image which is uniformly distributed throughout. But if the image is being tampered and noise is being added to hide this tampering, such manipulated blocks will be in another cluster, and the distance between the centres will be higher than the threshold.

Algorithm 4.2 Noise Variance Inconsistency

- 1: Open the image
 - 2: Break the image into blocks of size 32x32
 - 3: For every block
 - 4: Estimate noise value using a mask
 - 5: Apply K-means clustering on calculated noise values with $K=2$
 - 6: Centre1=Centre of the first cluster
 - 7: Centre2=Centre of the second cluster
 - 8: if $(\text{Centre1}-\text{Centre2})>0.7$
 - 9: return "Classified as Tampered Image"
-

4.2.5 Error Level Analysis (ELA)

There are a lot of advanced photo editing tools available, which are perfect enough to deceive human eyes. Although after doctoring the image, the characteristics are changed, but it looks almost similar to the originally captured image. Therefore, the human visual system is not capable enough to detect the traces of modifications simply by looking at the picture. In the process of manipulating the image it is being resaved multiple times or at least once, this is being captured by analysing the quality of the resaved image using Error Level Analysis. Figure 4.5 (a) represents the original digital photograph of books arranged on a shelf and Figure 4.5 (c) is the doctored version of the original image in which a toy dinosaur is added, and books are copied. Both the images look perfect in quality as far as the human vision system is considered. But the difference could be identified by using Error Level Analysis of the original and modified image as characterized by Figure 4.5 (b) and Figure 4.5 (d) respectively. Figure 4.5 (b) is having higher pixel values represented by white colour and has high Error Level Analysis values. Contrary to this Figure 4.5 (d) which represents Error Level Analysis of the modified image is darker, black, having very low pixel values and traces of modified portion can also be identified by looking at it.

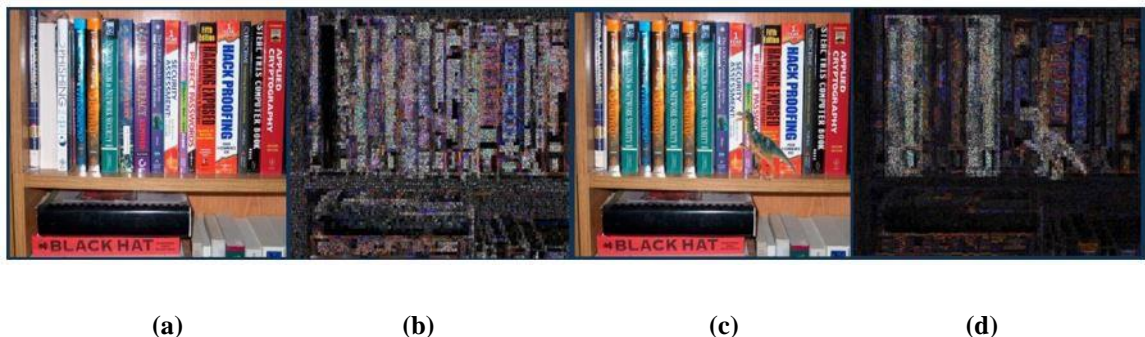


Figure 4.5: (a) Original photograph (b) Error Level Analysis of Original photograph (c) Modified and resaved photograph (d) Error Level Analysis of modified and resaved photograph [194]

Error Level Analysis is a forensic method used to determine if a picture has been digitally modified. Original image when doctored or photoshopped is already being resaved many times or at least one time. So, the quality of pixels in such types of images has already been degraded, and they are at their local minima. Manipulated image when resaved at 95% quality level does not create much difference in pixel values. Whereas when

an original image is resaved at a known lower quality level, it generates a significant difference in the pixel values. This difference serves as the identification parameter, whether the image is Real or Fake as described in Algorithm 3.

Algorithm 4.3 Error Level Analysis

- 1: Im=Original Image
 - 2: Resaved_Im=Resave Original Image with 95% quality
 - 3: Diff_Im=Calculate difference (Im,Resaved_Im)
 - 4: Extrema=Calculate minima and maxima for each band(R,G,B) in Diff_Im
 - 5: Max_diff=Get max(Extrema)
 - 6: if (Max_diff >=5)
 - 7: return “Classified as Real Image”
-

4.2.6 Ensemble with Max Voting

Ensemble learning is used to improve the performance of architecture by aggregating its different components together to achieve a common goal. There are many ensemble models such as: bagging, boosting, stacking, voting etc. To ensemble the predictions obtained from individual techniques, we use the max voting method. This method classifies a news instance as true if more than half of the base techniques predict it as true otherwise false. If two predictions are real, and two predictions are fake, the overall prediction is fake. We have two classes real and fake, so class variable $j \in \{0,1\}$ and 4 different predictions to ensemble $t \in \{1,2,3,4\}$. The decision of t^{th} technique is $d_{t,j} \in \{0,1\}$. For class j the sum $\sum_{t=1}^4 d_{t,j}$, tabulates the total number of votes for class j . Max voting chooses the class j , which maximizes the sum i.e. more than half according to Eq. (4.13).

$$J = \operatorname{argmax}_{j \in \{0,1\}} \sum_{t=1}^4 d_{t,j} \tag{4.13}$$

4.2.7 Implementation Details

To validate the framework proposed in section 3, we have done implementations in python 3.7 programming language on Windows 10 operating system using online GPU google colab. The deep learning model Hierarchical Attention Network is implemented using Keras library on top of TensorFlow architecture. We implemented all four modules first independently and then collectively using ensemble max voting on three different datasets. The performance of the work is evaluated in terms of precision, accuracy, recall and F1

score. For HAN and overall ensemble model evaluation, we have used 80% data for training, 10% for validation and 10% for testing. To test the individual performance of image caption and headline matching with news text module, Noise Variance Inconsistency and Error Level Analysis, we have used the complete datasets as these three are the rule-based techniques.

We used three different datasets for validating the proposed framework. Out of the three datasets, only one dataset ‘All Data’ has multimedia data in the form of image URLs in it, for the remaining two we scrapped the images of the corresponding news headline and body from the webpage using site URLs which was a parameter already given in the datasets using Beautiful Soup Python Library. Table 4.1 summarizes the specifics of the datasets.

Table 4.1: Dataset Details

Dataset Name	Details	Total entries	Fake news	Real news
Fake News Detection by Jruvika [176]	Hosted on Kaggle, contains site URL, Headline, Body and Label	3988	2121	1867
All Data [195]	Dataset contains 54 attributes including site URL, Headline, Body, Image URL, Label	20 015	11941	8074
Fake News Sample by Guilherme Pontes [178]	Hosted on Kaggle, contains 17 attributes including site URL, Headline, Body and Label	45569	20226	25343

Fake News Detection by Jruvika [176] is uploaded on the Kaggle website by Jruvika. It initially contains four attributes site URL, Headline, Body and Label (Real/Fake). The dataset originally contains 4009 news instances, after pre-processing the text part, i.e. after removing duplicate entries, missing/null entries and initial data cleaning we have 3988 rows with 2121 Fake and 1867 Real news instances. This dataset does not have images of the news. So, we use the Beautiful Soup Python library to scrap the corresponding images from site URLs, and if the URL is corrupted, we use the headline of the news to scrap the corresponding image. Now we have a fifth column in our dataset corresponding to image URLs.

All Data [195] dataset contains 54 different attributes including site URL, Headline, Body, Label (Real/Fake), image URL, title length, text length, sentiment word count and other metadata. We extracted five columns that are useful for us, apart from this some of

the image URLs are corrupted, so we use the site URL and Headline of news to scrape the image using the Beautiful Soup Python module. Now we have 20 015 total rows in the dataset out of which 11 941 corresponding to fake and 8074 corresponding to real news instances.

Fake News Sample by Guilherme Pontes [178], this dataset is hosted on Kaggle by Guilherme Pontes and contains news articles labelled into different categories hate, satire, clickbait, political, conspiracy, fake, reliable, rumour, unreliable etc. and has 17 different attributes of each news. We filtered rows with fake, rumour, unreliable into the Fake news category and reliable into the real news category. Images are scrapped using site URL and headline. After initial filtering, cleaning and pre-processing we have a total of 45 569 rows with five fields site URL, Headline, Body, image URL and Label (Real/Fake), out of which 25 343 are real news and 20 226 are fake news articles.

4.2.8 Result Analysis

The experimental performance evaluation of our framework on Fake News Detection Dataset in terms of accuracy, precision, recall and F1-score of all the four proposed methods individually and jointly using ensemble with max voting is detailed in Table 4.2. Figure 4.6(a) represents the comparison in terms of bar chart, Figure 4.6(b) highlights the confusing matrix values of the test samples. The area under the curve value is 0.93, which depicts the robustness of the method, as shown in Figure 4.6(c). It is evident from the results that the ensemble model outperforms individual models as it incorporates the best of all the models and has the mechanism to counter text and multimedia forgeries. Hence, it gives the best accuracy of 94.74 % on Fake News Detection Dataset.

Table 4.2: Result Analysis on Fake News Detection Dataset

Feature	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
HAN (Text)	91.48	92.39	89.47	90.91
Caption and Headline matching (Text + Image)	77.66	68.48	96.84	80.23
Noise Variance Inconsistency (Image)	81.47	72.49	97.37	83.11
Error Level Analysis (Image)	84.75	75.83	98.98	85.87
Ensemble with Max Voting (Text + Image)	94.74	95.68	93.16	94.40

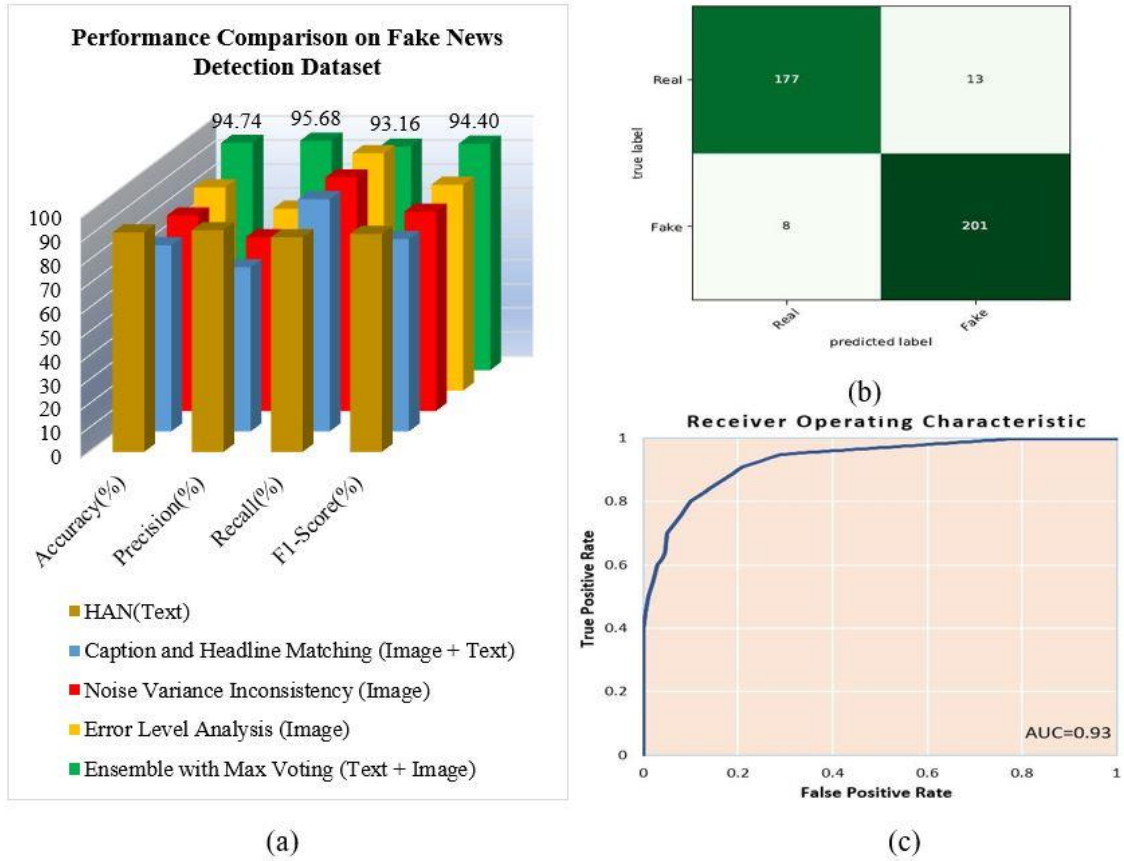


Figure 4.6: (a) Performance Comparison of individual and ensemble model (b) Confusion Matrix (c) ROC curve for Fake news detection dataset

The highest accuracy, precision, recall and F1 Score on All Data dataset is 95.50%, 94.53 %, 94.43% and 94.48% respectively, using the ensemble method as represented in Table 4.3. Statistical performance comparison of individual and ensemble model is made in Figure 4.7 (a) using bar chart. Confusion matrix values of test data and ROC-AUC curve for All Data dataset are shown in Figure 4.7 (b) and (c), respectively.

Table 4.3: Result Analysis on All Data Dataset

Feature	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
HAN (Text)	92.85	89.81	92.81	91.28
Caption and Headline matching (Text + Image)	79.48	70.12	85.60	77.09
Noise Variance Inconsistency (Image)	79.93	67.03	98.84	79.88
Error Level Analysis (Image)	83.44	71.62	97.63	82.63
Ensemble with Max Voting (Text + Image)	95.50	94.53	94.43	94.48

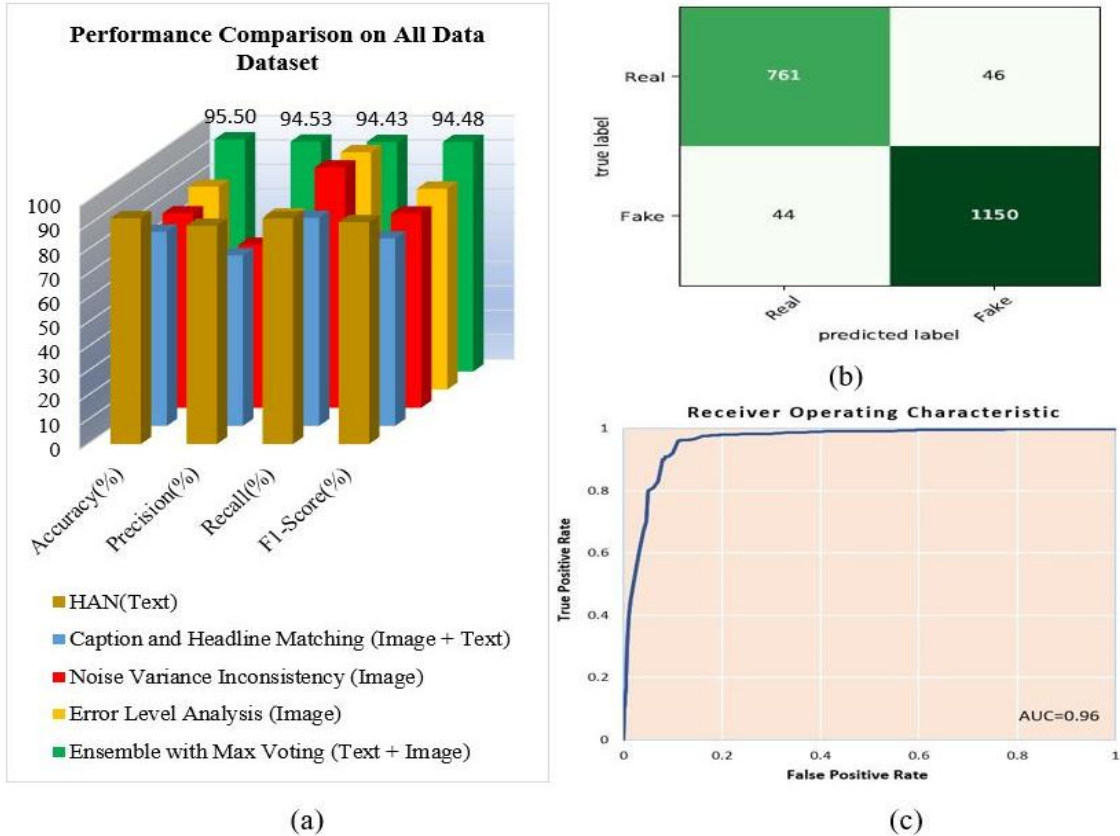
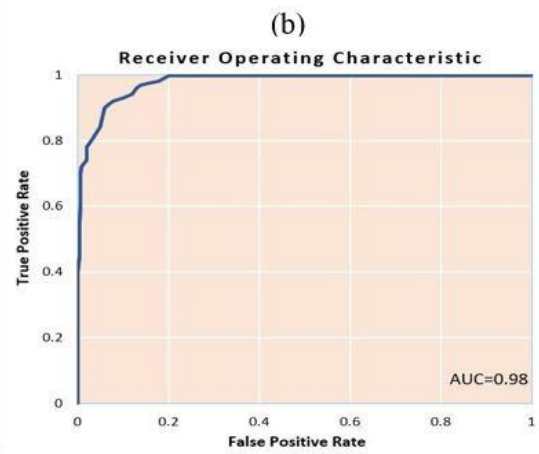
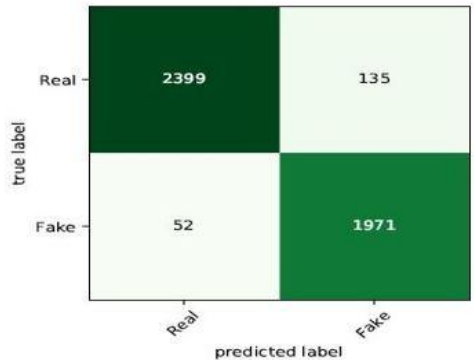
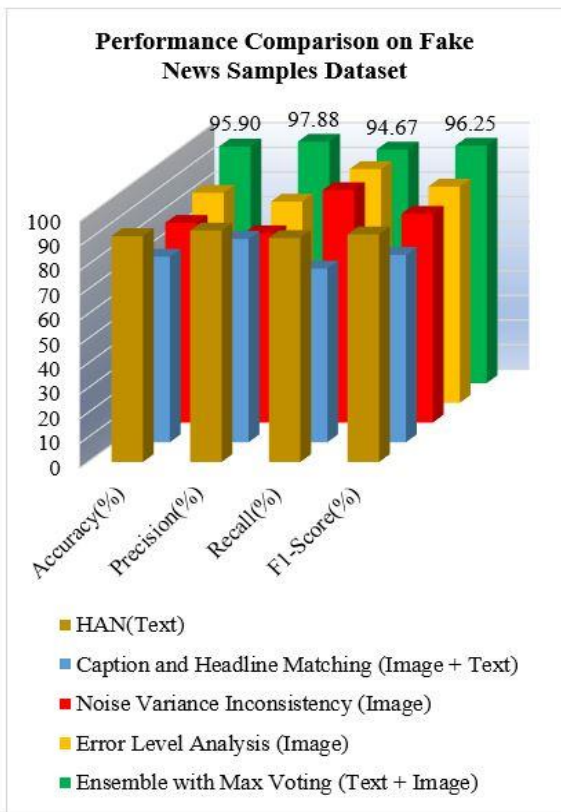


Figure 4.7: (a) Performance Comparison of individual and ensemble model (b) Confusion Matrix (c) ROC curve for All Data dataset

Table 4.4 analysis the performance metrics on the Fake News Samples dataset, highlights the fact that ensemble architecture gives the highest accuracy of 95.90 % on text and multimedia data. Figure 4.8 (a), (b) and (c) concentrate on the pictorial representation of the performance comparison bar chart, confusion matrix and ROC-AUC curve. Finally, the overall performance comparison on all the three datasets based on five comparison parameters is done in Figure 4.9.

Table 4.4: Result Analysis on Fake News Samples Dataset

Feature	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
HAN (Text)	91.57	93.84	90.80	92.29
Caption and Headline matching (Text +Image)	75.18	82.42	70.39	75.93
Noise Variance Inconsistency (Image)	80.95	76.78	94.24	84.61
Error Level Analysis (Image)	85.17	81.57	94.75	87.67
Ensemble with Max Voting (Text + Image)	95.90	97.88	94.67	96.25



(a)

(c)

Figure 4.8: (a) Performance Comparison of individual and ensemble model (b) Confusion Matrix (c) ROC curve for Fake News Samples dataset

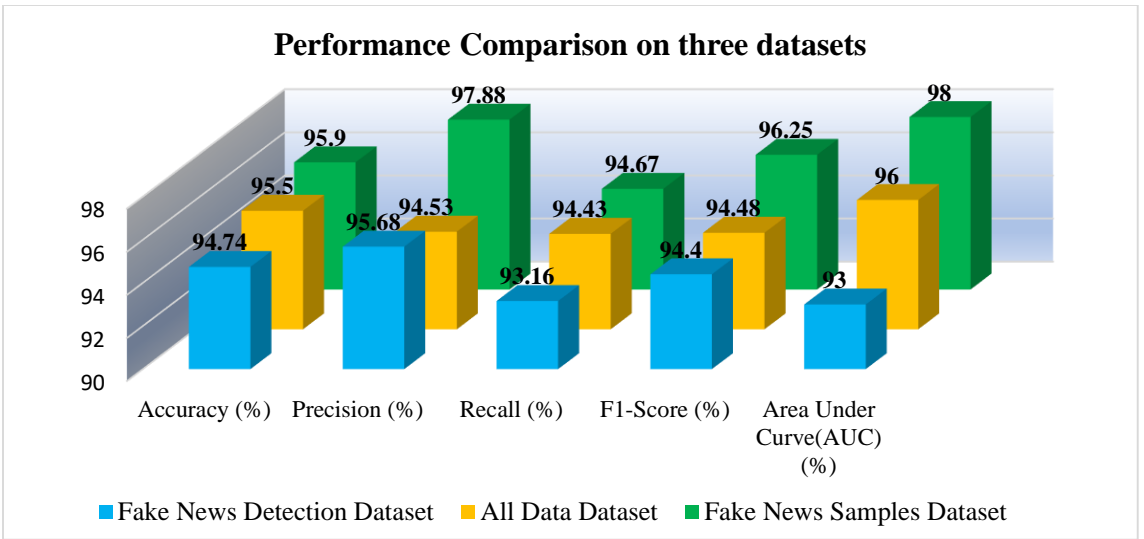


Figure 4.9: Overall Performance Comparison of three datasets

The above discussion supported with Tables 4.2, 4.3, and 4.4 along with Figures 4.6 ,4.7, 4.8 and 4.9 verifies that our proposed ensemble architecture which is a hybrid of four modules achieves promising performance and outstanding accuracies for the Fake News Classification task. The following are the main reasons accounting for such an exemplary performance by the proposed framework:

- The headline, body text and multimedia are the three main components of widely circulated news instances on web platforms. Multimedia in the form of images is the most circulated, visualized, analysed information format as it is more fascinating and exerts a long-lasting effect in human memory. So, the main reason for achieving promising results is that we have considered all possible types of forgery formats in textual and visual data including HAN deep model for textual information, clickbaits, image captioning, out of the context and doctored visual information and finally ensembled the aggregated results from multiple streams.
- Stylometric and content-specific hidden patterns of words and sentences in fake news writing style have been analysed using the deep architecture of Hierarchical Attention Network (HAN).
- The news accompanied by images of the wrong context (but not tampered) and appealing false headlines (clickbait) has been taken care of using Image Captioning and semantic matching of the caption as well as the headline with the news body text.
- Instead of changing the complete image, the prevalent visual forgery format is to crop, alter or superimpose a portion of the image with the help of available photo editing tools to spread the intended fake notion about the event. Once the picture is manipulated the false event is very well narrated in text and it becomes extremely difficult to find out this forgery even by using advanced deep learning techniques if visual information is not being considered. So, image forensic techniques of Noise Variance Inconsistency (NVI) and Error Level Analysis (ELA) have been used for achieving higher accuracies in the identification of manipulated (photoshopped) images.

- The successful, outperforming and efficient working of the designed architecture can also be justified with the fact that it a holistic system that is capable enough of addressing a wide range of fake formats in text and visual information prevalent on web platforms. The framework collectively addresses different forgery formats of the fake writing style of the text, images with wrong context and doctored images into a single multimodal binary classification system.

4.2.9 Ablation Study

Ablation study is the process of systematically analysing a framework in the presence and absence of a particular component by removing and aggregating them one by one. It helps in the optimization of the system architecture by determining the bottleneck as well as redundant components. To illustrate the contribution and effectiveness of each one of the individual modules Hierarchical Attention Network (HAN), Image Caption and Headline matching with News Text (CHM), Noise Variance Inconsistency (NVI) and Error Level Analysis (ELA) of the proposed ensemble architecture in final classification accuracy we performed ablation study. Individual techniques, the ensemble of two and ensemble of three modules are being experimented on Fake News Detection, All Data and Fake News Samples datasets with identical parameter settings as the overall proposed framework. Following the discussions in section 4.2.6 and Eq. (4.13) in the process of ensembling, a news instance is classified as real/true if more than half of the base techniques predict it as real/true otherwise false/fake. The accuracies and analytical elucidations are detailed in Table 4.5.

Table 4.5: Ablation Study of proposed Ensemble Multimodal Framework

Feature	Accuracy (%)			Analysis
	Fake News Detection Dataset	All Data Dataset	Fake News Samples Dataset	
<i>Individual Techniques</i>				
Hierarchical Attention Network (HAN)	91.48	92.85	91.57	Classification accuracies are exhibiting independent strength of the classifiers to categorize the news instances as “Real” or “Fake” based on particular feature engineering
Caption and Headline matching (CHM)	77.66	79.48	75.18	
Noise Variance Inconsistency (NVI)	81.47	79.93	80.95	

Error Level Analysis (ELA)	84.75	83.44	85.17	
<i>Ensemble of Two Techniques</i>				
HAN+CHM	76.69	77.27	75.02	Less accuracy because according to eq. (13) to classify a news instance as “Real” both techniques must categorize it as “Real”
HAN+NVI	79.00	79.47	79.34	
HAN+ELA	83.60	82.49	85.00	
<i>Ensemble of Three Techniques</i>				
HAN+CHM+NVI	86.90	85.78	84.10	Improved accuracy as the number of classifiers has increased, better utilization of textual and visual features, final classification is determined by more than half i.e. any two identical categories
HAN+CHM+ELA	87.22	88.50	87.89	
HAN+NVI+ELA	89.48	91.21	90.28	
<i>Overall Ensemble Multimodal Framework</i>				
HAN+CHM+NVI+ELA	94.74	95.50	95.90	Highest accuracies because the overall framework combines advantages of all techniques and detects an extensive category of textual and visual features

4.2.10 Parameter Analysis for Setting Thresholds

In rule-based algorithms parameter analysis is the process of determining a particular threshold value for boundary conditions. The threshold values used in Algorithm 1, 2 and 3 described in the proposed model section are being determined by extensive experimental analysis on three different datasets. Image caption and headline tokens are matched for semantic similarity with news content according to Algorithm 1 described in section 4.2.3. If these have a similar context and are talking about the same event, the chance is good that the news is real. The cosine similarity index is used for calculating the semantic similarity between words. The counter is a variable to count the semantically similar words. If this count is greater than a threshold value i.e., 15% of the total number of tokens in news text, the article is categorized as real. The thresholds for these two parameters are set after experimenting and analysing various combinations for all the three datasets as described in Table 4.6.

Noise Variance Inconsistency (NVI) analysis of the visual data is performed according to Algorithm 2 detailed in section 4.2.4. The subblock image noise is calculated and the blocks are divided into two clusters based on the calculated noise value. Less distance between cluster centres represents no image forgery. If the image is tampered, manipulated blocks will be in another cluster and the distance between the centres will be

higher than the threshold. Table 4.7 details the implementation and analysis that are done for setting the value of the threshold between cluster centres.

Table 4.6: Parameter Analysis for setting thresholds in Algorithm 1

Thresholds	Accuracy (%)			Analysis
	Fake News Detection Dataset	All Data Dataset	Fake News Samples Dataset	
Cosine Similarity>0.25 & counter>= 0.20	59.62	60.15	58.03	The accuracy trends analyzed for different datasets proves that the best performance of the method is achieved at thresholds Cosine Similarity>0.75 & counter>=0.15
Cosine Similarity> 0.55 & counter>= 0.15	71.22	69.67	71.32	
Cosine Similarity>0.75 & counter>=0.15	77.66	79.48	75.18	
Cosine Similarity>0.85 & counter>= 0.10	59.02	62.37	60.01	

Table 4.7: Parameter Analysis for setting thresholds in Algorithm 2

Thresholds	Accuracy (%)			Analysis
	Fake News Detection Dataset	All Data Dataset	Fake News Samples Dataset	
(Centre1-Centre2)>0.5	75.27	74.62	73.33	The technique performs best for all datasets when (Centre1-Centre2)>0.7 is taken as the threshold for classifying an image as Fake/Tampered
(Centre1-Centre2)>0.6	80.88	78.34	76.84	
(Centre1-Centre2)>0.7	81.47	79.93	80.95	
(Centre1-Centre2)>0.8	80.02	79.00	78.62	

Section 4.2.5 is furnished with the discussion of Error Level Analysis (ELA) image forensics method. The default setting of most photo editing tools provides the quality between 90%-95% upon resaving the image once. It means that 95% is the maximum quality value that an image can attain after first resave. Thus, we took 95% error level for resaving the images. The difference between the pixels of the original image and the resaved image serves as the identification parameter for classifying an image as Real or Fake according to Algorithm 3. The following Table 4.8 illustrates the variation in accuracies upon changing this threshold for three datasets.

Table 4.8: Parameter Analysis for setting thresholds in Algorithm 3

Thresholds	Accuracy (%)			Analysis
	Fake News Detection Dataset	All Data Dataset	Fake News Samples Dataset	
Max_diff >=3	77.69	75.00	76.54	Best performance of the method is achieved at Max_diff >=5 threshold. It signifies that if the maximum pixel difference between original and resaved image (at 95% quality) is 5 or more; the image is not tampered.
Max_diff >=4	82.00	82.80	81.29	
Max_diff >=5	84.75	83.44	85.17	
Max_diff >=7	83.67	81.24	83.28	

4.2.11 State-of-the-art Comparison

To compare the effectiveness of our proposed framework with contemporary techniques, we implemented the baseline methods with all the parameter settings as written in the research paper for the datasets that we have used in our work. The performance of each baseline is calculated in terms of accuracy, precision, recall and F1 score with the dataset split as 80% training, 10 % validation and 10% testing. State-of-the-art comparison on each of the Fake News Detection, All Data and Fake News Samples datasets are described in Table 4.9, Table 4.10 and Table 4.11, respectively. The following methods are used as baselines for state-of-the-art comparison:

- Ajao et al. [182] proposed hybrid CNN and RNN deep models for veracity analysis of the text.
- Agarwalla et al. [180] implemented Logistic Regression, SVM, NB with Lidstone Smoothing on the Body and Headline of news articles.
- Vishwakarma et al. [124] proposed a method of extracting text from images then search this text as well as the headline on the web and calculate a reality parameter (Rp) by checking the credibility of the top 15 Google search results.
- Event Adversarial Neural Network (EANN) devised by Wang et al. [120] used Text CNN and VGG 19 to extract multimodal features from text and images. For the sake of symmetry in comparisons, we remove the event discriminator from EANN and implement only binary fake/real classifiers on our datasets.
- To fuse textual and visual features, a recurrent neural network with an attention mechanism (att. RNN) proposed by Jin et al. [196] is used which concatenates the

features extracted by LSTM and VGG 19 for effective fake news detection. In our experiments for a fair comparison, all the experimental settings are identical to the base paper except the part dealing with social context information.

- A framework for detecting visual forgery from tampered and misleading images is proposed by Qi et al. [63], the authors utilized frequency domain and pixel domain information for training CNN and GRU deep models. We implemented this Multi-domain Visual Neural Network (MVNN) model on images of our three datasets independently and highlighted the results in performance comparison metrics.
- Lago et al. [197] experimented with textual and visual data for fake news detection using classical as well as advanced techniques. TF-IDF, cosine similarity and Jaccard’s similarity are used for text analysis; Classical image forensic methods such as JPEG ghosts, Color filter array, Mean-filter noise residue inconsistency, Block artifact grid detection as well as Splicebuster an advanced image splicing detector are used for image analysis using Random Forest, Logistic regression and CNN classifier to design a binary classification framework.
- Khattar et al. [121] experimented with multimodal data using a variational autoencoder system to learn probabilistic latent variables. Encoder, Decoder and a binary fake news detection component are the main parts of the proposed framework using Bi-LSTM and VGG19 techniques for text and image feature engineering respectively.

Table 4.9: Performance comparison with state-of-the-arts on Fake News Detection dataset

Method	Input data format	Proposed Feature/Classifier	Acc. (%)	P (%)	R (%)	F1 (%)
Ajao et al. [182]	Text	Hybrid CNN and RNN	86.00	82.37	85.66	83.98
Agarwalla et al. [180]	Text	Logistic Regression, SVM, NB with Lidstone Smoothing	83.87	81.22	82.67	81.93
Vishwakarma et al. [124]	Text + Image	Reality Parameter, Rule-based classifier	85.30	85.20	88.40	86.77
Wang et al. [120] (EANN)	Text + Image	CNN, VGG19	83.20	85.60	82.20	83.86
Jin et al. [196] (att. RNN)	Text + image	LSTM, VGG 19	80.75	79.33	83.17	81.20
Qi et al. [63] (MVNN)	Image	Frequency and pixel domain features, CNN, GRU	92.91	91.05	92.78	91.90
Lago et al. [197]	Text + Image	TF-IDF, cosine similarity, Jaccard’s similarity, Classical and	90.26	90.65	89.43	90.03

		advanced image forensics, Random Forest, Logistic Regression, CNN				
Khattar et al. [121] (MVAE)	Text + Image	BiLSTM, VGG19, Encoder, decoder and Fake News Detector	84.72	83.60	86.47	85.01
Ensemble with max voting	Text +Image	Ensemble of HAN, semantic matching, NVI, ELA	94.74	95.68	93.16	94.40

Table 4.10: Performance comparison with state-of-the-arts on All Data Dataset

Method	Input Data Format	Proposed Feature /Classifier	Acc. (%)	P (%)	R (%)	F (%)
Ajao et al. [182]	Text	Hybrid CNN and RNN	84.80	88.72	81.80	85.11
Agarwalla et al. [180]	Text	Logistic Regression, SVM, NB with Lidstone Smoothing	83.01	85.05	79.66	82.27
Vishwakarma et al. [124]	Text +Image	Reality Parameter, Rule based classifier	88.00	87.90	88.80	88.34
Wang et al. [120] (EANN)	Text +Image	CNN, VGG19	84.01	86.73	83.27	84.96
Jin et al. [196] (att. RNN)	Text +image	LSTM, VGG 19	79.78	87.20	79.60	83.22
Qi et al. [63] (MVNN)	Image	Frequency and pixel domain features, CNN, GRU	91.62	93.27	91.76	92.50
Lago et al. [197]	Text + Image	TF-IDF, cosine similarity, Jaccard's similarity, Classical and advanced image forensics, Random Forest, Logistic Regression, CNN	92.13	93.05	94.07	93.56
Khattar et al. [121] (MVAE)	Text + Image	BiLSTM, VGG19, Encoder, decoder and Fake News Detector	85.70	87.80	83.02	85.34
Ensemble with max voting	Text + Image	Ensemble of HAN, semantic matching, NVI, ELA	95.50	94.53	94.43	94.48

Table 4.11: Performance comparison with state-of-the-arts on Fake News Samples Dataset

Method	Input data format	Proposed feature/Classifier	Acc. (%)	P (%)	R (%)	F (%)
Ajao et al. [182]	Text	Hybrid CNN and RNN	83.22	79.02	80.66	79.83
Agarwalla et al. [180]	Text	Logistic Regression, SVM, NB with Lidstone Smoothing	82.66	79.01	81.20	80.09
Vishwakarma et al. [124]	Text + Image	Reality Parameter, Rule based classifier	87.11	86.08	89.16	87.59
Wang et al. [120] (EANN)	Text + Image	CNN, VGG19	83.32	84.37	81.45	82.88
Jin et al. [196] (att. RNN)	Text + Image	LSTM, VGG 19	82.39	85.30	82.61	83.93
Qi et al. [63] (MVNN)	Image	Frequency and pixel domain features, CNN, GRU	86.07	82.32	85.60	83.93
Lago et al. [197]	Text + Image	TF-IDF, cosine similarity, Jaccard's similarity, Classical	92.00	92.20	92.77	92.48

		and advanced image forensics, Random Forest, Logistic Regression, CNN				
Khattar et al. [121] (MVAE)	Text + Image	BiLSTM, VGG19, Encoder, decoder and Fake News Detector	88.45	87.04	90.22	88.60
Ensemble with max voting	Text+ Image	Ensemble of HAN, semantic matching, NVI, ELA	95.90	97.88	94.67	96.25

The above experimental results show that the proposed model for fake news detection on text and multimedia data is quite satisfactory. The accuracy over all the three datasets Fake News Detection, All Data and Fake News Samples is a meaningful improvement over the state-of-the-art. Image forensics, HAN deep learning and semantic caption and headline matching with the body of the news are the salient features that have helped in improving the accuracy of our results. Finally combining all the independent techniques to exploit the forgeries in text and multimedia data format helps us in achieving 95.90% highest detection accuracy.

4.2.12 Significant Outcomes

The significant outcomes of the proposed multimodal fake news detection framework are as under:

- The problem of online fake news in the multimodal format of text with wrong context images and tampered images is being identified.
- A novel framework is proposed to incorporate holistic fake news detection of all the modalities (text and images) and different forgery formats (Fake writing style of the text, images with wrong context and doctored images).
- Stylometric and content-specific hidden patterns of words and sentences in fake news text have been extracted using the deep architecture of the Hierarchical Attention Network, which is being trained and tested for different datasets.
- The news accompanied with images of the wrong context (but not tampered) and appealing headlines (clickbait) has been taken care of using Image Captioning and semantic matching of the caption as well as the headline with the news body text.
- Image forensic techniques of Noise Variance Inconsistency and Error Level Analysis have been used for the identification of manipulated (photoshopped) images.

- Textual and visual hybrid methods are ensembled using the max voting technique to classify a news instance as fake or real.
- Results of all the four modules (HAN, image caption and headline matching with news body, Noise Variance Inconsistency, Error Level Analysis) on three different datasets have been tested independently and then jointly using max voting ensemble technique.
- Ablation study to analyse the effect of each technique independently as well as collectively and parameter analysis for setting thresholds are also discussed.
- State-of-the-art comparisons on the same datasets under identical experimental conditions have also been highlighted.

4.3 Multi-modal Fusion Using Fine-tuned Self-attention and Transfer Learning for Veracity Analysis of Web Information

It has become significant to incorporate images, as nearly all news, fake or real, is accompanied by some visual data. A combination of Deep Learning Neural Network Models is used for veracity analysis of multi-modal data. We have used two modalities of textual and visual features in conjunction to classify a piece of news as fake or real. We use implicit features obtained from pre-trained models to train our custom neural network to obtain the final predictions using two independent architectures of Late Fusion and Early Fusion. We have used Inception-ResNet-v2 to extract visual features. The models BERT and ALBERT have been used to elicit textual attributes. Diverse forms of text input, like English articles, Chinese articles and Tweets have been used to make our model robust and usable across multiple platforms. The architecture of Multimodal Early Fusion and Late Fusion has been detailed in Figures 4.10 and 4.11 as well as the working steps are described in Algorithm 1 and 2 respectively.

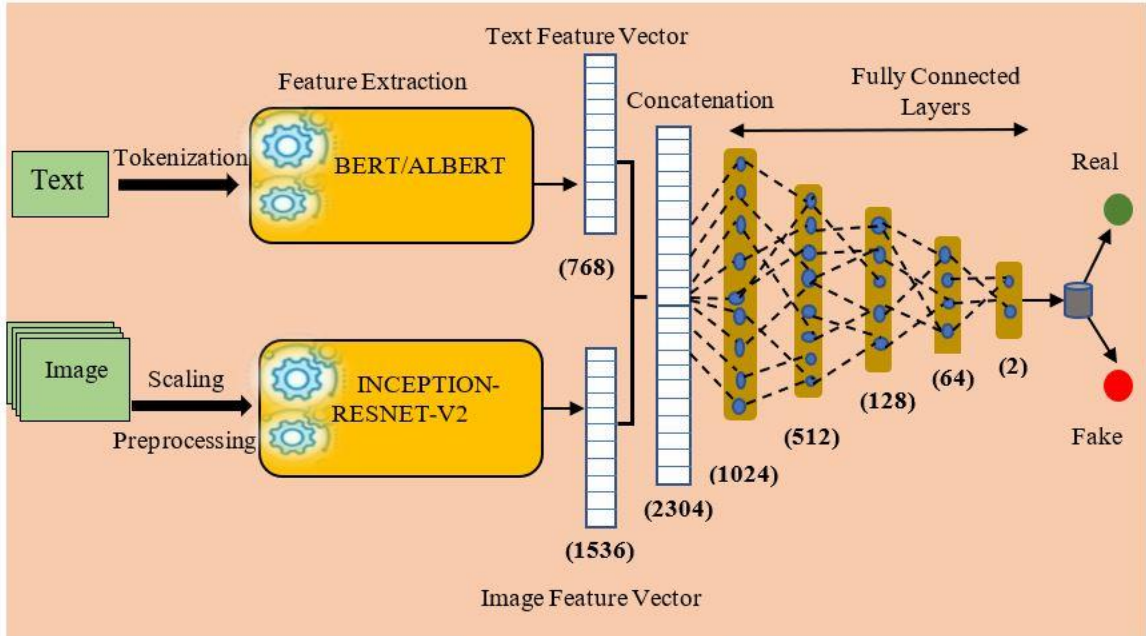


Figure 4.10: Architecture of Multimodal Early fusion

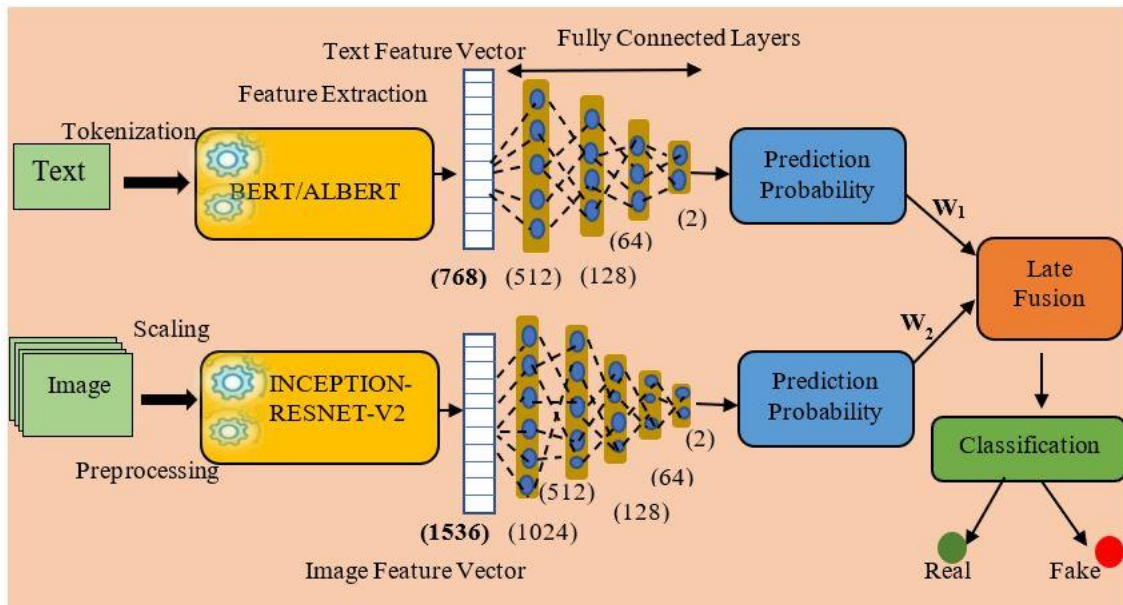


Figure 4.11: Architecture of Multimodal Late fusion

Algorithm 4.4: Multimodal Early Fusion

Parameter Initialization

- Input $A = \{a_1, a_2, \dots, a_n\}$ is set of text vectors, $B = \{b_1, b_2, \dots, b_n\}$ is set of corresponding images, $C = \{c_1, c_2, \dots, c_n\}$ is set of labels for A and B.
- n is the size of training set
- Split A, B and C into three subsets for 70% training, 10% validation and 20% testing.

1: For 1 to 30 epochs do

2: Extract feature vector using BERT/ALBERT model M_1 for text input A.

-
- 3: Extract feature vector using Inception-ResNet-v2 model M_2 for image input B.
 - 4: Convert feature vectors obtained from M_1 and M_2 into unidimensional.
 - 5: Perform concatenation of feature vectors.
 - 6: Add a series of fully connected layers (dense, batch normalization, relu, dropout=0.4)
 - 7: Apply binary sigmoid classifier and calculate prediction probabilities
 - 8: Apply binary cross-entropy loss and Adam optimizer
 - 9: end for
 - 10: Evaluate the performance on test set
-

Algorithm 4.5: Multimodal Late Fusion

Parameter Initialization

- Input $A = \{a_1, a_2, \dots, a_n\}$ is set of text vectors, $B = \{b_1, b_2, \dots, b_n\}$ is set of corresponding images, $C = \{c_1, c_2, \dots, c_n\}$ is set of labels for A and B.
- n is the size of training set
- Split A, B and C into three subsets for 70% training, 10% validation and 20% testing.

- 1: For 1 to 30 epochs do
 - 2: Extract feature vector using BERT/ALBERT model M_1 for text input A
 - 3: Add fully connected layers (dense, batch normalization, relu, dropout=0.4) to text feature vector
 - 4: Apply binary sigmoid classifier and calculate prediction probability p_1
 - 3: Extract feature vector using Inception-ResNet-v2 model M_2 for image input B
 - 4: Add fully connected layers (dense, batch normalization, relu, dropout=0.4) to image feature vector
 - 5: Apply binary sigmoid classifier and calculate prediction probability p_2
 - 6: Perform final prediction using weighted averaging late fusion ($\sum_{i=1}^2 w_i p_i / \sum_{i=1}^n w_i$)
 - 7: Apply binary cross-entropy loss and Adam optimizer
 - 8: end for
 - 9: Evaluate the performance on test set
-

4.3.1 Pre-processing

The multi-modal veracity analysis framework is implemented on three datasets with textual data and image URLs to access the associated images. URLs were loaded using the Beautiful Soup library to access the images. Some of the images were inaccessible, which subsequently reduced the dataset as the corresponding rows were removed. Any attribute having null was handled by removing the corresponding rows. The dataset after all the mentioned processing was saved in the local drives for later access.

As far as pre-processing is concerned, textual data was not pre-processed as the model used was BERT and ALBERT, which handles all types of words by its huge vocabulary size. The maximum sequence length for the BERT model for All data [195], Weibo [133] and MediaEval 2016 [198] datasets was fixed at 512, 400 and 20 respectively. The maximum sequence length was decided by taking the limit such that 95-98% of the word length comes under its range. For the dataset having tweets, pre-processing was performed by removing emojis, hyperlinks and tweet mentions from the textual part.

Images for every dataset were resized to (299,299) as this is the default size required by the Inception-ResNet-v2 model for feature extraction. All the images for every dataset were first scaled and converted to NumPy arrays. Then, minor pre-processing was done by using the Keras predefined function to obtain final images for input to our model. Features from textual data are extracted by using BERT and ALBERT language models based on the bidirectional encoding property of transformers and visual features are extracted by using the Inception-ResNet-v2 deep neural pre-trained model. After the feature vectors are obtained from both the modalities, separate standard scaling of both is done by using Sci-Kit Learns module "Standard Scaler".

4.3.2 BERT

BERT stands for Bidirectional Encoder Representations from Transformers. It was developed by research scientists working at Google AI Language [199]. It can perform various tasks including NLP, Next Sentence Prediction, Question-Answering, Classification, etc. What sets BERT apart from the other similar models is its ability to apply the bidirectional training property of a Transformer, which is a famous attention model, efficiently to language modeling framework. The Transformer is an attention mechanism that is used to learn the relationship between contexts of different words in text data. Transformers generally comprise two distinct mechanisms —one is an encoder that can read the input text, and another is a decoder that can produce the task prediction. But since BERT is only concerned with generating a language model, the decoder mechanism is not required. BERT is employed for a broad range of language-related tasks. All of them can be successfully performed only by augmenting a small-layer to the standard model. Classification-related tasks, such as sentiment analysis are quite similar in operation to the Next Sentence classification. We only need to add a classification layer on top of the output of the Transformer for the [CLS] token. The input representation for a token in BERT model is created by summing the equivalent token, segment and position embeddings. An apprehension of this structure is highlighted in Figure 4.12.

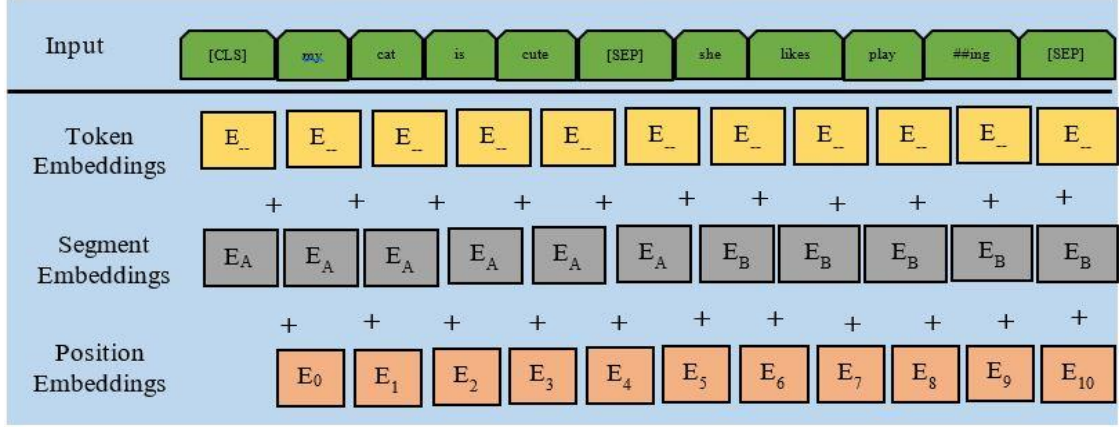


Figure 4.12: BERT Input Representation

Textual features associated with our dataset were obtained using BERT pre-trained model which was accessed from TensorFlow-Hub. BERT-base version was used for feature extraction as it was found to be suitable for our dataset. BERT-base extracts feature using attention mechanism by several encoding layers. Scaled Dot product attention and Multi Head attention detailed in Figure 4.13 (a) and (b) are the core functions of BERT encoder. An attention function can be defined as mapping a query and a set of key-value pairs to an output, where the query, keys, values and output are all vectors. In scaled dot product attention, the input comprises of keys and queries of dimension d_k and values of dimension d_v . To obtain the weights on the values the dot product of the query with all keys is divided by $\sqrt{d_k}$ and passed on to a SoftMax function as in Eq. (4.14). The set of queries, keys and values are arranged together into matrices Q, K and V. Multi-head attention can be described mathematically as in Eq. (4.15) and Eq. (4.16).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4.14)$$

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O \quad (4.15)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (4.16)$$

The projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ and $W^O \in \mathbb{R}^{h \times d_{model}}$ where h is the number of parallel attention layers or heads and d_{model} is the dimension of each head. Twelve encoding layers extract implicit features from the text by taking input as input_ids, input_masks, segment_ids.

BERT pre-trained model has an associated tokenizer which is built on large-sized vocabularies. In the Weibo dataset, character level tokenization is required as it is a Chinese language dataset where each character represents a word. For every sentence [CLS] and [SEP] tokens are appended, where CLS signifies Classification and SEP signifies special separating token.

- Input_word_ids: Encoded tokens using BERT-tokenizer
- Mask_ids: Separates useful and padded tokens (0 or 1)
- Segment_ids: Useful for pairwise training of sentences

These are tokenization inputs to the feed-forward BERT model which has a transformer architecture with 12 attention layers where each layer extracts features by attention mechanism to give output. The model is kept non-trainable to extract the bottleneck features which are obtained by freezing the weights of the entire model. BERT-base gives pooled output and sequence-output and pooled output is used for the further classification task. The vector length associated with every row is 768 sized one-dimension vector.

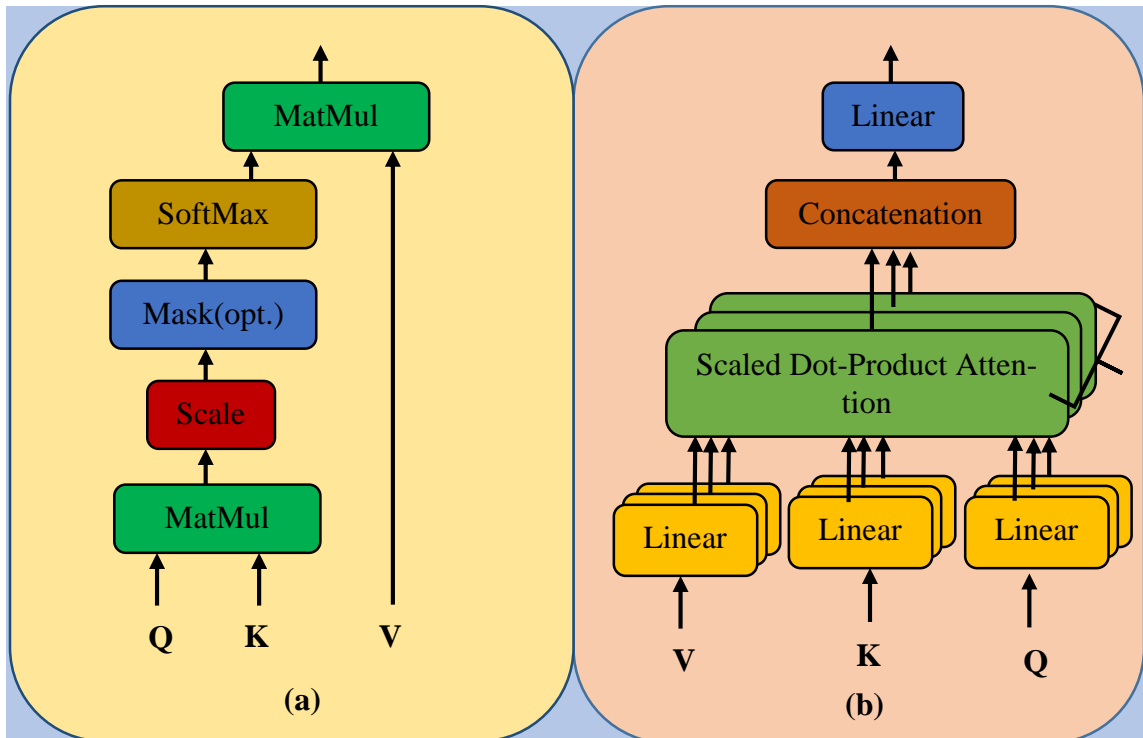


Figure 4.13: (a) Scaled Dot-Product Attention (b) Multi-Head Attention

In our proposed frameworks we harness the capabilities of BERT's multi-layer bidirectional Transformer encoder-based architecture grounded on the creative implementation of Transformer detailed in [200] by Vaswani et. al. The model is first initialized with the pre-trained parameters and then the last layer is replaced with a customized network to fine-tune the parameters with our labelled training datasets for the downstream task of binary fake news classification.

4.3.3 ALBERT

ALBERT, which stands for A Lite BERT [201], is a self-supervised learning model that was presented by Google researchers along with some upgrades to the BERT model. They came up with three major innovations that made it an even more refined model than BERT.

The foremost of them is **factorized embedding parameterization**. This played a major role for efficient allocation of the model's capacity. The size of the hidden layers was isolated from the vocab-embeddings size by using one-hot vectorization into first, an embedding space with lower dimensionality, and then to a hidden space altogether. Because of this, it was possible to increase the size of the hidden layer without much tuning of vocab-embeddings' parameter size. Another key feature was the **cross-parameter sharing**. This facilitated the sharing of parameters across all layers, keeping the depth of the model in control. Hence, ALBERT has 18 times lesser parameters than BERT. Lastly, they introduced the concept of **inter-sentence coherence loss**. The original BERT model was not very reliable when it came to the next sentence prediction tasks. The introduction of SOP loss to model this inter-sentence coherence in ALBERT made it possible to increase the performance in such tasks.

4.3.4 Inception-ResNet-v2

Combining the best features of both the Inception and ResNet models, Google had proposed the Inception-ResNet-v2 [202]. Inception-ResNet-v2 is a type of CNN (convolutional neural network) that is trained on the ImageNet dataset that has more than a million images. The network has a depth of 164 layers and can correctly classify images into about 1000 distinct groups such as pen, tree, box and many fruits. Henceforth, the network has

learned sufficient feature representation techniques for a large set of images. The network has a standard size for input images, that is, 299x299.

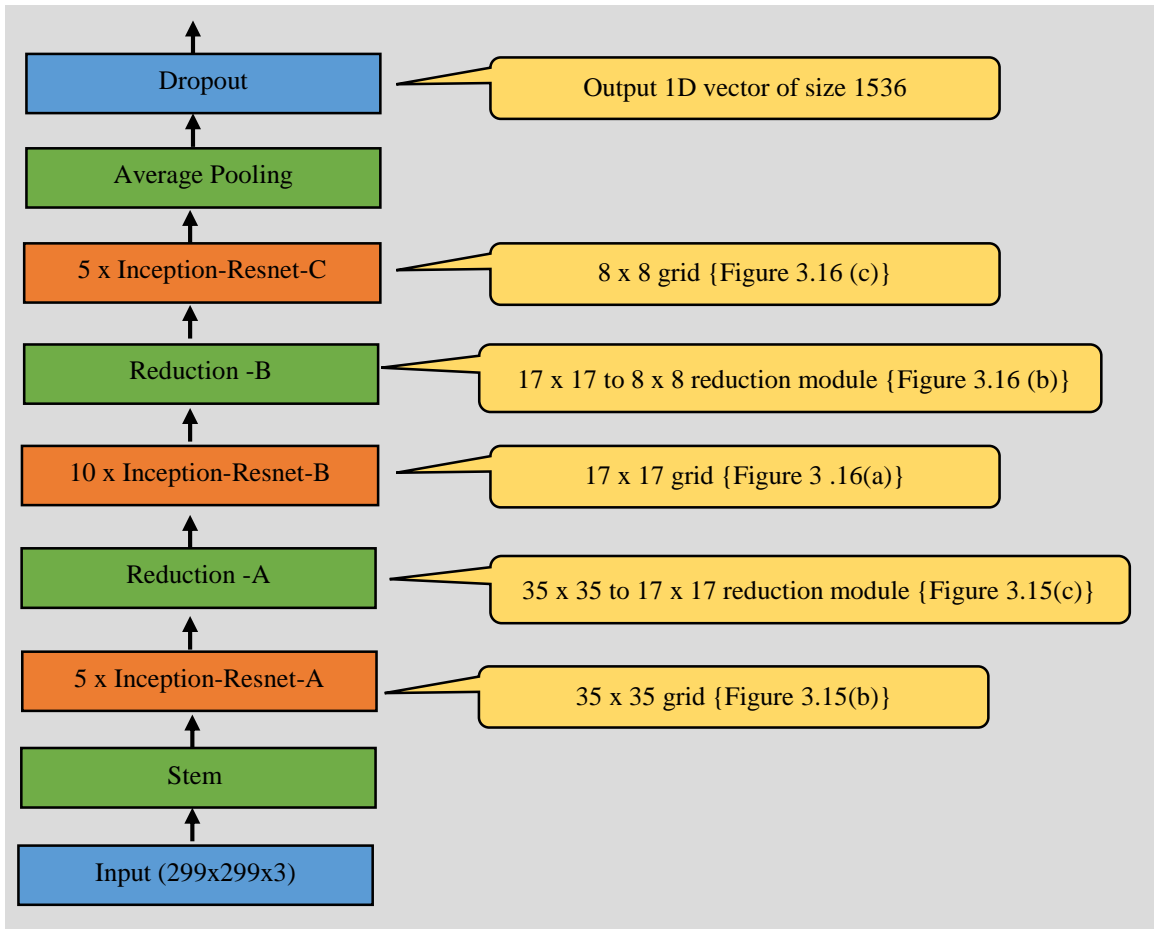


Figure 4.14: Schema for Inception-ResNet-v2 Network

Figure 4.14 represents the large-scale schema of Inception-ResNet-v2 network. Figure 4.15 (a), (b), (c) and Figure 4.16 (a), (b), (c) characterizes the comprehensive structure of its components. All the convolutions not marked with “V” in the figures are same-padded and convolutions marked with “V” are valid padded. The k, l, m and n represent the number of filters of the Reduction-A module for Inception-ResNet-v2 network and their values are 256, 256, 384 and 384 respectively. After pre-processing the images, the InceptionResnetv2 model is loaded from the Keras Applications Library. After the model is loaded the final prediction layer is removed which classifies the image to 1000 available classes. The output of the penultimate layer is taken as our feature vector. The vector obtained from the InceptionResNetv2 model is a one-dimensional vector of size 1536.

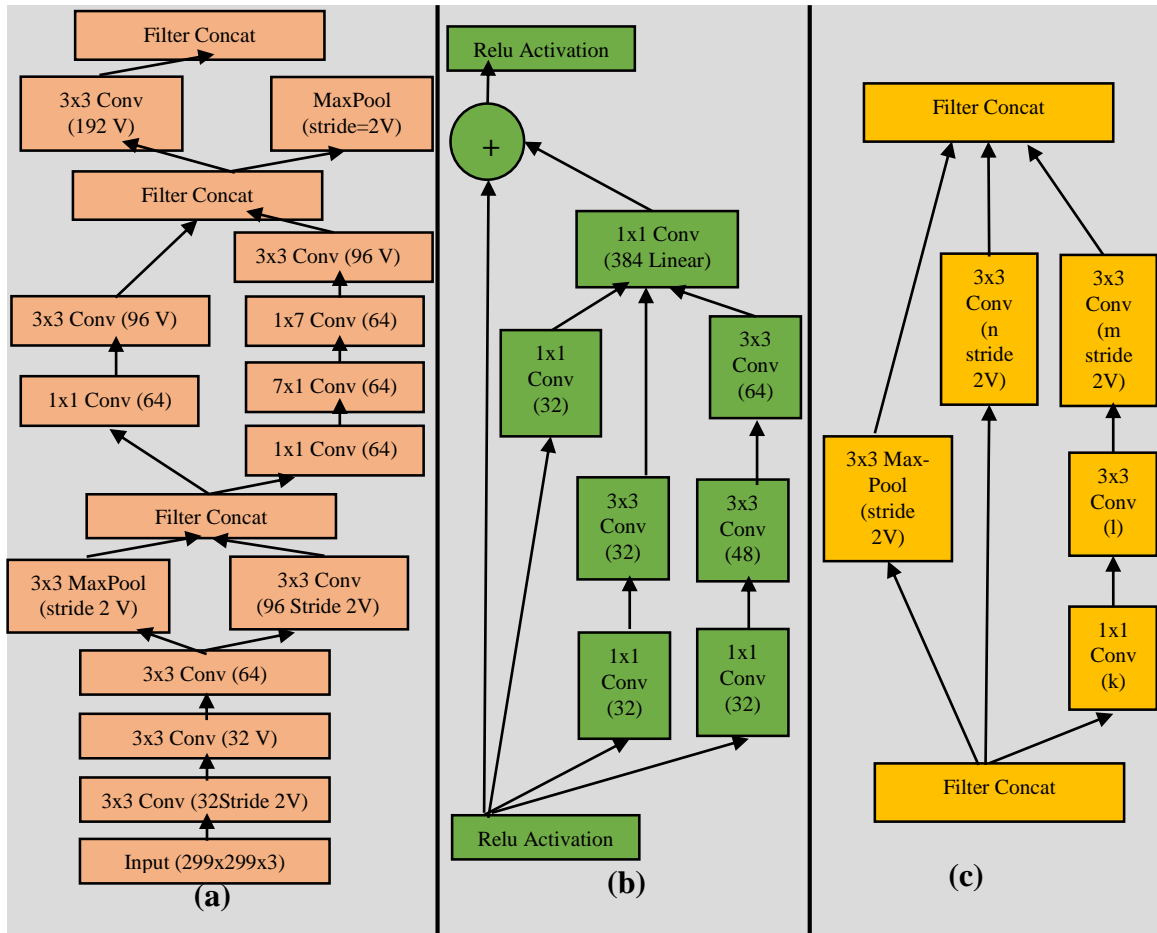


Figure 4.15: Schema for (a) Stem (b) Inception-ResNet-A (c) Reduction-A blocks of Inception-ResNet-v2 Network

Inception-Resnet model introduced the concept of residual connections that augment the output produced by the convolution operation of the basic inception model, to the input image. A necessary condition for this residual connection to work, the input image representation and the output produced after the convolution operation should have the same dimensions. Therefore, 1x1 convolutions are used after the original convolutions are performed, to keep the same depth. It is to be noted that after the convolution operation, the depth gets increased. InceptionResNet-v2 model is famous for attaining higher accuracy even at lower epoch cycles.

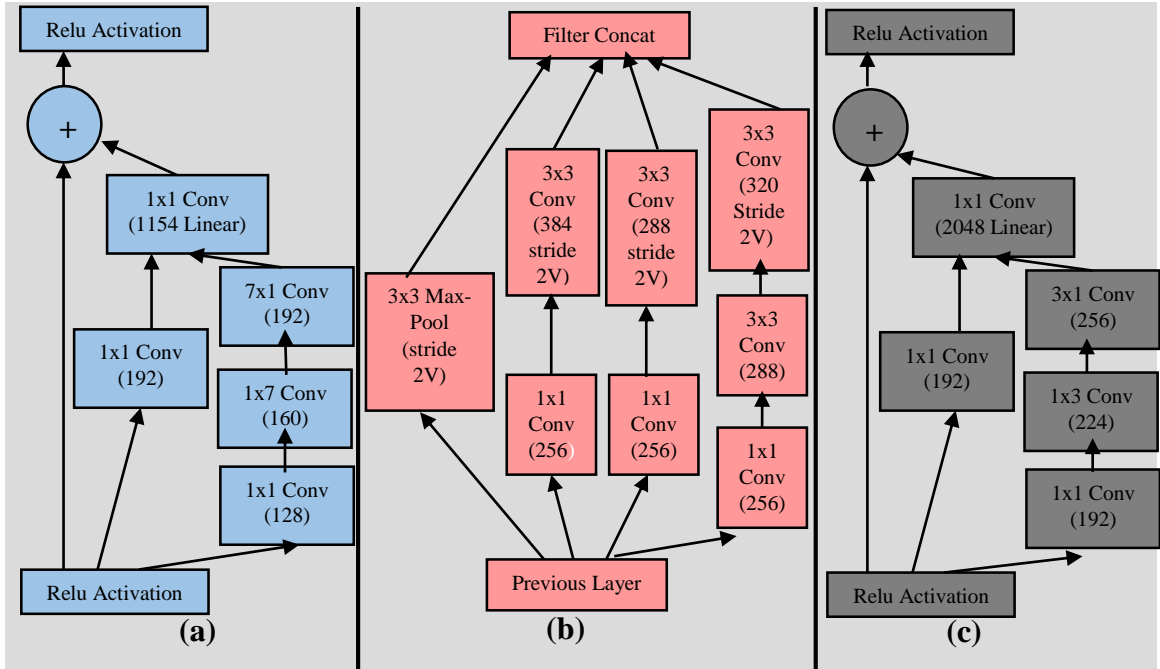


Figure 4.16: Schema for (a) Inception-Resnet-B (b) Reduction-B (c) Inception-Resnet-C blocks of Inception-ResNet-v2 Network

4.3.5 Multi-modal Fusion

Multi-modal fusion is the mechanism of combining information from diverse data channels into a single unit. According to the basic architectural levels, there are three basic approaches of fusion: recognition-based, decision based and hybrid multi-level fusion. The recognition-based fusion also popularly known as early fusion merge the feature vectors of each stream by means of integration (concatenation) mechanism. The main requirement of integrating feature vectors from multiple varied modalities is that their dimensions must be identical and this can be achieved by reshape operator. The decision-based fusion, popularly known as late fusion is the process of merging the final probability decisions coming from each branch. Hybrid multi-level fusion comprises of integrating the input modalities by distributing among the decision and recognition levels.

To perform a multi-modal fusion of independent text and image branches, first each of the modality is separately used to train two different models:

- A BERT/ALBERT model followed by custom layers is used to train the text model.

- A pre-trained Inception-ResNet-v2 followed by custom layers is used to train the image model.

Early fusion or recognition-based fusion is achieved by concatenating the feature vectors after converting them into identical dimensions. Late fusion or decision-based fusion is achieved by weighted averaging of the probabilities obtained by the independent predicted arrays from each parallel branch. The overall prediction is performed according to Eq. (4.17) by using two normalized weights (w_1, w_2); where p_i is the probability of i^{th} feature and w_i is the corresponding assigned weight. The weights assigned for text classifier is w_1 and for image classifier is w_2 .

$$\text{Final prediction} = \frac{\sum_{i=1}^2 w_i p_i}{\sum_{i=1}^2 w_i} \quad (4.17)$$

4.3.6 Implementation Details

In this sub-section, we extensively detail the experimentation settings, datasets, parameter selection and insights gained from the wide range of experiments performed. The implementation is done on Google Colab which offers up to 13.53 free RAM and 12 GB NVIDIA Tesla K80 GPU. The proposed framework is built and implemented in Python 3 on top of the Keras deep learning framework. All the three datasets are split in the ratio of 7:1:2 for the train, validate and test clusters respectively to make our experimentation better and effective. The performance scores and comparison of the proposed framework are listed in terms of F1-measure, accuracy, recall, precision evaluation metrics. Analysis of the results is being done in numerical as well as in graphical representation using the shapes of accuracy-loss trends with epochs and area under the curve plots individually for each dataset.

To substantiate the efficacy of our proposed framework we experimented with three different datasets comprising unique properties of language and writing style. The first dataset "All Data" contains English news articles, the second one "Weibo" contains Chinese news articles and the third one twitter dataset "MediaEval 2016" comprising of tweets on various topics. All three datasets have text as well as images. The following Table 4.12 highlights the details of the datasets used.

Table 4.12: Dataset Details

Dataset Name	Details	Attributes Used	Total entries	Fake news count	Real news count
All Data [195]	Dataset contains English News Articles with 54 attributes including site URL, Image URL, Body, Headline, Label	Title, Text, Image, Label	20015	11941	8074
Weibo [133]	Dataset contains Chinese News articles with four columns title, text, image URL and label	Title, Text, Image, Label	5250	2767	2438
MediaEval 2016 [198]	Dataset contains tweets related to multiple events with post_id, post_text, user_id, image_ids, username, timestamp and label	Tweet text, Image, Label	15519	6875	8644

All Data dataset has news articles in English language. It contains title, title-length, body/text, text_length, site_url, sentiment words count, title-length, image_url, and other metadata of a total of 54 attributes for almost 25000 news articles with binary labels of real and fake. The attributes which were used are the title of the news post, the body of the post, associated image and label. Cleaning and pre-processing were performed on the dataset before use. Entire rows that had null values for any of these four attributes were removed. Also, entire rows were removed if the image from image_url could not be fetched from the given source. The Title and Text attributes were concatenated into a single attribute for each row. After cleaning, the dataset had 20,015 rows of data with 11 914 fake news and 8 047 real news count.

Weibo dataset is the standard Chinese dataset having microblogs from the Chinese microblogging site Sina Weibo. It is one of the most popularly used dataset and has news articles in the Chinese Language (Mandarin). News articles are accompanied by an image URL, from which images were fetched and stored before use. The dataset contains title, text, image and binary label (Real/Fake) associated with each news article. Entire rows that had null values for any of these four attributes were removed. Also, entire rows were removed if the image from image_url could not be fetched from the given source. After initial preprocessing, the dataset contains 5250 total news articles out of which 2767 are fake and 2438 are real news articles. The Title and Text attributes were concatenated into a single attribute for each row. For preprocessing, tokenization was performed using a Chinese character level tokenizer. BERT model has a tokenizer which is the most suitable for this

task. Tokenization was done on a character level as the Chinese language is to character in the same way as the English language is to words. After this, the dataset was ready to be fed as input to our model.

This dataset is part of the MediaEval task of 2016 and is available online at the Image Verification corpus [198]. The dataset contains tweets related to multiple events. It has been extensively used since Twitter is amongst the most commonly used social media platforms and is highly susceptible to the creation as well as propagation of fake news. It has several attributes: `post_id`, `post_text`, `user_id`, `image_ids`, `username`, `timestamp` and `label`. The attributes which were used are `Post_text` - The body of the post, `Image_ids` - link to an image related to the post and `label` - Classified as real or fake. Cleaning and preprocessing were performed on the dataset before use. Entire rows that had null values for any of these three attributes were removed. Also, entire rows were removed if the image from `image_ids` could not be fetched from the given source. After cleaning the dataset contains a total of 15 519 tweets with images and binary labels out of which 6875 tweets are fake and 8644 tweets are real.

4.3.7 Model Parameter Description

The network is being trained and tested for three datasets. BERT-base and ALBERT-base pretrained networks are used for text feature extraction and Inception-ResNet-v2 pretrained model is used for image feature vector extraction. All the model parameters are kept identical for these three datasets except the maximum sequence length which is set to 512, 400 and 20 respectively for All Data, Weibo and MediaEval dataset. The network is trained with Adam optimizer having learning rate 0.0001, momentum parameter $\beta_1 = 0.9$ and $\beta_2=0.980$ up to 30 epochs with a batch size of 128. To train Early Fusion architecture text and image feature vectors of length 768 and 1536 are concatenated together followed by five dense layers for smoother dimensionality reduction. The neurons in dense layers are 1024, 512, 128, 64 and 2 correspondingly. Two neuron final dense layer with sigmoid activation supports binary classification. Throughout the custom layers dropout probability is 0.4 with batch normalization. The initial four dense layers are equipped with ReLU activation function and loss is calculated with binary cross entropy loss. The Late fusion

multi-modal architecture calculates independent prediction probability p_1 for text and p_2 for image stream. These probabilities are fused with weighted averaging late fusion framework analysed extensively for four different weight combinations $(w_1, w_2) = \{(0.4,0.6), (0.5, 0.5), (0.6,0.4), (0.7, 0.3)\}$. Feature vector extracted from text is followed by four fully connected dense layers of 512, 128, 64 and 2 neurons. Feature vector extracted from image part is followed by five fully connected dense layers of 1024, 512, 128, 64 and 2 neurons. All other hyperparameters are same for both the architectures to have an efficient analysis and fair comparison of the prediction preciseness.

4.3.8 Result Analysis

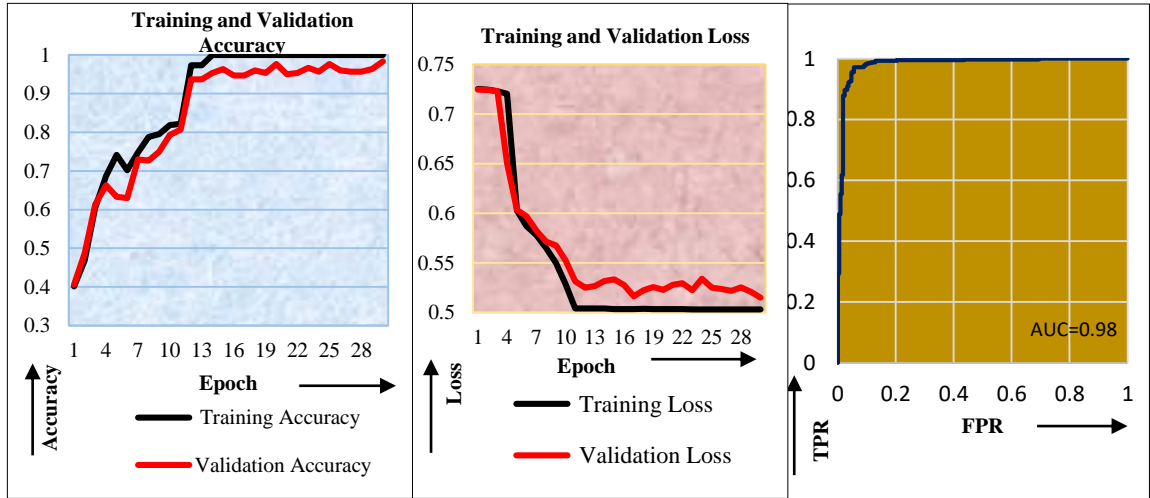
We employed performance measures like Accuracy, Precision, Recall, F1Score, Accuracy-epoch curve, loss-epoch curve and ROC Curve to analyse and assess the outcomes attained by performing the experimentations on three datasets. The results are presented in Table 4.13, Table 4.14 and Table 4.15 for each dataset sequentially. On each dataset, we applied multi-modal early fusion and late fusion with four different weight combinations for weighted averaging the results using both variants of the framework i.e. BERT and Inception-ResNet-v2 as well as ALBERT and Inception-ResNet-v2. For a clear understanding and insights gained through these results, pictorial representation in terms of graphs is also presented.

The highest accuracy achieved on All Data dataset is 97.19% with ALBERT and Inception-ResNet-v2 framework. Table 4.13 details in with the outcomes of applying early fusion and late fusion on the dataset. Figure 4.17 (a), (b) and (c) concentrates on the accuracy versus epoch curve, loss versus epoch curve and ROC curve.

Table 4.13: Result Analysis on All Data Dataset

Fusion	Weightage of text & image streams	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BERT+Inception Resnet-v2					
Early Fusion	Feature vector concatenation	94.10	95.02	94.88	94.95
Late Fusion	$(w_1, w_2) = (0.4,0.6)$	85.75	78.33	82.17	80.20
	$(w_1, w_2) = (0.5,0.5)$	90.33	88.06	92.64	90.29
	$(w_1, w_2) = (0.6,0.4)$	95.76	92.20	97.45	94.75

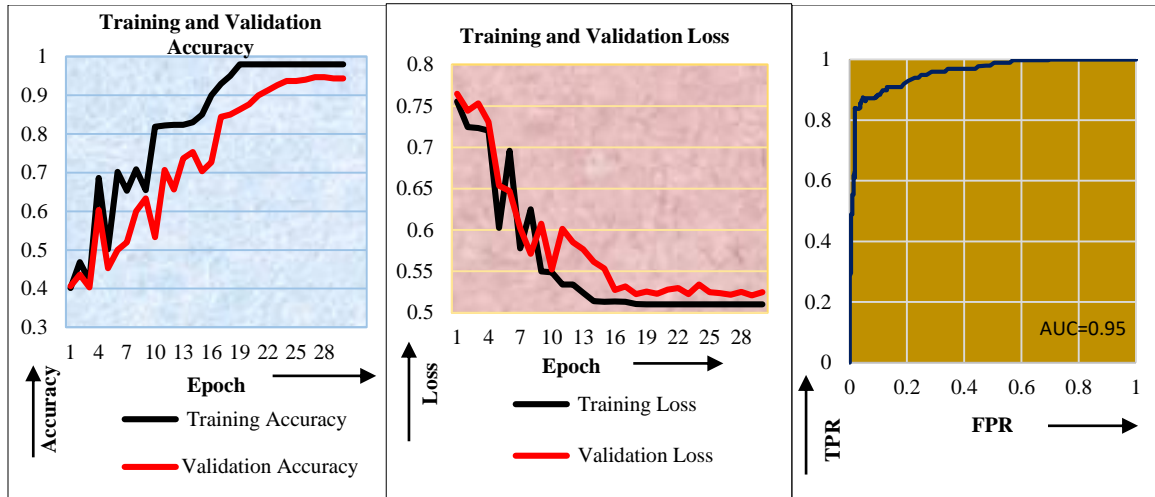
	$(w_1, w_2) = (0.7, 0.3)$	91.33	91.20	89.74	90.46
ALBERT+Inception Resnet-v2					
Early Fusion	Feature vector concatenation	96.41	95.48	94.38	94.93
Late Fusion	$(w_1, w_2) = (0.4, 0.6)$	87.06	85.77	89.23	87.47
	$(w_1, w_2) = (0.5, 0.5)$	92.70	90.26	92.11	91.17
	$(w_1, w_2) = (0.6, 0.4)$	97.19	97.00	99.00	97.99
	$(w_1, w_2) = (0.7, 0.3)$	91.29	92.66	88.24	90.39



(a) (b) (c)
Figure 4.17: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for All Data dataset

Table 4.14: Result analysis on Weibo Dataset

Fusion	Weightage of text & image streams	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BERT+Inception Resnet-v2					
Early Fusion	Feature vector concatenation	89.61	78.43	76.34	77.38
Late Fusion	$(w_1, w_2) = (0.4, 0.6)$	80.55	79.42	83.17	81.25
	$(w_1, w_2) = (0.5, 0.5)$	83.40	86.27	78.90	82.42
	$(w_1, w_2) = (0.6, 0.4)$	89.92	90.20	93.42	91.78
	$(w_1, w_2) = (0.7, 0.3)$	82.45	87.32	74.65	80.49
ALBERT+Inception Resnet-v2					
Early Fusion	Feature vector concatenation	91.00	91.27	89.79	90.52
Late Fusion	$(w_1, w_2) = (0.4, 0.6)$	83.24	86.47	94.33	90.23
	$(w_1, w_2) = (0.5, 0.5)$	83.67	87.24	92.22	89.66
	$(w_1, w_2) = (0.6, 0.4)$	94.28	94.07	95.82	94.94
	$(w_1, w_2) = (0.7, 0.3)$	88.04	87.00	90.22	88.58



(a) (b) (c)
Figure 4.18: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for Weibo dataset

Table 4.15: Result Analysis on MediaEval Dataset

Fusion	Weightage of text & image streams	Accuracy	Precision	Recall	F1-score
BERT+Inception Resnet-v2					
Early Fusion	Feature vector concatenation	66.69	59.52	40.86	48.46
Late Fusion	(w1, w2) = (0.4,0.6)	54.22	59.80	54.17	56.84
	(w1, w2) = (0.5,0.5)	59.67	68.52	52.86	59.68
	(w1, w2) = (0.6,0.4)	67.04	74.91	62.50	68.14
	(w1, w2) = (0.7,0.3)	61.51	61.27	65.77	63.44
ALBERT+Inception Resnet-v2					
Early Fusion	Feature vector concatenation	67.79	60.30	76.00	67.24
Late Fusion	(w1, w2) = (0.4,0.6)	57.61	67.25	52.28	58.82
	(w1, w2) = (0.5,0.5)	69.07	70.86	52.47	60.29
	(w1, w2) = (0.6,0.4)	75.33	78.00	62.50	69.39
	(w1, w2) = (0.7,0.3)	63.01	74.52	52.09	61.32

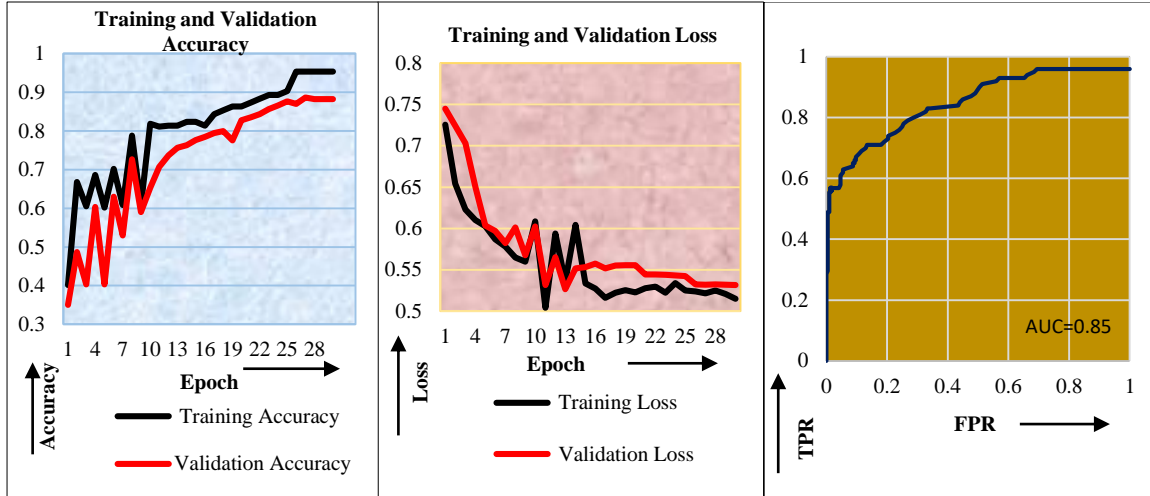


Figure 4.19: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for MediaEval dataset

It is evident from Table 4.14 and Table 4.15 that highest accuracy achieved on Weibo and MediaEval dataset are 94.28% and 75.33% respectively. Figure 4.18 and Figure 4.19 elaborates the effectiveness of the training and validation process as well as the robustness of the framework in terms of ROC curve. The above discussion reinforced with Tables 4.13, 4.14, and 4.15 along with Figures 4.17, 4.18 and 4.19 verifies that our proposed multi-modal fusion architecture using fine-tuned self-attention and transfer learning accomplishes promising performance and outstanding precisions for veracity analysis of web information.

4.3.9 State-of-the-art Comparison

To compare the efficacy of our designed architecture with contemporary methods, we calculated the performance of five different baseline methods in terms of accuracy, precision, recall and F1 score with the dataset split as 7:1:2 for training, validation and testing. State-of-the-art juxtaposition on each of the All Data, Weibo and MediaEval datasets are outlined in Table 4.16, Table 4.17 and Table 4.18, correspondingly. The approaches used as baselines for state-of-the-art comparison are as follows:

- Vishwakarma et al. [124] implemented reverse search method of extracted text from images on internet to calculate a reality parameter by using top search results.

- Wang et al [120] devised multi-modal fake news detector by incorporating text CNN and VGG19 along with event discriminator module. To have a fair comparison, we removed the event discriminator module and compared on binary classifier with our datasets.
- Early fusion of textual features extracted by LSTM and visual features extracted by VGG19 is being proposed and implemented by Jin et al. [203].
- Lago et al. [197] proposed fake news detection multi-modal framework by a combination of classical and advanced techniques such as Mean-filter noise residue inconsistency, Splice buster, JPEG ghosts, TF-IDF, cosine similarity, Logistic Regression, CNN etc.
- Khattar et al. [121] designed variational autoencoder system with three core parts encoders, decoder and a binary fake news detection using Bi-LSTM and VGG19 methods for veracity analysis of online articles.

Table 4.16: State-of-the-art Comparison on All Data Dataset

Reference	Input Modality	Proposed Method /Classifier	Acc. (%)	P (%)	R (%)	F (%)
Vishwakarma et al. [124]	Text +Image	Rule based classifier, Reality Parameter	88.00	87.90	88.80	88.34
Wang et al. [120]	Text +Image	CNN + VGG19	84.27	86.72	83.28	84.96
Jin et al. [203]	Text +Image	LSTM + VGG 19	79.66	87.19	79.61	83.22
Lago et al. [197]	Text +Image	Classical and advanced image forensics, CNN, LR, RF, Jaccard's similarity, TF-IDF, cosine similarity	92.34	93.06	94.06	93.56
Khattar et al. [121]	Text +Image	BiLSTM + VGG19	85.70	87.81	83.03	85.34
Proposed Method	Text + Image	ALBERT + Inception-ResNet-v2	97.19	97.00	99.00	97.99

Table 4.17: State-of-the-art comparison on weibo Dataset

Reference	Input Modality	Proposed Method /Classifier	Acc. (%)	P (%)	R (%)	F (%)
Vishwakarma et al. [124]	Text +Image	Rule based classifier, Reality Parameter	90.88	91.24	89.16	90.19
Wang et al. [120]	Text +Image	CNN + VGG19	82.73	84.70	81.22	82.92
Jin et al. [203]	Text +Image	LSTM + VGG 19	78.80	86.20	68.60	76.40

Lago et al. [197]	Text +Image	Classical and advanced image forensics, CNN, LR, RF, Jaccard's similarity, TF-IDF, cosine similarity	80.02	79.66	83.24	81.41
Khattar et al. [121]	Text +Image	BiLSTM + VGG19	82.41	85.42	76.93	80.95
Proposed Method	Text + Image	ALBERT + Inception-ResNet-v2	94.28	94.07	95.82	94.94

Table 4.18: State-of-the-art comparison on MediaEval Dataset

Reference	Input Modality	Proposed Method /Classifier	Acc. (%)	P (%)	R (%)	F (%)
Vishwakarma et al. [124]	Text +Image	Rule based classifier, Reality Parameter	73.45	75.20	71.00	73.04
Wang et al. [120]	Text +Image	CNN + VGG19	71.50	82.20	63.80	71.84
Jin et al. [203]	Text +Image	LSTM + VGG 19	68.20	78.00	61.50	68.77
Lago et al. [197]	Text +Image	Classical and advanced image forensics, CNN, LR, RF, Jaccard's similarity, TF-IDF, cosine similarity	75.00	76.32	71.67	73.92
Khattar et al. [121]	Text +Image	BiLSTM + VGG19	74.52	80.10	71.90	75.77
Proposed Method	Text + Image	ALBERT+ Inception-ResNet-v2	75.33	78.00	62.50	69.39

The above discussion concludes that our proposed model for veracity analysis of web information is reasonably promising. The accuracy over all the three datasets All Data, Weibo and MediaEval is a decent development over the parallel methods. Fine-tuned self-attention and transfer learning are the prominent technologies that have facilitated refining the preciseness of our results. Finally, merging multiple data streams with late fusion helps us achieve 97.19% highest fake news detection accuracy.

4.3.10 Significant Outcomes

The significant outcomes of the work are as follows:

- The problem of fraudulent content in text and image multi-modal data format is being identified and addressed for veracity analysis of web information contents.
- Textual feature extraction and model training is being done using BERT and ALBERT bidirectional self-attention-based transformers.
- Visual data is being analysed by harnessing transfer learning capabilities using fine-

tuned Inception-ResNet-v2 deep neural network architecture.

- The proposed framework focused on two independent multi-modal fusion frameworks of BERT and Inception-ResNet-V2 as well as ALBERT and Inception-ResNet-V2.
- The multi-modal fusion of textual and visual branches is extensively experimented and analysed using concatenation of feature vectors and weighted averaging of probabilities named Early Fusion and Late Fusion respectively.
- Three broadly accepted datasets All Data, Weibo and MediaEval 2016 that incorporate English news articles, Chinese news articles, and Tweets correspondingly are used to test and compare our designed framework's outcomes with previous notable work in the domain.

Chapter 5

Conclusion and Future Scope

This chapter provides a summary of proposed works, significant findings and contributions. Further, we also suggest some future directions, short-term and long-term perspectives for combating the issue of information pollution on social media and other web platforms.

5.1 Conclusions

We developed four approaches that deal with the features from multimodal (Text and Image) and unimodal data using supervised as well as semi-supervised advanced deep-learning, image forensics and other techniques for veracity analysis of online news articles. The approaches are summarized as follows:

- In this work, considering the painstaking and inconsistent task of huge volumes of online data annotation, we tried to propose a semi-supervised text fake news classification framework based on temporal ensembling convolutional neural network architecture. The system is trained using title and body part of the news articles with different size convolutional filters and then extracted feature vectors are concatenated together. The validation experiments are repeated for five different proportions 10%, 20%, 30%, 40% and 50% respectively, of labelled and rest unlabelled training samples for each of the three datasets. The experimental result analysis and comparison against peer technologies in terms of precision, accuracy, recall, F1score and ROC curve advocates the promising performance of the proposed work.
- In the second approach, we tried to design a semi-supervised text fake news detection framework based on Graph Convolutional Networks. Embedding the text arti-

cles in Euclidean space, similarity graph constructing using Word Mover's Distance and Graph Classification are the three landmark components of the designed architecture. Extensive experimental analysis has been done by repeating the training and testing of the model for different proportions of labelled and unlabelled data for each dataset. The promising results compared in terms of accuracy, precision, recall, F1 score and ROC Curves very well advocates the worthiness of the method for veracity analysis. The proposed framework exhibits encouraging results by out-doing numerous state-of-the-art methods on multiple standard datasets.

- In this third framework, the fake news problem is explored and a hybrid multimodal framework is suggested to battle against it which uses an ensemble of both textual and visual features. The datasets are converted to a suitable form by pre-processing them and scrapping images from the web with the corresponding news. HAN deep learning model is used to extract hidden patterns at the word level and sentence level text in the headline and body part of the news. One of the novel features is to generate an image summary using the automatic caption generator tool. The generated image caption and news headline are then matched for semantic similarity with the body of the news. The test for forgery on the image is applied by checking whether extra local noise is being added during tampering the image or not. Another feature applied is to check Error Level Analysis by duplicating images with less quality and then computing the difference. Finally, textual and visual features are ensembled using the max voting method and a robust framework for fake news detection is designed.
- Finally, we have considered both the textual and visual characteristics of the input news instances. To use these features in veracity analysis, the latest pre-trained deep learning models BERT, ALBERT and Inception-ResNet-v2 have been used. These models are fine-tuned according to our application and dataset requirements to get precise results. Fine-tuning for leveraging existing knowledge in a deep learning model can be potentially beneficial for automatic detection and analysis of fake news. Extensive evaluation of the proposed Multimodal Early and Late fusion

framework on three renowned datasets of English news articles, Chinese news articles and Tweets demonstrated our proposed model's commendable performance in identifying the fake news articles.

5.2 Future Scope

A lot of work has been done in the past years to make online content more reliable and trustworthy. Some of the key areas still remain unaddressed. Quick and real-time detection of the source is useful to control the spread of false information and reduce the adverse impact on society. Real-time collected datasets, automatic detection of rumors and finding its original source is challenging issue. The following section highlights the potential future direction of work.

- **Cross-platform detection:** As people have accounts on various social networking websites and sometimes, they spread the rumor across their different social networks, in such cases source detection becomes somewhat difficult. Along with this, propagation of false information from one web community to another i.e. cross-platform spread and detection has become a significant challenge for tracking in front of the researchers.
- **Real-time learning:** Deployment of a web-based application for fact-checking which can learn in real-time from new manually fact-checked articles and provides real-time detection of fraudulent information.
- **Unsupervised models:** Current work is mainly done by using supervised learning approaches. Unsupervised models need to be developed due to massive unlabelled data from social media.
- **Datasets:** The establishment of convincing gold standard datasets in this field is highly required as most of the research is being done on customized datasets. Because of the lack of publicly available large-scale datasets a benchmark comparison between different algorithms cannot be done.
- **Multilingual platform:** Most of the work focuses on linguistic features in English language text. Other popular and regional languages (multilingual platform for fake news detection) are not considered yet.

- **Complex and dynamic network structure:** The veracity classification task becomes a prediction task if we are doing it before its resolution and requires a huge amount of supporting evidence. The issue further complicates because of the complex and dynamic network structure of social platforms.
- **Early detection:** Detecting fake news at the early stage is a highly challenging task before it becomes widespread so that timely actions can be taken for its mitigation and intervention. After fake news has become widespread and gained users' trust, it's almost impossible to change people's perception.
- **Cross-domain analysis:** Most of the existing approach focuses only on one way of deception detection, mainly in the form of content, propagation, style, etc. Cross-domain analysis, including multiple aspects such as topic-website-language-images-URL, helps in identifying unique non-varying characteristics, provides early accurate detection of fraudulent content.
- **Deep learning:** Deep learning technologies can address all formats of information text, image, speech and video. Deep architecture is customizable to a new class of problem and it bypasses feature engineering, which is the most time-consuming but necessary part of a machine-learning framework. However, the disadvantage of deep learning technologies is that they require a considerable amount of time for model training with a relatively massive amount of data and do not provide interpretations of what the model has actually learned, so inside the model it is almost a black box type of processing.
- **Multimedia false information detection:** Fabricated and manipulated audio, images and videos need developing data analytics, computer vision and signal processing techniques. To discover signature characteristics of manipulated and fabricated multimedia machine learning and deep learning algorithms are highly required.
- **Bridging echo chambers:** Social media is prone to form echo chambers when a user's existing beliefs, views are reinforced and he is not aware of the opposite beliefs. Therefore, further research is required to bridge the conflicting echo cham-

bers in order to effectively exchange the opposing beliefs to readers so that polarization can be reduced. It also helps in truth discovery by making users think judiciously and rationally on multiple dimensions.

References

- [1] M. Nunes and J. Correia, “Improving trust using online credibility sources and social network quality in P2P marketplaces,” in *IEEE 8th Iberian Conference on Information Systems and Technologies (CISTI)*, Lisboa, Portugal, 2013.
- [2] C. Shao, G. L. Ciampaglia, A. Flammini and F. Menczer, “Hoaxy: A Platform for Tracking Online Misinformation,” in *25th international conference companion on world wide web*, Montréal Québec Canada, 2016.
- [3] S. Zannettou, M. Sirivianos, J. Blackburn and N. Kourtellis, “The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans,” *Journal of Data and Information Quality (JDIQ)*, vol. 11, no. 3, pp. 1-37, 2019.
- [4] S. Kumar and N. Shah, “False information on web and social media: A survey.,” *arXiv preprint arXiv:1804.08559*, 2018.
- [5] X. Zhou and R. Zafarani, “Fake News: A Survey of Research, Detection Methods, and Opportunities,” *arXiv preprint arXiv:1812.00315 2*, 2018.
- [6] S. Shelke and V. Attar, “Source detection of rumor in social network—a review,” *Online Social Networks and Media*, vol. 9, pp. 30-42, 2019.
- [7] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata and R. Procter, “Detection and resolution of rumours in social media: A survey,” *ACM Computing Surveys*, vol. 51, no. 2, pp. 1-36, 2018.
- [8] N. Newman, R. Fletcher, A. Kalogeropoulos, D. A. L. Levy and R. K. Nielsen, “Reuter Institute Digital News Report 2018,” Reuters Institute for the Study of Journalism, Oxford, UK, 2018.
- [9] Y. Chi, S. Zhu, K. Hino, Y. Gong and Y. Zhang, “iOLAP: A framework for analyzing the internet, social networks, and other networked data,” *IEEE transactions on multimedia*, vol. 11, no. 3, pp. 372-382, 2009.
- [10] S. Vosoughi, D. Roy and S. Aral, “The spread of true and false news online,” *Science*, vol. 359, no. 6380, pp. 1146-1151, 2018.
- [11] B. Horne and S. 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 *International AAAI Conference on Web and Social Media*, Montreal, Canada, 2017.
- [12] D. Martens and W. Maalej, “Towards understanding and detecting fake reviews in app stores,” *Empirical Software Engineering*, vol. 24, no. 6, pp. 3316-3355, 2019.
- [13] E. Elmurngi and A. Gherbi, “Detecting fake reviews through sentiment analysis using machine learning techniques,” in *Sixth International Conference on Data Analytics*, Barcelona, Spain, 2017.

- [14] E. Elmurngi and A. Gherbi, "An Empirical Study on Detecting Fake Reviews Using Machine Learning Techniques," in *IEEE seventh international conference on innovative computing technology (INTECH)*, Luton , UK, 2017.
- [15] M. Dong, L. Yao, X. Wang, B. Benatallah, C. Huang and X. Ning, "Opinion fraud detection via neural autoencoder decision forest," *Pattern Recognition Letters*, vol. 123, pp. 21-29, 2020.
- [16] H. Ahmed, I. Traore and S. Saad, "Detecting opinion spams and fake news using text classification," *Security and Privacy (WILEY)*, vol. 1, no. 1, pp. 1-15, 2018.
- [17] M. Viviani and G. Pasi, "Credibility in social media: opinions, news, and health information—a survey," *Wiley interdisciplinary reviews: Data mining and knowledge discovery* , vol. 7, no. 5, p. e1209, 2017.
- [18] K. Starbird, "Examining the Alternative Media Ecosystem through the Production of Alternative Narratives of Mass Shooting Events on Twitter," in *International AAAI Conference on Web and Social Media*, Montreal, Canada, 2017.
- [19] "The Onion - America's Finest News Source," [Online]. Available: <https://www.theonion.com/>.
- [20] "SatireWire," [Online]. Available: <http://www.satirewire.com/> .
- [21] I. Reilly, "F for Fake: Propaganda! Hoaxing! Hacking! Partisanship! and Activism! in the Fake News Ecology," *The Journal of American Culture* , vol. 41, no. 2, pp. 139-152, 2018.
- [22] E. R. X. H. Sina Mohseni, "Open Issues in Combating Fake News: Interpretability as an Opportunity," *arXiv preprint arXiv:1904.03016* , 2019.
- [23] "Global social media ranking 2019 | Statista," [Online]. Available: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>.
- [24] "Chinese panic-buy salt," [Online]. Available: <https://www.theguardian.com/world/2011/mar/17/chinese-panic-buy-salt-japan>.
- [25] "Bogus' AP tweet about explosion at the White House wipes billions off US markets - Telegraph," [Online]. Available: <https://www.telegraph.co.uk/finance/markets/10013768/Bogus-AP-tweet-about-explosion-at-the-White-House-wipes-billions-off-US-markets.html>.
- [26] "Mexico 'Twitter terrorism' charges cause uproar - BBC News," [Online]. Available: <https://www.bbc.com/news/world-latin-america-14800200>.
- [27] "FACT CHECK: Did 600 Murders Take Place in Chicago Last Weekend?," [Online]. Available: <https://www.snopes.com/fact-check/600-murders-in-chicago/>.
- [28] "Donald Trump ends school shootings by banning schools | 8Satire," [Online]. Available: <https://www.8satire.com/donald-trump-ends-school-shootings-banning-schools/>.
- [29] "FakeNews: Was Kamal Nath the driver of Rajiv Gandhi?," [Online]. Available: <https://in.news.yahoo.com/fakenews-kamal-nath-driver-rajiv-gandhi-094054585.html>.

- [30] “Fake News: North Korea Opening its doors to Christians,” [Online]. Available: <https://medium.com/@kpeterson16/fake-news-north-korea-opening-its-doors-to-christians-c07ca719ff10>.
- [31] “Fact Check: Do paracetamol tablets contain ‘machupo’ virus? - Times of India,” [Online]. Available: <https://timesofindia.indiatimes.com/news/fact-check-do-paracetamol-tablets-contain-machupo-virus/articleshow/68425709.cms>.
- [32] “Reality Check: Was Hillary Clinton photographed with Osama Bin Laden? - BBC News,” [Online]. Available: <https://www.bbc.com/news/world-41821923> .
- [33] M. V. Bronstein, G. Pennycook, A. Bear, D. G.Rand and T. D.Cannon, “Belief in Fake News is Associated with Delusionality, Dogmatism, Religious Fundamentalism, and Reduced Analytic Thinking,” *ournal of applied research in memory and cognition*, vol. 8, no. 1, pp. 108-117, 2019.
- [34] A. P. Ghaisani, Q. Munajat and P. W. Handayani, “Information Credibility Factors on Information Sharing Activites in Social Media,” in *IEEE Second International Conference on Informatics and Computing (ICIC)*, Jayapura, Indonesia, 2017.
- [35] “Crowd booster,” [Online]. Available: <https://www.crunchbase.com/organization/crowdboosters#section-overview>.
- [36] N. Diakopoulos, M. Naaman and F. Kivran-Swaine, “Diamonds in the rough: Social media visual analytics for journalistic inquiry,” in *IEEE Symposium on Visual Analytics Science and Technology*, Salt Lake City, UT, USA, 2010.
- [37] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu and H. Qu, “Whisper: Tracing the Spatiotemporal Process of Information Diffusion in Real Time,” *IEEE Transactions on Visualization and Computer Graphics* , vol. 18, no. 12, pp. 2649 - 2658, 2012.
- [38] “Social Media Analytics & Monitoring Platform - Talkwalker,” [Online]. Available: <https://www.talkwalker.com> .
- [39] “Google Analytics | Google Developers,” [Online]. Available: <https://developers.google.com/analytics/> .
- [40] “Hootsuite - The Best Way To Manage Social Media,” [Online]. Available: <https://signupnow.hootsuite.com>.
- [41] “Home for Vertical Stories | Snaplytics,” [Online]. Available: <https://www.snaplytics.io/>.
- [42] Y. Li, H. Bao, Y. Zheng and Z. Huang, “Social analytics framework to boost recommendation in online learning communities,” in *IEEE 15th International Conference on Advanced Learning Technologies*, Hualien, Taiwan, 2015.
- [43] M. AlRubaian, M. Al-Qurishi, M. Al-Rakhami, M. M. Hassan and A. Alamri, “CredFinder: A real-time tweets credibility assessing system,” in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, San Francisco, CA, USA, 2016.

- [44] “Twitter Trails_ Tool for monitoring the propagation of rumors,” [Online]. Available: <http://twittertrails.com/>.
- [45] S. Finn, P. T. Metaxas and E. Mustafaraj, “Investigating Rumor Propagation with TwitterTrails,” *arXiv preprint arXiv:1411.3550*, 2014.
- [46] A. Gupta, P. Kumaraguru, C. Castillo and P. Meier, “TweetCred: Real-Time Credibility Assessment of Content on Twitter,” in *International Conference on Social Informatics*, 2014.
- [47] C. Shao, P.-M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer and G. L. Ciampaglia, “Anatomy of an online misinformation network,” *PloS one*, vol. 13, no. 4, 2018.
- [48] “Emergent,” [Online]. Available: <http://www.emergent.info/> .
- [49] “RumorLens | University of Michigan School of Information,” [Online]. Available: <https://www.si.umich.edu/research/research-projects/rumorlens> .
- [50] P. Resnick, S. Carton, S. Park, Y. Shen and N. Zeffer, “RumorLens : A System for Analyzing the Impact of Rumors and Corrections in Social Media,” in *Computational Journalism Conference*, New York, USA, 2014.
- [51] M. Egele, G. Stringhini, C. Kruegel and G. Vigna, “Towards Detecting Compromised Accounts on Social Networks,” *IEEE Transactions on Dependable and Secure Computing* , vol. 14, no. 4, pp. 447 - 460, 2017.
- [52] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin and C. Collins, “#FluxFlow: Visual analysis of anomalous information spreading on social media,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1773 - 1782, 2014.
- [53] “REVEAL | Social Media Verification,” [Online]. Available: <https://revealproject.eu/> .
- [54] “InVID Project - Video Verification,” [Online]. Available: <https://www.invid-project.eu/> .
- [55] N. Hassan, F. Arslan, C. Li and M. Tremayne, “Toward Automated Fact-Checking: Detecting check-worthy factual claims by ClaimBuster,” in *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax NS, Canada, 2017.
- [56] “Truth or Fiction - Fact Check,” [Online]. Available: <https://www.truthorfiction.com/> .
- [57] “Snopes.com | The definitive fact-checking site and reference source for urban legends, folklore, myths, rumors, and misinformation,” [Online]. Available: <https://www.snopes.com/> .
- [58] “FactCheck.org - A Project of The Annenberg Public Policy Center,” [Online]. Available: <https://www.factcheck.org/> .
- [59] “Fact-checking U.S. politics | PolitiFact,” [Online]. Available: <https://www.politifact.com/> .
- [60] K. Shu, D. Mahudeswaran and H. Liu, “FakeNewsTracker: a tool for fake news collection, detection, and visualization,” *Computational and Mathematical Organization Theory*, vol. 25, no. 1, pp. 60-71, 2019.

- [61] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing and Management*, vol. 57, no. 2, pp. 1020-1025, 2020.
- [62] S. Zannettou, M. Sirivianos, J. Blackburn and N. Kourtellis, "The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans," *ACM Journal of Data and Information Quality*, vol. 11, no. 3, pp. 1-37, 2019.
- [63] P. Qi, J. Cao, T. Yang, J. Guo and J. Li, "Exploiting Multi-domain Visual Information for Fake News Detection," in *IEEE International Conference on Data Mining (ICDM)*, Beijing, China, 2019.
- [64] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Systems with Applications*, p. 112986, 2019.
- [65] "BUSINESS INSIDER," [Online]. Available: <https://www.businessinsider.com/2009/1/us-airways-crash-rescue-picture-citizen-journalism-twitter-at-work?IR=T>.
- [66] "WIKIPEDIA The Free Encyclopedia," [Online]. Available: [https://en.wikipedia.org/wiki/Sully_\(film\)](https://en.wikipedia.org/wiki/Sully_(film)).
- [67] M.-A. Kaufhold, M. Bayer and C. Reuter, "Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning," *Information Processing and Management*, vol. 57, no. 1, p. 102132, 2020.
- [68] K. Zahra, M. Imran and F. O. Ostermann, "Automatic identification of eyewitness messages on twitter during disasters," *Information Processing and Management*, vol. 57, no. 1, p. 102107, 2020.
- [69] B. Dore, "Foreign Policy-Fake News , Real Arrests," April 2020. [Online]. Available: <https://foreignpolicy.com/2020/04/17/fake-news-real-arrests/>.
- [70] Y. Liu, C. Gao, X. She and Z. Zhang, "A bio-inspired method for locating the diffusion source with limited observers," in *IEEE Congress on Evolutionary Computation (CEC)*, Vancouver, BC, Canada, 2016.
- [71] A. Louni and K. P. Subbalakshmi, "Who Spread That Rumor: Finding the Source of Information in Large Online Social Networks with Probabilistically Varying Internode Relationship Strengths," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 2, pp. 335-343, 2018.
- [72] D. Shah and T. Zaman, "Rumors in a network: Who's the culprit?," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5163 - 5181, 2011.
- [73] J. Choi, S. Moon, J. Woo, K. Son, J. Shin and Y. Yi, "Rumor source detection under querying with untruthful answers," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, Atlanta, GA, USA, 2017.
- [74] J. Choi, S. Moon, J. Shin and Y. Yi, "Estimating the rumor source with anti-rumor in social networks," in *IEEE 24th International Conference on Network Protocols (ICNP)*, Singapore, 2016.
- [75] K. Zhu and L. Ying, "Information Source Detection in the SIR Model: A Sample-Path-Based Approach," *IEEE/ACM Transactions on Networking*, vol. 24, no. 1, pp. 408 - 421, 2016.

- [76] X. Zhang, Y. Zhang, T. Lv and Y. Yina, "Identification of efficient observers for locating spreading source in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 442, pp. 100-109, 2016.
- [77] S. Khaled, N. El-Tazi and H. M. O. Mokhtar, "Detecting fake accounts on social media," in *IEEE International Conference on Big Data*, Seattle, Washington, USA, 2018.
- [78] F. C. Akyon and M. E. Kalfaoglu, "Instagram fake and automated account detection," in *Innovations in Intelligent Systems and Applications Conference (ASYU)*, Izmir, Turkey, 2019.
- [79] A. Louni and K. P. Subbalakshmi, "Who Spread That Rumor: Finding the Source of Information in Large Online Social Networks With Probabilistically Varying Internode Relationship Strengths," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 2, pp. 335-343, 2018.
- [80] J. Choi, S. Moon, J. Woo, K. Son, J. Shin and Y. Yi, "Rumor source detection under querying with untruthful answers.," in *IEEE Conference on Computer Communications*, Atlanta, USA, 2017.
- [81] L. Zhao, Q. Wang, J. Cheng, Y. Chen, J. Wang and W. Huang, "Rumor spreading model with consideration of forgetting mechanism: A case of online blogging LiveJournal," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 13, pp. 2619-2625, 2011.
- [82] L. Zhao, J. Wang, Y. Chen, Q. Wang, J. Cheng and H. Cui, "SIHR rumor spreading model in social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 7, pp. 2444-2453, 2012.
- [83] H. Qiyi, M. Fang and F. Wenjie, "Rumor Spreading and Monitoring Deployment in Online Social Networks," in *IEEE 17th International Conference on Communication Technology (ICCT)*, Chengdu, China, 2017.
- [84] G. J., L. W. and C. X., "The effect of the forget-remember mechanism on spreading," *The European Physical Journal B*, vol. 62, no. 2, pp. 247-255, 2008.
- [85] V. Indu and S. M. Thampi, "A nature - inspired approach based on Forest Fire model for modeling rumor propagation in social networks," *Journal of network and computer applications*, vol. 125, pp. 28-41, 2019.
- [86] M. Mendoza, B. Poblete and C. Castillo, "Twitter under crisis: Can we trust what we RT?," in *Proceedings of the first workshop on social media analytics.*, Washington, USA, 2010.
- [87] C. Jianhong, S. Qinghua and Z. Zhiyong, "Agent-Based Simulation of Rumor Propagation on Social Network Based on Active Immune Mechanism," *Journal of Systems Science and Information*, vol. 5, no. 6, pp. 571-584, 2017.
- [88] J.-J. Cheng, Y. Liu, B. Shen and W.-G. Yuan, "An epidemic model of rumor diffusion in online social networks," *The European Physical Journal B*, vol. 86, no. 1, pp. 1-7, 2013.
- [89] M. Nekovee, Y. Moreno, G. Bianconi and M. Marsili, "Theory of rumour spreading in complex social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 374, no. 1, pp. 457-470, 2007.

- [90] T. Mondal, P. Pramanik, I. Bhattacharya, N. Boral and S. Ghosh, “Analysis and Early Detection of Rumors in a Post Disaster Scenario,” *Information Systems Frontiers*, vol. 20, no. 5, pp. 961-979, 2018.
- [91] J. Ma, W. gao, S. Joty and K.-F. Wong, “An Attention-based Rumor Detection Model with Tree-structured Recursive Neural Networks,” *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 42, 2020.
- [92] J. Ma, W. Gao and K.-F. Wong, “Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 2017.
- [93] D. T. Vu and J. J. Jung, “Rumor Detection by Propagation Embedding Based on Graph Convolutional Network,” *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, pp. 1053 - 1065, 2021.
- [94] Z. Wu, D. Pi, J. Chen, M. Xie and J. Cao, “Rumor detection based on propagation graph neural network with attention mechanism,” *Expert Systems with Applications*, vol. 158, pp. 1-16, 2020.
- [95] T. H. Do, X. Luo, D. M. Nguyen and N. Deligiannis, “Rumour Detection Via News Propagation Dynamics and User Representation Learning,” in *IEEE Data Science Workshop (DSW)*, Minneapolis, MN, USA, 2019.
- [96] K. Eismann¹, “Difusion and persistence of false rumors in social media networks: implications of searchability on rumor self-correction on Twitter,” *Journal of Business Economics*, pp. 1-31, 2020.
- [97] X. Zhou, R. Zafarani, K. Shu and H. Liu, “Fake News: Fundamental Theories, Detection Strategies and Challenges,” in *Twelfth ACM international conference on web search and data mining.*, Melbourne, Australia, 2019.
- [98] C. Xueqin, Z. Fan, Z. Kunpeng, T. Goce, Z. Ting and Z. Fengli, “Information Diffusion Prediction via Recurrent Cascades Convolution,” in *IEEE 35th International Conference on Data Engineering (ICDE)*, Macao, 2019.
- [99] Y. Ishida and S. Kuraya, “Fake News and its Credibility Evaluation by Dynamic Relational Networks: A Bottom up Approach,” *Procedia Computer Science*, vol. 126, p. 2018, 2228-2237.
- [100] F. Wang, Y. Moreno and Y. Sun, “Structure of peer-to-peer social networks,” *Physical Review E*, vol. 73, no. 3, p. 036123, 2006.
- [101] G. Csányi and B. Szendrői, “Structure of a large social network,” *Physical Review E*, vol. 69, no. 3, p. 036131, 2004.
- [102] S. Edunov, D. Logothetis, C. Wang, A. Ching and M. Kabiljo, “Darwini: Generating realistic large-scale social graphs,” *arXiv preprint arXiv:1610.00664*, 2016.
- [103] A. Prat-Pérez, J. Guisado-Gámez, X. F. Salas, P. Koupy, S. Depner and D. B. Bartolini, “Towards a property graph generator for benchmarking,” in *Fifth international workshop on graph data-management experiences & systems*, Chicago, IL, USA, 2017.

- [104] T. G. Kolda, A. Pinar, T. Plantenga and C. Seshadhri, “A Scalable Generative Graph Model with Community Structure,” *SIAM Journal on Scientific Computing*, vol. 36, no. 5, pp. C424-C452, 2014.
- [105] A. Alexandrov, K. Tzoumas and V. Markl, “Myriad: scalable and expressive data generation,” *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 1890-1893, 2012.
- [106] D. Chakrabarti, Y. Zhan and C. Faloutsos, “R-MAT: A Recursive Model for Graph Mining,” in *Proceedings of the 2004 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics*, USA, 2004.
- [107] M. Hamann, U. Meyer, M. Penschuck, H. Tran and D. Wagner, “I/O-Efficient Generation of Massive Graphs Following the LFR Benchmark,” *ACM Journal of Experimental Algorithmics*, vol. 23, no. 1, pp. 1-33, 2018.
- [108] G. Bagan, A. Bonifati, R. Ciucanu, G. H. L. Fletcher, A. Lemay and N. Advokaat, “gMark: Schema-Driven Generation of Graphs and Queries,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 4, pp. 856-869, 2016.
- [109] Y. A. Belov and S. I. Vovchok, “Generation of a Social Network Graph by Using Apache Spark,” *Automatic Control and Computer Sciences*, vol. 51, no. 7, pp. 678-681, 2017.
- [110] K. Lakkaraju and G. Sukthankar, “Synthetic Generators to Simulate Social Networks(No. SAND2014-19132C),” in *Sandia National Lab.(SNL-NM)*, Albuquerque, NM (United States), 2014.
- [111] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li and J. Liu, “Content based fake news detection using knowledge graphs,” in *International semantic web conference*, Monterey, California, USA, 2018.
- [112] A. Gautam and K. R. Jerripothula, “SGG: Spinbot, Grammarly and GloVe based Fake News Detection,” in *IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, New Delhi, India, 2020.
- [113] Q. Li and W. Zhou, “Connecting the Dots Between Fact Verification and Fake News Detection,” in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, 2020.
- [114] M. K. Elhadad, K. F. Li and F. Gebali, “Detecting Misleading Information on COVID-19,” *IEEE Access*, vol. 8, pp. 165201-165215, 2020.
- [115] T. Pomari, G. Ruppert, E. Rezende, A. Rocha and T. Carvalho, “Image splicing detection through illumination inconsistencies and deep learning,” in *25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018.
- [116] S. Elkasrawi, S. S. Bukhari, A. Abdelsamad and A. Dengel, “What you see is what you get? Automatic Image Verification for Online News Content,” in *12th IAPR Workshop on Document Analysis Systems (DAS)*, Santorini, Greece, 2016.
- [117] F. Marra, D. Gragnaniello, D. Cozzolino and L. Verdoliva, “Detection of GAN-generated Fake Images over Social Networks,” in *IEEE Conference on Multimedia Information Processing and Retrieval*, Florida, USA, 2018.

- [118] Z. Jin, J. Cao, Y. Zhang, J. Zhou and Q. Tian, "Novel Visual and Statistical Image Features for Microblogs News Verification," *IEEE Transactiona on Multimedia*, vol. 19, no. 3, pp. 598-608, 2017.
- [119] Y. Yang, L. Zheng, J. Zhang, Q. Cui, X. Zhang, Z. Li and P. S. Yu, "TI-CNN: Convolutional Neural Networks for Fake News Detection," *arXiv preprint arXiv:1806.00749*, 2018.
- [120] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su and J. Gao, "EANN: Event Adversarial Neural Networks for Multi-Modal," in *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, London, United Kingdom, 2018.
- [121] D. Khattar, J. S. Goud, M. Gupta and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *The World Wide Web Conference*, San Francisco , USA, 2019.
- [122] P. Shah and Z. Kobti, "Multimodal fake news detection using a Cultural Algorithm with situational and normative knowledge," in *IEEE Congress on Evolutionary Computation (CEC)*, Glasgow (UK), 2020.
- [123] A. Giachanou, G. Zhang and P. Rosso, "Multimodal Fake News Detection with Textual, Visual and Semantic Information," in *International Conference on Text, Speech, and Dialogue*, Brno, Czech Republic, 2020.
- [124] D. K. Vishwakarma, D. Varshney and A. Yadav, "Detection and veracity analysis of fake news via scrapping and authenticating the web search," *Cognitive Systems Research*, vol. 58, pp. 217-229, 2019.
- [125] P. Meel and D. K. Vishwakarma, "HAN, image captioning, and forensics ensemble multimodalfake news detection," *Information Sciences*, vol. 567, pp. 23-41, 2021.
- [126] S. Vosoughi, D. Roy and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146-1151, 2018.
- [127] K. Shu, D. Mahudeswaran, W. Suhan, D. Lee and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media," *Big data*, vol. 8, no. 3, pp. 171-188, 2020.
- [128] V. Indu and S. M. Thampi, "A nature - inspired approach based on Forest Fire model for modeling rumor propagation in social networks," *Journal of Network and Computer Applications*, vol. 125, pp. 28-41, 2019.
- [129] O. Varol, E. Ferrara, F. Menczer and A. Flammini, "Early detection of promoted campaigns on Social Media," *EPJ Data Science*, vol. 6, no. 1, pp. 1-13, 2017.
- [130] C. Castillo, M. Mendoza and B. Poblete, "Information Credibility on Twitter," in *Proceedings of the 20th international conference on World wide web*, Hyderabad, India, 2011.
- [131] M. D. Vicario, W. Quattrociocchi, A. Scala and F. Zollo, "Polarization and Fake News: EarlyWarning of Potential Misinformation Targets," *ACM Transactions on the Web*, vol. 13, no. 2, pp. 1-22, 2019.

- [132] S. Hamidian and M. Diab, “Rumor Detection and Classification for Twitter Data,” in *SOTICS 2015 : The Fifth International Conference on Social Media Technologies, Communication, and Informatics*, Barcelona, Spain, 2015.
- [133] Z. Jin, J. Cao, H. Guo, Y. Zhang and J. Luo, “Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs,” in *Proceedings of the 25th ACM international conference on Multimedia*, Mountain View California ,USA, 2017.
- [134] M. Alrubaian, M. Al-Qurishi, M. M. Hassan and A. Alamri, “A Credibility Analysis System for Assessing Information on Twitter,” *IEEE Transactions on Dependable And Secure Computing*, vol. 15, no. 4, pp. 661-674, 2018.
- [135] O. Ajao, D. Bhowmik and S. Zargari, “Sentiment Aware Fake News Detection on Online Social Networks,” in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton , United Kingdom, 2019.
- [136] S. Kumar, R. Asthana, S. Upadhyay, N. Upreti and M. Akbar, “Fake news detection using deep learning models: A novel approach,” *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 2, 2020.
- [137] F. Marra, D. Gragnaniello, D. Cozzolino and L. Verdoliva, “Detection of GAN-generated Fake Images over Social Networks,” in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)* , Florida, USA, 2018.
- [138] B. D. Horne, J. Nørregaard and S. Adali, “Robust Fake News Detection Over Time and Attack,” *ACM Transactions on Intelligent Systems and Technology* , vol. 11, no. 1, pp. 1-23, 2019.
- [139] Y. Liu and Y.-F. B. Wu, “FNED: A Deep Network for Fake News Early Detection on Social Media,” *ACM Transactions on Information Systems*, vol. 38, no. 2, pp. 1-33, 2020.
- [140] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff and B. Stein, “A Stylometric Inquiry into Hyperpartisan and Fake News,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018.
- [141] Á. I. Rodríguez and L. L. Iglesias, “Fake News Detection Using Deep Learning,” *arXiv preprint arXiv:1910.03496*, 2019.
- [142] H. Jwa, D. Oh, K. Park, J. M. Kang and H. L. 1, “exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT),” *Applied Sciences*, vol. 9, no. 19, 2019.
- [143] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru and S. Satoh, “SpotFake: A Multi-modal Framework for Fake News Detection,” in *IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, Singapore, 2019.
- [144] J. E. v. Engelen and H. H. Hoos, “A survey on semi-supervised learning,” *Machine Learning*, vol. 109, no. 2, p. 373–440, 2020.
- [145] S. Laine and T. Aila, “Temporal Ensembling for Semi -supervised Learning,” in *International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

- [146] X. Shi, H. Su, F. Xing, Y. Liang, G. Qu and L. Yang, "Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis," *Medical Image Analysis*, vol. 60, 2020.
- [147] J. K. Rout, A. Dalmia, K.-K. R. Choo, S. Bakshi and S. K. Jena, "Revisiting semi-supervised learning for online deceptive review detection," *IEEE Access*, vol. 5, pp. 1319-1327, 2017.
- [148] C. Helwe, S. Elbassuoni, A. A. Zaatari and W. El-Hajj, "Assessing Arabic Weblog Credibility via Deep Co-learning," in *Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy, 2019.
- [149] G. B. Guacho, S. Abdali, N. Shah and E. E. Papalexakis, "Semi-supervised Content-based Detection of Misinformation via Tensor Embeddings," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Barcelona, Spain, 2018.
- [150] J. Gaglani, Y. Gandhi, S. Gogate and A. Halbe, "Unsupervised WhatsApp Fake News Detection using Semantic Search," in *IEEE 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2020.
- [151] S. Hosseinimotlagh and E. E. Papalexakis, "Unsupervised content-based identification of fake news articles with tensor decomposition ensembles," in *Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*, Los Angeles, California, USA, 2018.
- [152] W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau and B. S. Lee, "Unsupervised rumor detection based on users' behaviors using neural networks," *Pattern Recognition Letters*, vol. 105, pp. 226-233, 2018.
- [153] S. Yang, K. Shu, S. Wang and F. W. ., H. L. Renjie Gu, "Unsupervised Fake News Detection on Social Media: A generative approach," in *AAAI Conference on Artificial Intelligence*, Hawaii, USA, 2019.
- [154] S. C. R. Gangireddy, D. P, C. Long and T. Chakraborty, "Unsupervised Fake News Detection: A Graph-based Approach," in *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, New York, USA, 2020.
- [155] M. Miyabe, A. Nadamoto and E. Aramaki, "How do rumors spread during a crisis? Analysis of rumor expansion and disaffirmation on Twitter after 3.11 in Japan.," *International Journal of Web Information Systems*, vol. 10, no. 4, pp. 394-412, 2014.
- [156] N. P. Nguyen, G. Yan, M. T. Thai and S. Eidenbenz, "Containment of misinformation spread in online social networks," in *Proceedings of the 4th Annual ACM Web Science Conference*, New York, USA, 2012.
- [157] S. Kate, M. Jim, O. Mania, A. Peg and M. R. M., "Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing," in *ICoNference*, Berlin, Germany, 2014.
- [158] Y. Wu, H. Huang, J. Zhao, C. Wang and T. Wang, "Using Mobile Nodes to Control Rumors in Big Data Based on a New Rumor Propagation Model in Vehicular Social Networks," *IEEE Access*, vol. 6, pp. 62612-62621, 2018.

- [159] S. Wen, J. Jiang, Y. Xiang, S. Yu, W. Zhou and W. Jia, "To shut them up or to clarify: Restraining the spread of rumors in online social networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 12, pp. 3306-3316, 2014.
- [160] K. Shu, D. Mahudeswaran, S. Wang, D. Lee and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media," *arXiv preprint arXiv:1809.01286*, 2018.
- [161] S. Tschatschek, A. Singla, M. G. Rodriguez and A. Merchant, "Fake News Detection in Social Networks via Crowd Signals," in *Companion Proceedings of the The Web Conference 2018*, Lyon, France, 2018.
- [162] J. Roozenbeek and S. v. d. Linden, "The fake news game: actively inoculating against the risk of misinformation," *Journal of Risk Research*, vol. 22, no. 5, pp. 570-580, 2019.
- [163] K. P. K. Kumar and G. Geethakumari, "Detecting misinformation in online social networks using cognitive psychology," *Human-centric Computing and Information Sciences*, vol. 4, no. 1, pp. 1-22, 2014.
- [164] S. Vosoughi, M. ' . Mohsenvand and D. Roy, "Rumor Gauge: Predicting the Veracity of Rumors on Twitter," *ACM transactions on knowledge discovery from data*, vol. 11, no. 4, pp. 1-36, 2017.
- [165] O. Kyle, S. Olga and W. Frederick, "Collective Classification for Social Media Credibility Estimation," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, Hawaii, 2019.
- [166] K. Shu, S. Wang and H. 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*, Melbourne, Australia, 2019.
- [167] K. Shu, S. Wang and H. Liu, "Understanding User Profiles on Social Media for Fake News Detection," in *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*., Miami, FL, USA, 2018.
- [168] M. Lukasik, P. Srijith, D. Vu, K. Bontcheva, A. Zubiaga and T. Cohn, "Hawkes Processes for Continuous Time Sequence Classification: an Application to Rumour Stance Classification in Twitter," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2016.
- [169] X. Liu, A. Nourbakhsh, Q. Li, R. Fang and S. Shah, "Real-time Rumor Debunking on Twitter," in *24th ACM international on conference on information and knowledge management.*, Melbourne; Australia, 2015.
- [170] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu and H. Liu, "Unsupervised Fake News Detection on Social Media: A Generative Approach," in *Proceedings of the AAAI conference on artificial intelligence.*, Hawaii, USA , 2019.
- [171] M. Aldwairi and A. Alwahedi, "Detecting fake news in social media networks," *Procedia Computer Science* , vol. 141, pp. 215-222, 2018.

- [172] D. K. Vishwakarma, D. Varshney and A. Yadav, "Detection and veracity analysis of fake news via scrapping and authenticating the web search," *Cognitive Systems Research*, vol. 58, pp. 217-229, 2019.
- [173] B. Botnevik, E. Sakariassen and V. Setty, "BRENDA: Browser Extension for Fake News Detection," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, China, 2020.
- [174] D. Varshney and D. K. Vishwakarma, "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences," *Applied Intelligence*, pp. 1-22, 2021.
- [175] P. Meel and D. K. Vishwakarma, "A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles," *Expert Systems with Applications*, vol. 177, pp. 1-14, 2021.
- [176] "Fake News Detection," 2017. [Online]. Available: <https://www.kaggle.com/jruvika/fake-news-detection>.
- [177] "Fake News Data," 2018. [Online]. Available: <https://www.kaggle.com/c/fake-news/data>.
- [178] "Fake News Sample," 2018. [Online]. Available: <https://www.kaggle.com/pontes/fake-news-sample>.
- [179] A. P. S. Bali, M. Fernandez, S. Choubey, M. Goel and P. K. Roy, "Comparative Performance of Machine Learning Algorithms for Fake News Detection," in *International Conference on Advances in Computing and Data Sciences*, Springer, Singapore, 2019.
- [180] K. Agarwalla, S. Nandan, V. A. Nair and D. D. Hema, "Fake News Detection using Machine Learning and Natural Language Processing," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 7, no. 6, pp. 844-847, 2019.
- [181] H. Karimi and J. Tang, "Learning Hierarchical Discourse-level Structure for Fake News Detection," in *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019.
- [182] O. Ajao, D. Bhowmik and S. Zargari, "Fake news identification on twitter with hybrid cnn and rnn models," in *Proceedings of the 9th International Conference on Social Media and Society*, Copenhagen, Denmark, 2018.
- [183] M. J. Kusner, Y. Sun, N. I. Kolkin and K. Q. Weinberger, "From word embeddings to document distances," in *International conference on machine learning (PMLR)*, Lille, France, 2015.
- [184] Y. Rubner, C. Tomasi and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, no. 2, pp. 99-121, 2000.
- [185] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
- [186] "real_or_fake," 2018. [Online]. Available: <https://www.kaggle.com/rchitic17/real-or-fake>.
- [187] "BBC News," [Online]. Available: <https://www.bbc.com/news/world-41821923>.

- [188] “Snopes,” [Online]. Available: <https://www.snopes.com/fact-check/iss-test-marijuana-space/>.
- [189] “Business insider India,” [Online]. Available: <https://www.businessinsider.in/politics/a-photo-of-trump-and-other-leaders-staring-at-putin-is-going-viral-but-its-fake/articleshow/59536672.cms>.
- [190] “Hoaxes,” [Online]. Available: http://hoaxes.org/weblog/comments/frozen_venice.
- [191] “FACTCHECK,” [Online]. Available: <https://www.factcheck.org/2020/02/fake-coronavirus-cures-part-2-garlic-isnt-a-cure/>.
- [192] “CNBC,” [Online]. Available: <https://www.cnbc.com/2020/02/07/containing-coronavirus-means-fighting-misinformation-singapore-minister.html>.
- [193] D. Tutorials, “Web Scrapping with Python,” 2020. [Online]. Available: <https://www.datacamp.com/community/tutorials/amazon-web-scrapping-using-beautifulsoup>.
- [194] N. Krawetz, “A picture’s worth : Digital Image Analysis and Forensics,” Hacker Factor Solutions, Black Hat Briefings , 2007.
- [195] “All Data,” [Online]. Available: <https://drive.google.com/file/d/0B3e3qZpPtccsMFo5bk9Ib3VCc2c/view>.
- [196] Z. Jin, J. Cao, H. Guo, Y. Zhang and J. Luo, “Multimodal Fusion with Recurrent Neural Networks for,” in *Proceedings of the 25th ACM international conference on Multimedia*, Mountain View California ,USA, 2017.
- [197] F. Lago, Q.-T. Phan and G. Boato, “Visual and Textual Analysis for Image Trustworthiness Assessment within Online News,” *Security and Communication Networks*, pp. 1-14, 2019.
- [198] “MediaEval 2016,” 2016. [Online]. Available: <https://github.com/MKLab-ITI/image-verification-corpus/tree/master/mediaeval2016>.
- [199] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [200] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, “Attention is all you need,” *arXiv preprint arXiv:1706.03762*, 2017.
- [201] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” *arXiv preprint arXiv:1909.11942* , 2019.
- [202] C. Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, 2017.
- [203] Z. Jin, J. Cao, H. Guo, Y. Zhang and J. Luo, “Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs,” in *Proceedings of the 25th ACM international conference on Multimedia*, Mountain View California ,USA, 2017.

Author Biography

Priyanka Meel

(2K18/Ph.D./IT/08)

Department of Information Technology,
Delhi Technological University, Delhi, India



Priyanka Meel received a Bachelor of Technology (B.Tech.) Degree from the Indian Institute of Information Technology and Management, Gwalior, India, in 2011 and Master of Technology (M.Tech.) Degree from the Indian Institute of Information Technology and Management, Gwalior, India, in 2013. She joined as an Assistant Professor in the Department of Information Technology, Delhi Technological University, New Delhi, India, in 2016. She enrolled for a Doctor of Philosophy (Ph. D.) degree in the Department of Information Technology, Delhi Technological University, New Delhi, India in 2018. Currently, she is working towards the fulfillment of Ph. D. degree. Her current research interests include Artificial Intelligence, Data Analytics, Fake News Detection, Image Processing, Pattern Analysis, Machine Learning, and Deep Learning.