# Multimodal Fake News Analysis and Detection

Student Name: Mudit Dhawan
Roll Number: 2018159

BTP report submitted in partial fulfillment of the requirements
for the Degree of B.Tech. in Electronics and Communication Engineering
on December 11, 2021

**BTP Track**: Research Track

**BTP Advisors**
Dr. Ponnurangam Kumaraguru
Dr. Rajiv Ratn Shah

Indraprastha Institute of Information Technology
New Delhi

## Student's Declaration

I hereby declare that the work presented in the report entitled **Multimodal Fake News Analysis and Detection** submitted by me for the partial fulfillment of the requirements for the degree of *Bachelor of Technology* in *Electronics & Communication Engineering* at Indraprastha Institute of Information Technology, Delhi, is an authentic record of my work carried out under guidance of **Dr. Ponnurangam Kumaraguru** and **Dr. Rajiv Ratn Shah**. Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

**..............................** **Place & Date: .............................**
**Mudit Dhawan**

## Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

**..............................** **.............................**
**Dr. Ponnurangam Kumaraguru** **Dr. Rajiv Ratn Shah**

**Place & Date: .............................** **Place & Date: .............................**

## Abstract

Fake News has become the curse of our time. Online social media networks provide a low-cost platform to facilitate information and fact sharing, but it fails to offer any quality control. As the number of people receiving their daily news through these platforms increases, it becomes a significant problem for the government and other organizations. Fake News articles leverage the multimedia content posted on the platforms and mislead the reader through fabricated image(s) or text (title and text body) accompanying it. Many organizations have started an initiative to provide de-bunked fake news, i.e., fact-checked and verified counterfeit news items floated on various social media platforms by human fact-checkers. Though this human intervention is a good start towards eradicating this evil, it can not be feasible at a larger scale providing human fact-checked information for every post on social media. The scalability of this human fact-checked information isn't the only issue, but the promptness of such accurate information becomes crucial in this digital age. To address this problem, we aim to analyze multimodal fake content from platforms supporting online journalism (including various social media platforms) to extract meaningful features and design an all-inclusive early-stage Automated Fake News Detection System.

Keywords: Fake News, Online Social Media Networks(OSNs), Machine Learning, Multimodal Anlaysis, Multimodal Embeddings, Multimodal fusion

## Acknowledgments

First and foremost, I would like to express my heartfelt gratitude towards Dr. Ponnurangam Kumaraguru and Dr. Rajiv Ratn Shah for always guiding me irrespective of their busy schedule. I learned a lot from both of them and would always be grateful to them for giving me the opportunity to work on this ambitious project. Further, I would like to sincerely thank Ms. Shivangi Singhal (Ph.D. Scholar at IIIT-D) and other members of Precog research group for their constant support and help throughout the project. This thesis would not have been possible without their valuable inputs.

## Work Distribution

The report describes the work I have done till now as my contribution to my B.Tech. Project. This work has been done in collaboration with Ms. Shivangi Singhal (Ph.D. Scholar at IIIT-D).

# Contents

**1 Introduction** **1**

1.1 What is Fake News? . . . 1

1.2 What is Fact-Checked or Debunked Fake News? . . . 1

1.3 Issue of domain shift and emergent fake news . . . 1

1.4 Problem Statement . . . 2

1.4.1 Main objective . . . 3

1.5 Motivation . . . 3

1.6 Contributions . . . 4

**2 Literature Survey** **5**

2.1 Text-Based Fake News Detection . . . 5

2.2 Image(s)-Based Fake News Detection . . . 5

2.3 Multi-Modal Single-Image Fake News Detection . . . 6

2.4 Multi-Modal Multiple-Images Fake News Detection . . . 6

2.5 Explainable Fake News Detection . . . 7

2.6 Multi-Task Approach to Fake News Detection . . . 7

2.7 Time and Data Sensitive Fake News Detection . . . 8

2.8 Other Novel Frameworks . . . 9

**3 Multi-Modal Fake News Analysis** **10**

3.1 Regional Language Fact-Checked Dataset . . . 10

3.1.1 Fake Modality . . . 11

3.1.2 Movement of Fake News . . . 11

3.1.3 Cross Lingual Dynamics of Fake News . . . 14

**4 Proposed Fake News Detection Systems** **17**

4.1 Leveraging Inter-modality Discordance for Multimodal Fake News Detection . . . 17

4.1.1 Introduction . . . 17

4.1.2 Problem Statement . . . 18

4.1.3 Methodology . . . 18

4.1.4 Experiments and Results . . . . . . . . . . . . . . . . . . . . . . . . . . . 22

4.1.5 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

4.2 Cross-Attention Based Model for Explainable Fake News Detection . . . . . . . . 27

4.2.1 Methodology . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

4.2.2 Dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29

4.2.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29

4.2.4 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

**5 Future Works 32**

**6 Limitations 33**

# Chapter 1

# Introduction

## 1.1 What is Fake News?

Fake news has prevailed for a long time in society, and during the course, multiple definitions were proposed for it. All the definitions align with a general consensus that defines *Fake news* as a piece of intentionally and verifiably wrong information, which could mislead the readers [2]. The authors of such bogus news often benefit from the fake content through a shift in power, discrediting an opponent, money, or other malicious incentives. It posses an enormous threat on the inherent values of democracy, journalism, and freedom of expression [66].

## 1.2 What is Fact-Checked or Debunked Fake News?

The fight against Fake News is spearheaded by many organizations, which try to assess the information posted on social media and other news outlets based on the facts that it contains and presents a thorough investigation into whether the news is fraudulent or credible. The International Fact-Checking Network (IFCN)[1] provides some basic ethics that the Fact-Checking platforms need to follow. It provides a certification based on the platform's ability to abide by them and present impartial fact-checked news. Such organizations provide a structured report containing the *claim* of the news sample, information about the *claimant* followed by the *investigation* carried out by a member of the fact-checking organization, and the *verdict* on the veracity of the news sample.

## 1.3 Issue of domain shift and emergent fake news

Previous research in fake news detection has focused on publicly available large-scale labeled datasets. However, this experimental setting is far from what happens in real life. The issue of emergent fake news comes into the picture: news related to topics or events previously unseen

[1] https://www.poynter.org/ifcn/

| **Topic** | **Count** | **Month** |
|---|---|---|
| elephant | 13 | January |
| sochi | 274 | Feburary |
| malaysia | 310 | March |
| underwater | 113 | April |
| bringback | 131 | May |
| passport | 44 | June |
| columbianChemicals | 185 | September |
| syrianboy | 1786 | November |

Table 1.1: Event distribution in the Mediaeval 2016 dataset for the year 2014. The topics present were *malaysia, passport, sochi, bringback, columbianChemicals, elephant, underwater, and syrianboy*, which occurred in non-overlapping months, thus indicating the problem of domain shift

by the trained model in its training data. One simple solution to tackle such a problem could be re-training the model for the new event, but that would still require high amounts of labeled data for this new event, which would further increase the manual labor hours needed for a working system. Another problem of detecting a domain shift would add to the issues, as one would need to accurately predict when a model needs to be re-trained, as there would be no ground-truth labels to compare the outdated model. Even if we can mitigate the above problem of detecting new events and gathering labeled data, we would still need to re-train or fine-tune the fake news classification model, which would require high computational and memory resources to store such event-specific models. Though most publicly available datasets are collected based on a single topic with a fixed time frame, the problem of domain shift is not that prevalent within each dataset. However, the Mediaeval dataset [5] does shed some light on the above problem, where the topic of the news present in the dataset changes to 8 different events in a span of 12 months in a non-overlapping fashion. The table 1.1 shows the event distribution present in the 2014 year subset of the mediaeval2016 dataset.[2]

## 1.4 Problem Statement

Social Media has become the bread and butter for an average person's news intake. As of August 2018, around two thirds (68%) of Americans get their news from social media.[3] In this new ecosystem of instant information sharing, the traditional text-only news has transformed into multimedia-content rich story-telling news to engage more readers. There are thousands of news articles proliferating social media platforms that continue to shape the readers' views. The effect of fake news amplifies due to the echo-chamber effect [21] and validity effect [4] because of its repetition in a user's social network. Previous studies on the propagation patterns of fake news on social media shows that fake news (mostly political) travels faster (retweets on Twitter and shares on Facebook) than real news [57].

A news article posted on social media can have two broad types of features (1) Content-based

[2] https://github.com/MKLab-ITI/image-verification-corpus/tree/master/mediaeval2016

[3] https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

and (2) Social Network/ Propagation based features. Considering the catastrophic damage fake news can cause in its first few hours [16], it becomes essential to develop early-stage Fake news detection systems. Nevertheless, propagation-based features can be leveraged to detect patterns in users' activity to classify better the news posted by the user in a time-sensitive manner.

### 1.4.1 Main objective

The aim of the project is to design an all-inclusive Fake News detection system. It would be able to utilize the following information to provide *explainable and early-stage* predictions:

1. Textual content (headline, main body)
2. Visual content (variable number of images)

The proposed system would try to mimic the human fact-checking style. In a scenario where no previous knowledge or required fact-base is available to the system, the prediction would be based on the Textual, Visual, Source-based features (Content), and the Social Network features of the article. Otherwise, it would take into consideration the previously established facts to better classify Fake News Articles.

## 1.5 Motivation

Fake news has proved to be one of the virtues of low-cost news and information accessing and disseminating platforms on the internet. Corrupt and malicious people use these platforms to mislead people with false information for some gain in return, be it monetary or political interest. Fake news is crafted in such a manner that an ordinary reader without any domain knowledge might fall into the trap and believe the news content, which leads to the formation of thought and perspective based on bogus knowledge. Previous studies have extensively explored fake news detection from the point of traditional text-based news. However, in this age of social media platforms and other digital news outlets, analyzing the multimedia content for fabricated news items becomes essential. News articles or posts leverage these visual cues to create a better storytelling environment to attract a reader's attention. Fake News creators take advantage of this fact and add controversial or fabricated images to mislead and deceive the readers for rapid dissemination. The number of posts on social media every hour is enormous, and to have them fact-checked by humans is not merely possible. Even if it were possible on such a large scale, there would be no control over the fact-checker's bias in such a large pool, and their own beliefs might creep into the process. We need an automated system to flag such fake news posts to tackle this significant problem. For this, the model should look through the effort put in by fake news creators to disguise it as a real one. Such an all-inclusive fake news detection system would be a big step in the fight against fake news.

## 1.6 Contributions

The main contributions of our work can be summarized as follows:

- To the best of our knowledge, we are the first to propose a fake news detection framework that leverages all the different multimedia components present in a news article (i.e., headline, text-body/ content, multiple visual cues).
  - The embedding technique proposed in this model is a general multimodal feature extraction framework that can be integrated with different fusion and classification models.
- We pose multi-modal fake news detection as a *joint-learning* problem with the main task of binary classification, along with an auxiliary task of modeling Inter-Modality Discordance.
  - To this end, we propose an early-stage Inter-modality Discordance based Multimodal Fake News Detection Model. In this end-to-end multi-modal framework, modal-specific discriminative features are incorporated by utilizing the cross-entropy loss, along with a modified version of contrastive loss that explores inter-modality discordance.
- We propose a deep multimodal attention-based model that learns to detect fake news in earliness. We use the attention maps generated for both text and image separately to add a novel explainability factor to our Fake News Detection System.
  - We propose a Cross-Attention Module that generates a multimodal feature query to extract important visual and textual features to enable coherent feature extraction.
- We empirically show that our two proposed Multimodal Fake News Detection models can effectively identify fake news and outperform the state-of-the-art multimodal fake news detection models on a popular large-scale real-world dataset.
  - Our joint-learning approach outperforms the state-of-the-art by an average F1-score of *6.3%*.
- We create a public GitHub repository that contains Pytorch implementations of previously published Multimodal Fake News Detection systems to help new researchers entering the field.[4]

[4] https://github.com/MUDITDHAWAN/Multimodal-Fake-News-Detection-Systems

# Chapter 2

# Literature Survey

## 2.1 Text-Based Fake News Detection

Previous studies have extensively looked at explicit feature extraction for textual data, including statistical or semantic features. However, these hand-engineered features are challenging to generate and highly event/ domain-specific [44]. Feng et al. [12] proposed a new angle to the previously used shallow lexico-syntactic patterns by adding context-free grammar parse trees. Gupta et al. [15] used linguistic features such as frequency of swear words, negative or positive sentiment words, pronouns with other information extracted from tweets to give a real-time credibility score to tweets. To solve the problem of hand-engineered features, Ma et al. [32] employed LSTM networks to model the time-series textual information for an event and predict its veracity. Ruchansky et al. [44] proposed a hybrid model (CSI), which incorporated LSTMs to extract textual features in its Capture module along with the Score module user-based features. Rubin et al. [43] introduced Rhetorical Structure Theory and a Vector-Space model to analytically capture a story's coherence in terms of functional relationships among text units.

## 2.2 Image(s)-Based Fake News Detection

Research into single-modal manipulated content has been dominated by text-based content direction. However, in recent years there has been a shift to image-based fake news detection. Qi et al. [40] divide the manipulated images available on different platforms into two categories: (i) tampered images and (ii) misleading images. Tampered images can be detected using just the image present in the source, whereas to understand if the image present is misleading or not, it is necessary to take into account the test accompanying the image. Jin et al. [22] explored the role of images in automatic fake news detection. The authors presented a list of visual and statistical features that can quantitatively describe image distribution patterns of images and help with manipulated content detection.

On similar lines, Qi et al. [40] presented a novel Multi-domain Visual Neural Network (MVNN) to exploit an image's inherent characteristics in two domains. They have used CNN-based architecture to capture complex re-compression and other tampering artifacts in the frequency domain and CNN-RNN frameworks to extract features of different semantic levels in the pixel(spatial) domain. Huh et al. [20] proposed a learning algorithm for detecting visual image manipulations. They introduce the notion of "self-consistent" images (i.e., content produced by a single image pipeline, which they detect using the automatically recorded photo EXIF metadata. Their self-consistency algorithm focuses on detecting and localizing image splices, using a Siamese Network with a Resnet-50 backbone.

## 2.3 Multi-Modal Single-Image Fake News Detection

There has been extensive research to find hand-crafted features for the textual information present in traditional news articles, as discussed above. However, in this age of social media, fake news curators take advantage of the multimedia content to mislead fellow readers. They use images or videos to engage with the readers and fabricate a lie. Unfortunately, this sub-problem is yet to receive the recognition it deserves to tackle fake news in the digital age.

Yang et al. [62] proposed Ti-CNN architecture to extract explicit and latent representations for both text and images. For latent features, convolutional layers and max-pooling layers were used for both the modalities. Whereas the explicit text-based features were calculated using text statistics, and image-based explicit features contained information about the number of faces in the images and its resolution. Khatter et al. [23] used a Variational Autoencoder-based architecture to extract shared multimodal data representations (MVAE), which learns correlations across the tweets' text and images extracted using Bi-LSTM network and VGG-19, respectively. Wang et al. [58] proposed a novel event-discriminator module in their EANN architecture, where its role was to remove event-specific features using adversarial-style learning. For textual and feature extraction Text-CNN and VGG-19 networks were used respectively and then concatenated to find a multimodal representation. Zhou et al. [65] focused on the relationship between the visual and textual information in a news article and introduced a new similarity-aware fake news detection method. The feature extraction methods employed in this method were slightly different as they first converted the image into text using a pre-trained image-to-sequence model. They then used cosine similarity to capture cross-modal similarity on the text-CNN model's features on the two types of textual data. Shivangi et al. [53,54] leveraged pre-trained transformer-based textual encoders [10, 63] and VGG for image feature extraction to classify fake news via transfer learning.

## 2.4 Multi-Modal Multiple-Images Fake News Detection

A news article generally comprises two text components (i.e., title and content) and image(s) to support/ refute the claim in a more holistic manner. If there are multiple images present in an article, the first image is typically known as the top-image, and the rest of the image(s) are called other-images. Studies discussed in section 2.3 consider only the top-image (first image) and discard other images during classification.

Recently, a single study by Giachanou et al. [13] introduced a new direction by exploiting the information from multiple images along with the headline of the articles. The authors proposed a multi-image module that consists of a CNN-LSTM-based architecture and BERT [10] as a visual and textual feature extractors, respectively. For the CNN pipeline, a pre-trained VGG-16 [52] network to extract semantic and image-tag features. the semantic features are fed into an LSTM [19] layer to encapsulate the feature vector from the multiple input images. The outputs from the hidden states of the LSTM cells are max-pooled to form a final image feature to be used in fake news detection. They also incorporate similarity amongst the different components by calculating the cosine similarity between the title and the top-10 tags found for each image. Next, they concatenate the textual (title), similarity (amongst the title and top-10 image tags), and visual (max-pooled hidden state) features followed by an attention layer to predict the label for the given article.

## 2.5 Explainable Fake News Detection

Explainable AI has become an essential field in this data-hungry deep learning paradigm [14]. Models with millions of parameters are used as black-boxes to solve classification tasks with high accuracies, without any insight into how the prediction was made. Explainable Fake News detection system dive into the reasoning behind why a particular news article was classified as bogus. Yang et al. [60] proposed the XFake model to interpret the fakeness of a news article. They used explicit feature extraction modules for attribute, semantic, and linguistic statement analysis of only-text based articles. These extracted features were used as inputs for an ensembling technique, XGBoost shallow model, to explain the classification by calculating each attribute's feature importance. Shu et al. [50] proposed a novel way to incorporate other users' comments on the article using a sentence-comment co-attention module to capture top-k check worthy sentences to explain the classification.

## 2.6 Multi-Task Approach to Fake News Detection

Many researchers have recently looked into how fake news classification tasks can be augmented by jointly optimizing them with other relevant tasks such as topic classification and stance classification to boost performance for the main task at hand. Studies that look into the multi-task learning methodologies have highlighted its effectiveness in improving performance on a single task with shared information from other related tasks, which helps learn more meaningful representations and acts as a regularizer for the primary task [31]. Ma et al. [33] took advantage of previous studies that established false rumors to invoke more controversies than true news, which they included into their model as a stance detection task. Since then, many studies have dug deeper into this association between rumor and stance detection to create a more generalizable classifier for news posted on online social media platforms [28, 59]. Other than stance, researchers have also experimented with topic and sentiment classification. A novel deep learning-based method, SAME, incorporated multimodal information using Graph Affinity and Local Similarity Metrics and an adversarial loss to bridge the data distribution gap between various modalities [7]. The FDML model introduced an auxiliary task of topic classification to improve the performance, based on insights into topic-based credibility distributions; that is, specific topics attract higher proportions of false news than others [30].

## 2.7 Time and Data Sensitive Fake News Detection

Researchers have applied different versions of existing techniques to implement domain adaptations to perform time-sensitive and data-efficient fake news detection to handle the domain shift problem with every emerging event.

- **Incremental Training or Fine-tuning on a new dataset**
  It is one of the most straightforward ways of dealing with new data, and in the age of large models pre-trained on terabytes of data and then fine-tuning on the downstream task in hand, it is also called fine-tuning. The previously trained model on outdated or different topic data is used as a checkpoint to start training with the updated data. However, this method fails to generalize well over multiple sequential tasks. This problem is termed as catastrophic forgetting, where the model overwrites the weights learned for the former tasks with the updated weights to fit the next sequential task, thereby decreasing its performance on the previous tasks [26]. Not only does the problem of unlearning previous tasks/ news domains arises in this method, but also there is a general lack of large-scale labeled datasets for the model to learn the new event to mitigate the negative impacts of false news spreading on fast and connected social media platforms.

- **Continual learning**
  Current state-of-the-art models fail to perform well in scenarios where the data is acquisitioned in a continual manner instead of a stationary one-time data-collection drive. Such incremental data sources with non-stationary data distributions lead to catastrophic forgetting. Multiple techniques have been devised to alleviate neural networks and other machine learning algorithms of this problem and achieve the dream of lifelong learning, which can be divided into three broad categories [37]: (i) regularization based approaches [29], (ii) dynamic architectures [45], and (iii) complementary learning systems and memory replay [27]. In a recent paper utilizing the advancements in graphical neural networks to attack the problem of fake news detection without the content (top or event dependent information), the authors apply Gradient Episodic Memory (GEM) and Elastic Weight Consolidation (EWC) to tackle the problem of catastrophic forgetting [18].

## 2.8 Other Novel Frameworks

Recent advancements in Fake News Detection have produced some remarkable insight into the problem. Qian et al. [41] proposed the TCNN-URG model. The User Response generation Module learns to generate user responses to a news article using previous user responses to assist the Fake news detection. To extract the semantic information from the text, they have used a two-layer convolution neural network. The URG module works on the principle of conditional variational autoencoders. It learns to generate new user responses based on the historical responses and conditioned on the news article's features in question. Ma et al. [34] used a GAN-style network where two generators were trained opposite to a rumor discriminator. The generators added opposing or skeptical voices against valid claims and supportive voices towards rumors so that the discriminator is forced to learn more substantial rumor indicative representations from the data. Zellers et al. [64] used the adversarial training regimes used in natural language generation to machine-written fake news articles. They proposed a GPT [42] style model, Grover, which specialized in generating long conditionally generated text by modeling a news article as a joint probability distribution of the *domain, date, authors, headline, and body.*

# Chapter 3

# Multi-Modal Fake News Analysis

To build a better fake news detection system it was crucial to first understand the fake news sample being posted on different social media platforms, to asses them on the type of attributes that can be essential for classification.

## 3.1 Regional Language Fact-Checked Dataset

For this, the Indian Regional Languages Fact-checked dataset presented by Shivangi et al. [55]. The regional dataset contains 22,435 samples from 14 different languages spoken in India. Table 3.1 lists the different languages included in the dataset, along with the number of samples. These samples were collected from IFCN certified fact-checking websites for Indian regional languages- ALT News,[1] Boom,[2] Digiteye,[3] Fact Checker,[4] Fact Crescendo,[5] Factly,[6] News Mobile,[7] Vishwas,[8] and Webqoof[9]. Each organization provides a different set of features for the de-bunked news. However, as stated in Section 1.2, the most critical features that define the structure of the report are Claim, Claimant, Investigation, and Conclusion, which were present for each sample. The news samples were social media posts and news articles from different platforms such as Facebook, Twitter, Youtube, etc.

[1] https://www.altnews.in/

[2] https://www.boomlive.in/

[3] https://credibilitycoalition.org/credcatalog/project/digital-eye-india/

[4] https://www.factchecker.in/

[5] https://www.factcrescendo.com/

[6] https://factly.in/

[7] http://newsmobile.in/articles/category/nm-fact-checker/

[8] https://www.vishvasnews.com/fact-check-team/

[9] https://www.thequint.com/news/webqoof

| **Language** | **Total Samples** |
|---|---|
| English | 9058 |
| Hindi | 5155 |
| Tamil | 1314 |
| Malayalam | 1251 |
| Marathi | 1108 |
| Gujarati | 964 |
| Bengali | 956 |
| Telugu | 799 |
| Urdu | 596 |
| Punjabi | 428 |
| Odia | 287 |
| Sinhala | 256 |
| Assamese | 135 |
| Burmese | 128 |

Table 3.1: Number of Fact-Checked samples for different regional languages

### 3.1.1 Fake Modality

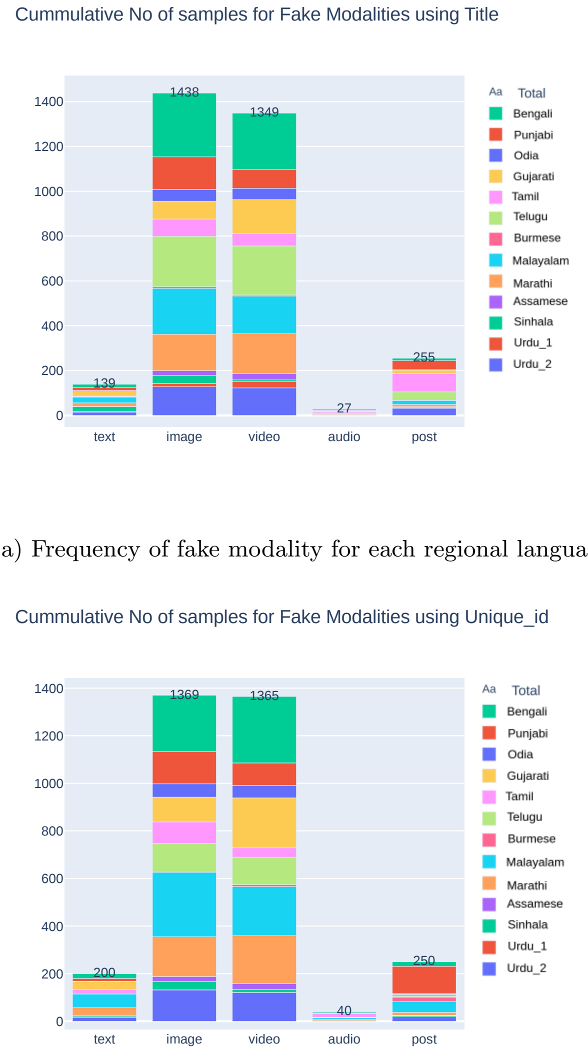
With the advent of social media and its position as the top-most news distributing medium, it becomes essential to understand fake news fabrication from a whole new perspective. Social media provides fast dissemination of information, and the multimedia content posted on the platform attracts the reader's attention. Information in the form of fabricated images, videos, text, or a combination of all is creating havoc. Most of the fake news datasets do not provide information, such as which elements of the news sample were fabricated. We utilized the title and unique-id of the de-bunked fake news, as it gave a reasonable estimate of the distribution of fabricated modality.

Figures 3.1a and 3.1b show that images and videos were the prime modalities, which the authors used to spread fake content. This analysis supports the previous studies, which indicate that text or even images alone cannot be used reliably enough to detect fake news. The fabricated modality distribution re-iterates the importance of multimodal fake news detection systems and the need to learn better feature representations.

### 3.1.2 Movement of Fake News

For propagation-based fake news detection systems, studying the movement of fake information across different social media platforms provides a good starting point for model formulation. These patterns help uncover platform and community dependent features and our understanding of community-wise platform preference for information acquisition and design more targeted features for classification.

We looked at the social media platform on which fake information for each regional language was spread. We looked at social media platforms' mentioned in the samples' claim and links provided



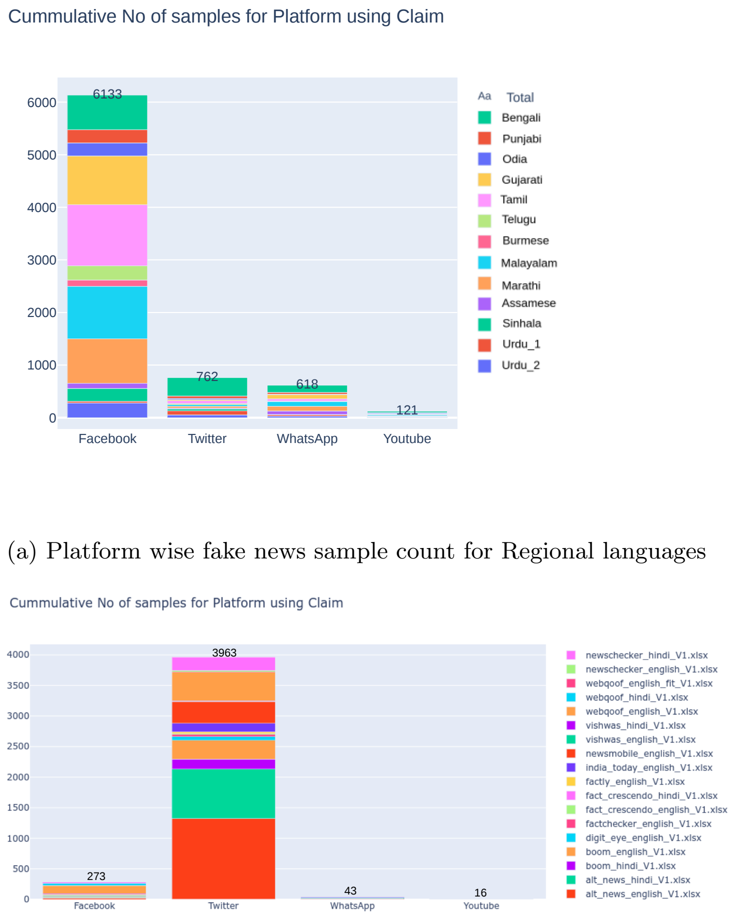
(a) Frequency of fake modality for each regional language using Title

(b) Frequency of fake modality for each regional language using *unique_id*

Figure 3.1: Frequency of fake modality for each regional language using (a) Title of the Fact-Checked article (b) the *unique_id* assigned by the website

for each de-bunked news article. The most common social media platform disseminating fake information was Facebook, for every regional language except for Sinhalese, where Twitter was the most common (figure 3.3a). An opposite trend was obtained in the analysis for English and Hindi samples where Twitter was the most common platform for fake information circulation (figure 3.3b). These negative results hint towards the difference in preference of social media platforms for different language communities.

On manual inspection, it was noted that in the claim, the first platform mentioned was where the fake news originated. In the first platform analysis to find where the bogus information originated, we used string matching. The results indicated that Fake News curators preferred Facebook for regional languages. The bogus information which originated on Facebook sometimes



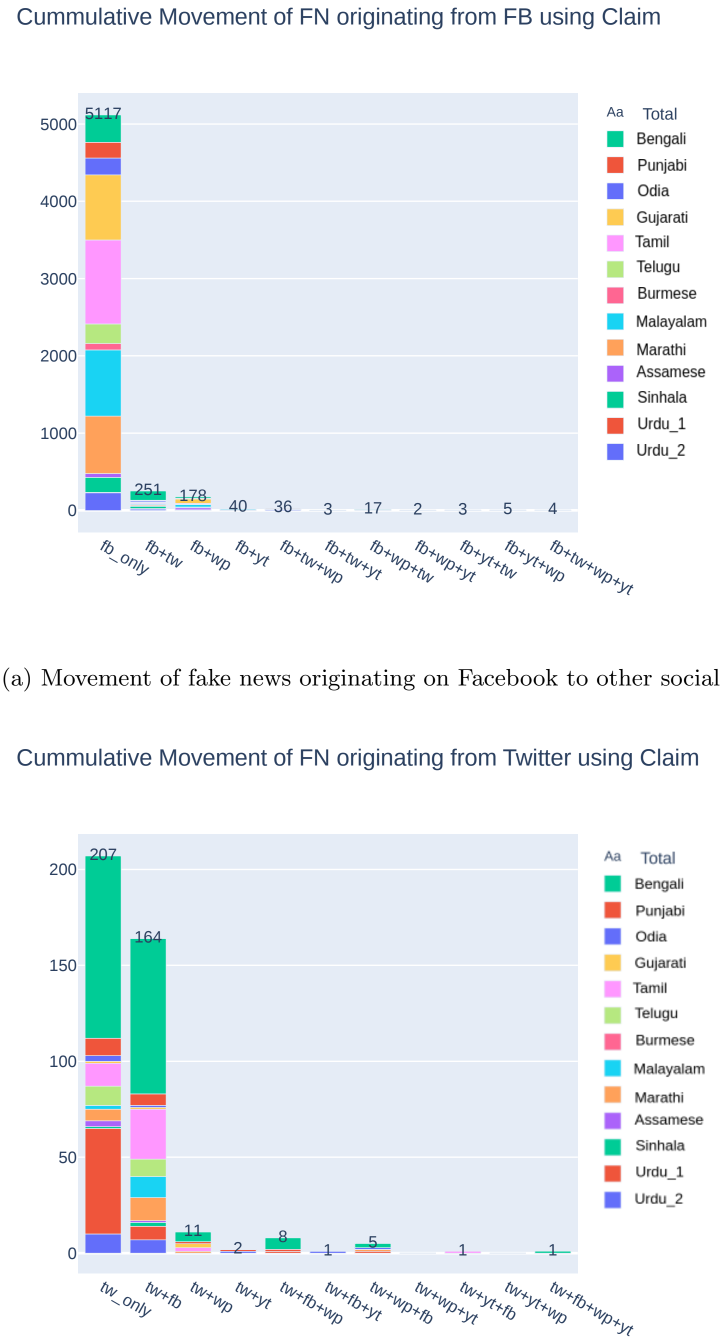
(a) Platform wise fake news sample count for Regional languages

(b) Platform wise fake news sample count for English and Hindi

Figure 3.2: Social Media platform distribution of (a) regional languages, (b) English and Hindi

appeared on Twitter as well, though with less probability. Figure 3.3 highlight the movement of fake news articles originating from Facebook and Twitter, respectively.

The analysis mentioned above gives a rough estimate of people's common platform to read fake news belonging to different language communities. This would help test the system in real-world scenarios and make a language-specific design to classify fake news better.



(a) Movement of fake news originating on Facebook to other social media platforms

(b) Movement of fake news originating on Twitter to other social media platforms

Figure 3.3: Movement of fake news originating on Facebook and Twitter to other social media platforms

### 3.1.3 Cross Lingual Dynamics of Fake News

In subection 3.1.1, it was established that images are one of the vital modalities that are used for the spread of false information. Therefore, it becomes necessary to analyze it further and find the correlation between fake content and different regional languages. For this part of the

analysis tried to answer one of the two research questions posed in [1]:
*Can image content transcend language silos? If so, how often does this happen, and amongst which languages?*

To answer the question mentioned above, we looked at a subset of the regional dataset. Tweet ids which contained false information were present in the dataset. We used a python library Twarc[10] to collect the fake samples along with their content. Out of the 3403 tweets, only 2480 of them had associated images. Therefore, we used this subset of multimodal tweets to perform further analysis.

In the first step, we clustered similar images using Facebook's PDQ algorithm.[11] The PDQ algorithm is a photo hashing method, which converts images into 256 bits with hamming distance, which was developed to solve the copy-detection problem. Internally, it uses Jarosz filters and Discrete Cosine Transforms along with different downsampling techniques to calculate a fixed 256 length vector and a quality measure for a given image. Further, the clustering of images is based on the hamming distance of the respective images based on a threshold that can be tuned.

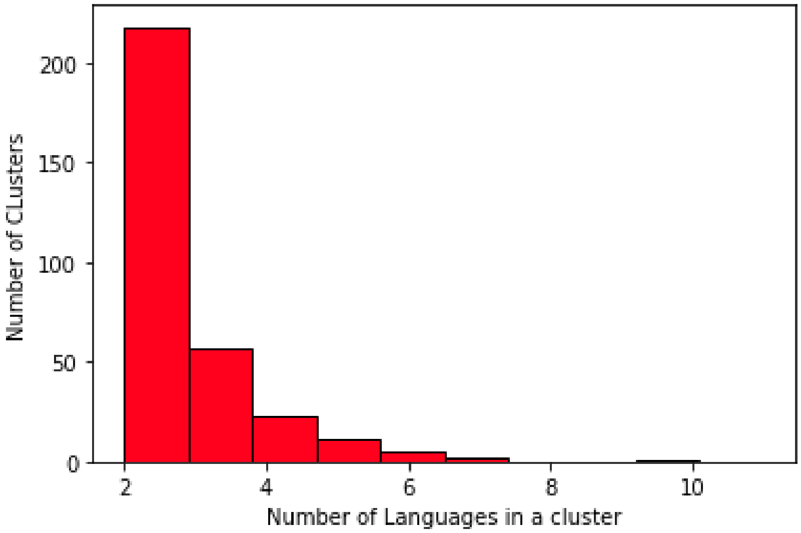


Figure 3.4: The number of different languages (the fact-check language associated with the image) present in a single image cluster

We used the default value of 23 to find clusters in the dataset. To understand whether similar images were used with posts of different regional languages to convey fake news, we calculated the number of different languages (the fact-check language associated with the image) present in a single image cluster. Figure 3.4 shows that some of the images could transcend language barriers, with a single image appearing in two different language communities 200+ times. To answer the second part of the question, we looked at the different language pairs inside an image cluster. The most common pairs of language we found that shared similar visual contents were Punjabi-Urdu and Marathi-Gujarati. The geographic location of the speakers and the

[10] https://github.com/DocNow/twarc
[11] https://github.com/facebook/ThreatExchange/tree/master/pdq

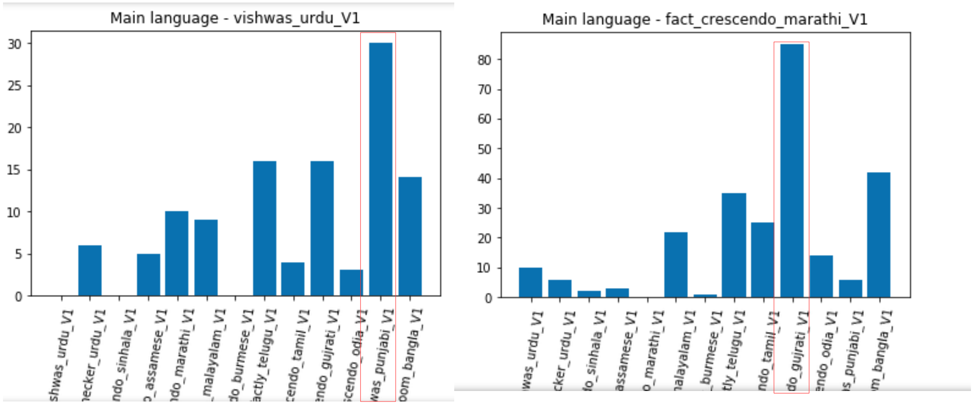


Figure 3.5: Example of diffusion of visual information from one language community to another. The most common pairs of language that shared similar visual contents were Punjabi-Urdu (Left) and Marathi-Gujarati (Right)

linguistical similarity of these languages could explain this cross-lingual diffusion [1].

# Chapter 4

# Proposed Fake News Detection Systems

We propose two novel Fake News Detection methods based on the extensive literature survey and social media dataset analysis. Both the methods surpass the state-of-the-art performance on the FakeNewsNet dataset [51].

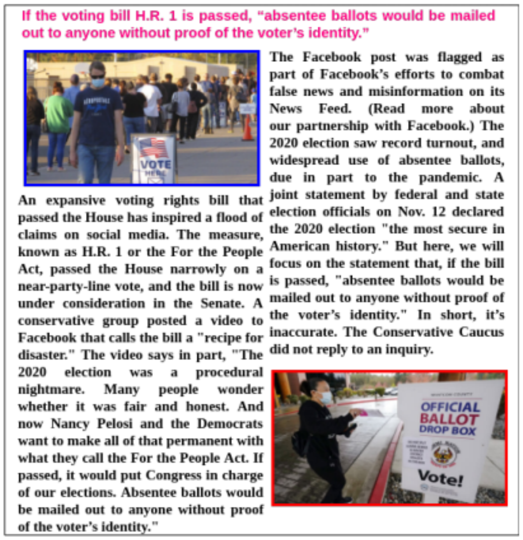
## 4.1 Leveraging Inter-modality Discordance for Multimodal Fake News Detection

### 4.1.1 Introduction

Existing studies have made significant strides towards multimodal fake news detection with less emphasis on exploring the discordance between the different multimedia present in a news article (Figure 4.1). We hypothesize that fabrication of either modality will lead to a dissonance between the modalities, and resulting in misrepresented, misinterpreted, and misleading news. Through this work, we inspect the authenticity of news from online media outlets by exploiting the relationship (discordance) between textual and multiple visual cues. We develop an inter-modality discordance-based fake news detection framework to achieve the goal. In this end-to-end multimodal framework, modal-specific discriminative features are incorporated by utilizing the cross-entropy loss, along with a modified version of contrastive loss that explores inter-modality discordance.[1]

**Motivation to add multiple images:**
We believe incorporating multiple images is beneficial due to following reasons: *(i)* understanding story in a text often requires reader to develop mental imagery skills [25,67]. Images can facilitate creation of such mental representations [9, 11] and can result in deeper learning [35, 36, 46, 47], *(ii)* images assist in clarification of ambiguous relations in the text, often termed as "multimedia

[1] This work was accepted at ACM Multimedia Asia 2022.


VOTE
HERE
OFFICIAL
BALLOT
DROP BOX
Vote!

If the voting bill H.R. 1 is passed, “absentee ballots would be mailed out to anyone without proof of the voter’s identity.”

An expansive voting rights bill that passed the House has inspired a flood of claims on social media. The measure, known as H.R. 1 or the For the People Act, passed the House narrowly on a near-party-line vote, and the bill is now under consideration in the Senate. A conservative group posted a video to Facebook that calls the bill a "recipe for disaster." The video says in part, "The 2020 election was a procedural nightmare. Many people wonder whether it was fair and honest. And now Nancy Pelosi and the Democrats want to make all of that permanent with what they call the For the People Act. If passed, it would put Congress in charge of our elections. Absentee ballots would be mailed out to anyone without proof of the voter’s identity."

The Facebook post was flagged as part of Facebook’s efforts to combat false news and misinformation on its News Feed. (Read more about our partnership with Facebook.) The 2020 election saw record turnout, and widespread use of absentee ballots, due in part to the pandemic. A joint statement by federal and state election officials on Nov. 12 declared the 2020 election "the most secure in American history." But here, we will focus on the statement that, if the bill is passed, "absentee ballots would be mailed out to anyone without proof of the voter’s identity." In short, it’s inaccurate. The Conservative Caucus did not reply to an inquiry.

Figure 4.1: A News article present on online news sharing platforms (part of our evaluation dataset). The text written in pink and black color depicts the headline ($H$) and the news content ($C$). Whereas the image highlighted in blue is represents the top-image ($I_1$) (first-image) that accompanies the text. In addition, the image highlighted in red represents other-images ($I_2, ..., I_k$) present within the given news sample.

effect” [36], *(iii)* while words can be viewed as descriptive representations, images, in contrast, are depictive external representations, showcasing the meaning that the text represents [3].

### 4.1.2 Problem Statement

In a given dataset, we have a set of $n$ news articles, $N{=}\{H_i, C_i, M_i, y_i\}_{i=1}^{n}$. Each news sample $N_i$ consists of four elements, *i.e.* headline ($H_i$), content (text body) ($C_i$), image-set ($M_i$) and the corresponding label ($y_i$). In this report, we encapsulate the cross-modal synergies to investigate multi-modal fake news detection as a joint-learning problem with the main task of *Binary Classification* where $N_i$ can be categorized as either fake *(y=1)* or real *(y=0)*, along with an auxiliary task of modeling *Inter-Modality Discordance*. We hypothesize that fabrication introduced in any will lead to a dissonance between them, *i.e.* the news feature vectors from a fake (real) sample, when projected onto a common multimodal space, will be distant (closer) and portray the negligible (significant) relationship between modalities [6] and in the process learn rich shared representations to assist the main classification task.

### 4.1.3 Methodology

This section introduces our proposed approach along with a descriptive model diagram (Figure 4.2). The architecture comprises of four main components *(i)* Unimodal Multiple Visual Feature Extractor, *(ii)* Unimodal Textual Feature Extractor, *(iii)* Inter-modality discordance score Module, and *(iv)* (Multi+Uni)modal Fake News Detector.

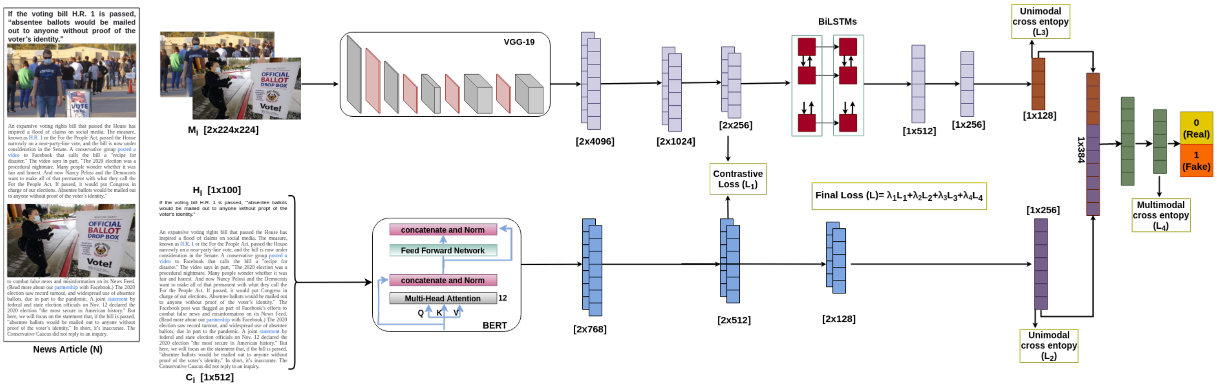


Figure 4.2: Illustration of the proposed model. It comprises of a primary task i.e. multimodal fake news detction. We introduce three auxiliary learning tasks i.e. measuring inter-modality discordance score via modified contrastive loss, multiple visual feature extractor and, textual feature extractor.

### Unimodal Multiple Visual Feature Extractor

The goal of the module is to obtain feature representations for a sequence of images. Taking inspiration from Giachanou et al. [13], we use a Convolution-LSTM based system that extracts sequential information from multiple images in a two-fold manner and can even handle variable number of images present in any news article. Let $M_i$=$\{I_1, I_2, ..., I_k\}$ represent a set of images present in the news article, where $M_i$ denotes the image-set of the $N_i$ news sample, and $k$ denotes the corresponding count of images present in it. The LSTM layer of the module handles the variable number of images for each sample. Although to fully utilize the power of vectorization, we perform dynamic batch-wise zero-padding to the maximum count of images present in a given batch. The pre-processed images are first passed through a VGG-19 network pre-trained on the ImageNet database. The second to last layer of the VGG-19 [52] network serves as a feature embedding that captures the semantic meaning for each image present in the news article. Next, to sufficiently capture the sequential nature of these images, i.e., the order in which they are present in a news article, the CNN-based features are routed through a Bidirectional Long-Short Term Memory (BiLSTM) [19,49] Layer. The hidden state of the last BiLSTM cell is then passed through a full-connected layer to obtain a fixed-length feature vector for a sequential image-set of variable length.

### Unimodal Textual Feature Extractor

This module extracts contextual representations from a news sample's headline and content (text body). Earlier textual encoders such as GloVe [38] and Word2Vec did not consider the context surrounding a token. For example, words such as "bank" can have multiple word senses, one referring to a financial institution and another to a riverside. Word embeddings such as BERT [10], GPT [42], and ELMo [39] were introduced to overcome these shortcomings. In our work, for each news article ($N_i$), the content ($i$) and Headline ($H_i$) is segregated into tokens using the WordPiece algorithm [48] to form respectively. These tokens, along with some special tokens ([CLS]), are then combined with positional encodings to generate the final input to a pre-trained BERT (Bidirectional Encoder Representations from Transformers). It is pre-trained on two language modeling tasks: (i) masked language modeling (MLM) and (ii) next sentence prediction task. BERT generates a 768-dimensional embedding vector for each token in the input sequence. We use the embedding vector corresponding to the [CLS] token as it provides a good sentence-level representation [10].

### (Multi+Uni)Modal Fake News Prediction

This module performs our primary task of Multimodal Fake News detection and two more supervised tasks to extract model-specific features from the news article. It leverages information from the textual and multiple visual entities of a news sample to form a multimodal feature vector. To form the multimodal feature vector, we need to perform multimodal fusion. Existing fusion strategies can be categorized into (i) early fusion, which merges unimodal features to form joint representation before attempting to classify, (ii) late fusion, where individual modalities pass through more targeted unimodal processing pipelines to classify the content for each modality independently. For our purpose, we use hybrid fusion, a strategy that lies midway between the two extremes (i.e., early and late fusion). We opted for such a fusion technique as early fusion helps capture the cross-correlations between the data features, whereas late fusion helps develop more robust unimodal features. Thereby the combination improves the overall model performance.

### Inter-modality Discordance Score Module

Inter-modality Discordance Score Module performs our proposed sub-task. It captures the relationship (discordance) between various components present in a news article for assisting multimodal fake news detection. The core idea behind this module is that the average distance (similarity metric) amongst the different components of a fake news article is greater than the for a true news article in a multimodal space. We posed this learning problem as a semi-supervised one with the help of a modified contrastive loss. A recent study by Claire Wardle, First Draft News Research Director,[2] divided misinformation and disinformation into seven types: (i) satire or parody, (ii) false connection, (iii) misleading content, (iv) false context, (v) imposter content, (vi) manipulated content, and (vii) fabricated content. Using this specialized module, catching fake stories where the headline and visual cues do not support the content (type: ii), and genuine content is shared with false visual-contextual information (type: iv) becomes more

[2] https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79

straightforward.

```
Algorithm 1: Inter-modality Discordance Loss (Training Phase)
    Input: ${M, y, H_i, C_i, I_{i_1}, I_{i_2}, ..., I_{i_k}}$
    Output: Loss
$P_i = [H_i, C_i, I_{i_1}, I_{i_2}, ..., I_{i_k}]$
for each $P_i$ do
    $r_c = \frac{1}{2+k} \sum P_i$;
    $distance = \frac{1}{2+k} \sum \|r_{P_i}, r_c\|$;
    if $y == 1$ then
        Loss = max(0, M - distance);
    end
    else
        Loss = distance;
    end
end
```

The detailed outline to calculate the inter-modality score is summarized in Algorithm 1, where inputs $(H_i, C_i, I_{i_1}, I_{i_2}, ..., I_{i_k})$ depict the intermediate normalized feature representations for the headline, content, and image-set. $r_c$ denotes the centroid value, and the distance signifies the average distance between the components of a news sample from the centroid. The distance metric chosen for measuring similarity is the euclidean distance (*L2-norm*). $M$ indicates margin value which acts as a threshold. When the average distance amongst the different components of a fake news article is distant enough, there is no penalization. However, when that distance is less than M, the loss will represent a positive value, and the model parameters will be updated to produce more distant feature vectors.

### Model Integration and Joint Learning

Our proposed architecture is trained using three fully-supervised-learning-based cross-entropy loss functions for the extracted textual, visual, and multimodal features(Section 4.1.3). Along with this, the auxiliary task of modeling inter-modality discordance is performed using a modified version of contrastive loss, which is explained in Section 4.1.3. It is to be noted that all four tasks are performed during model training, but the primary task is considered when assessing the model's performance. The four-loss functions considered during training are as follows:

$$L_1 = \begin{cases} \frac{1}{n} \sum_{i=0}^{m} d(r_m, r_c), & \text{if real sample} \\ max(0, m - \frac{1}{n} \sum_{i=0}^{m} d(r_m, r_c)), & \text{otherwise} \end{cases} \tag{4.1}$$

$$L_2 = \frac{1}{n} \sum_{i=1}^{n} y_T^i \log \hat{y}_T^i + (1 - y_T^i) \log(1 - \hat{y}_T^i) \tag{4.2}$$

$$L_3 = \frac{1}{n}\sum_{i=1}^{n} y_I^i \log \hat{y}_I^i + (1 - y_I^i)\log(1 - \hat{y}_I^i) \tag{4.3}$$

$$L_4 = \frac{1}{n}\sum_{i=1}^{n} y_C^i \log \hat{y}_C^i + (1 - y_C^i)\log(1 - \hat{y}_C^i) \tag{4.4}$$

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

$$L = \alpha L_1 + \beta L_2 + \gamma L_3 + \delta L_4$$

where $\alpha, \beta, \gamma, \delta = 1$ in our design.

### 4.1.4 Experiments and Results

In this section, we present a series of experiments to demonstrate the efficacy of our proposed method. Specifically, we aim to answer: the following evaluation questions:

- **RQ1:** Is the proposed model able to improve multimodal fake news detection by incorporating multiple images and modality discordance.
- **RQ2:** How effective modality discordance hypothesis is in multimodal fake news detection.

#### Dataset

To evaluate the performance of our proposed architectures, we use two standard public benchmark multimodal fake news datasets- the FakeNewsNet Dataset [51]. This repository contain two sub-datasets collected from Politifact[3] and Gossipcop.[4] Politifact is a US-based fact-checking website that debunks statements regarding politics and recently Covid-19 as well. GossipCop fact-checks information related to entertainment published in various magazines.

FakeNewsNet Repository (clean version): Giachanou et al. [13] performed multimodal fake news detection by using a clean version of the dataset released by Shu et al. [51]. They performed dataset cleaning in which all the news samples with non-news content images and duplicates are removed by manual intervention for the GossipCop dataset. In our study, for a fair comparison with the state-of-the-art, we used the cleaned version provided by the authors [13] that consists of 2,745 fake and 2,714 real samples having at least one image associated with them.

The final dataset statistics are provided in Table 4.1

[3] https://www.politifact.com/

[4] https://www.gossipcop.com/

|  |  | # News Articles | # of images |
|---|---|---|---|
| Politifact (raw) | Real | 624 (399) | 5607(5027) |
|  | Fake | 432 (346) | 5423(4462) |
|  | Overall | 1056 (745) | 11030(9489) |
| GossipCop (raw) | Real | 16,817 (10970) | 405367(381117) |
|  | Fake | 5,323 (4223) | 137717(125361) |
|  | Overall | 22,140 (15193) | 543084(506478) |
| GossipCop (clean) | Real | 2,714 (952) | 13567(1718) |
|  | Fake | 2,745 (2526) | 44306(4364) |
|  | Overall | 5,459 (3478) | 57873(6082) |

Table 4.1: The dataset statistics used during the experiments. Values in (.) signifies the final count of samples used during experimentation.

### Baselines

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

- SpotFake+ [53]: The algorithm leverages the XLNet language model to extract contextual text information [63]. The image features are learned from a pre-trained VGG-19 [52] network. The features obtained from both modalities are fused in an additive manner to build the desired news representation.

- SAME [7]: The system exploits the sentiments hidden within the user comments to detect fake news. The method integrates information from the text, images, and user profile given for a particular sample. An adversarial framework is added to preserve semantic relevance and consistency across different representations.

- SAFE [65]: This model design aims to capture the similarity among modalities to exploit the multimodal information and extract better representations for misinformation detection. They use cosine similarity for this task. The text features are extracted by passing the initial representations through TextCNN [24]. For visual features, images are first passed through an image2sentence model and then through a TextCNN module.

- Multi-image Multimodal Method : This is the first research that explores multiple images in tandem with text to perform fake news detection. To extract visual features from multiple images, they utilize a pre-trained VGG-19 network to obtain top-10 tags corresponding to each image along with the intermediate embeddings after passing through an LSTM layer. Authors also exploit similarity information, i.e., text-image similarity, by calculating cosine similarity between them the modalities.

### Experimental Setup

#### Dataset Split and Evaluation Metric:

We randomly split the dataset into *80%* training and *20%* testing. Table 4.2 depicts the distri-

|  | Politifact (raw) | GossipCop (raw) | GossipCop (clean) |
|---|---|---|---|
| Train | 316 (271) | 8752(3367) | 766 (2032) |
| Test | 83 (75) | 2218(856) | 186 (494) |
| Overall | 399 (346) | 10970(4223) | 952 (2526) |

Table 4.2: The train-test split statistics of the datasets used during the experiments. Values in (.) signifies the count of fake samples.

bution of fake and real samples in each split.

To evaluate the performance of our proposed model, we use Accuracy, Precision, Recall, and F1 score, which are considered standard evaluation metrics while performing classification experiments.

**Implementation Details and Hyper-parameters:**
We implemented our proposed methodology via Pytorch. All the experiments are conducted on Nvidia GeForce RTX 2080Ti and 3090 GPUs.

All hyper-parameters are carefully tuned in the validation set with the help of an early stopping strategy. The complete list of hyper-parameter values is made. public[5]

### Multimodal Fake News Detection RQ1

|  |  | **LIWC**$^\dagger$ | **VGG-19**$^\mp$ | **Att-RNN**$^\ddagger$ | **SAFE**$^\ddagger$ | **Proposed Method** |
|---|---|---|---|---|---|---|
| **Politifact (raw)** | Acc. | 0.822 | 0.649 | 0.769 | 0.874 | **0.913** |
|  | F1 | 0.815 | 0.720 | 0.826 | 0.896 | **0.902** |
| **GossipCop (raw)** | Acc. | 0.836 | 0.775 | 0.743 | 0.838 | **0.850** |
|  | F1 | 0.466 | 0.862 | 0.846 | **0.895** | 0.743 |

Table 4.3: Comparison of our proposed model with the text$^\dagger$, image$^\mp$ and single-image multimodal$^\ddagger$ fake news baselines.

|  |  | Giachanou *et al.* [13] | Proposed Method |
|---|---|---|---|
| **GossipCop (clean)** | Acc. |  | **0.880** |
|  | F1 | 0.795 | **0.915** |

Table 4.4: Comparison of our proposed model with the multi-image multimodal fake news detection baselines.

We compare our proposed method with the existing state-of-the-art models described in Section 4.1.4. A comparative table depicting the results of our proposed methodology and the baselines for all the evaluation metrics are shown in Table 4.3 and 4.4. From the table, we make the following observations:

- Our proposed model beats the current state-of-the-art for multi-image fake news detection [13] by a relatively large margin of **12%** on the F1-score.

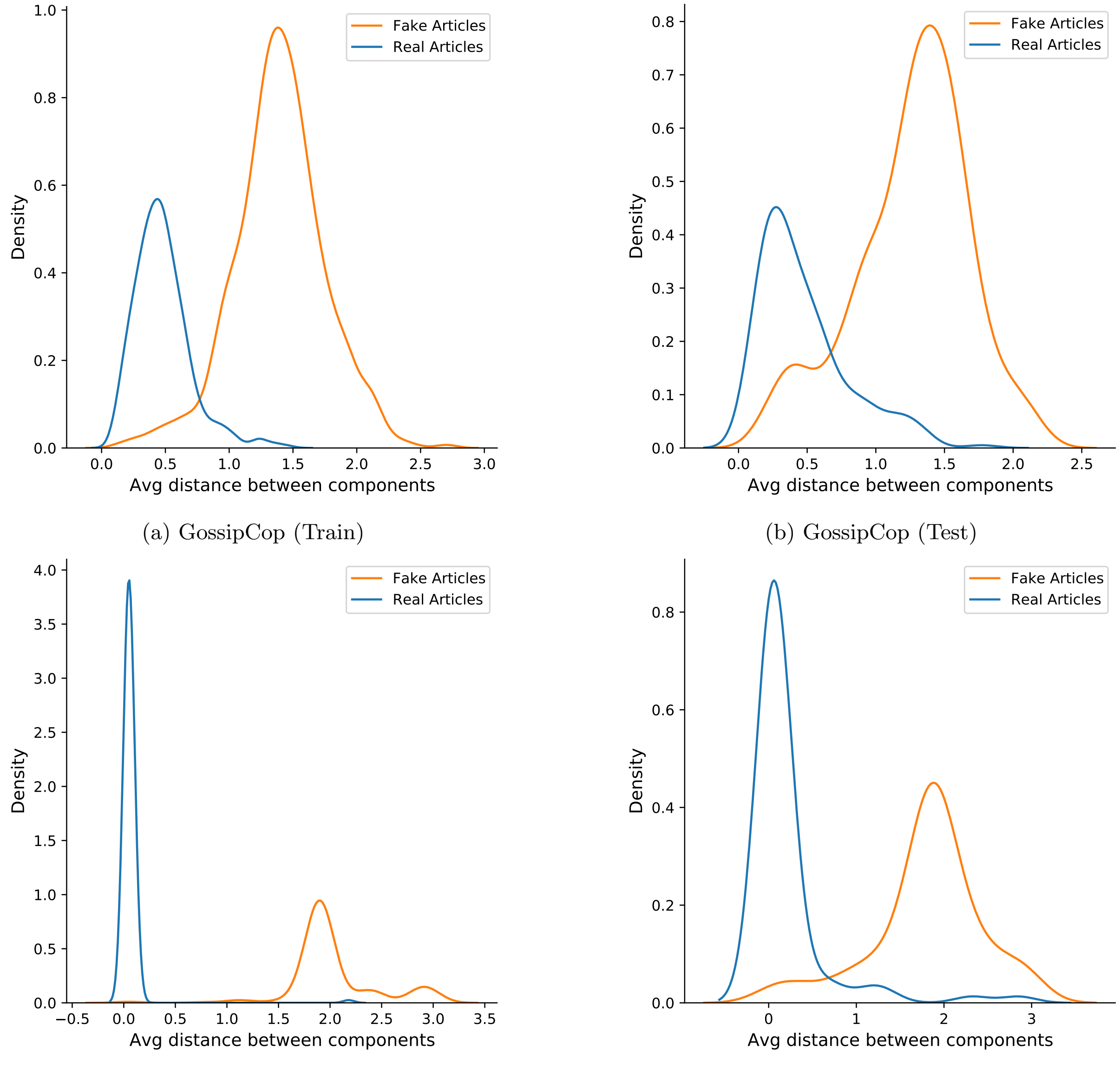
[5] https://github.com/mudit-dhawan/FND

- Additionally, our model outperforms SAFE [65], the current state-of-the-art for single-image fake news detection method, on the Politifact dataset on all the evaluation metrics by **3.6%** on accuracy and **2.2%** on Recall and F1-score.
- **Investigating the inconsistent performance of the proposed model on Gossipcop (raw):**
  - Previous studies [13, 54] have pointed out the presence of non-news images (i.e., logo, gifs, icons of the news websites and advertisements) within the news articles for the Gossipcop (raw) dataset. We hypothesize that such noisy images lead to a decrease in the performance of our proposed method. To investigate the case, we first performed an intersection on the Gossipcop (raw) and Gossipcop (clean) datasets to get the resultant set comprising of non-news images. We observe that, on average, *77.77%* of the images are non-news in the raw dataset for a sample.
  - We also found that, on average, for a sample, there is a *80%* probability that at least two out of three images passed to the model will be noisy and *97%* probability for at least one out of three images to be noisy. In conclusion, our proposed model is designed to capture the discordance between the modalities. Since there exists no relationship between the incoming noisy images and text, our model fails to learn any representative pattern about the news article leading to inconsistent results.

### Evaluating Inter-Modality Discordance (RQ2)

In order to answer the second research question, we visualize the average distance between the components from the centroid $r_c$ of a news sample via Kernel Density Estimate (KDE) plots. Figure 3 shows the distribution of discordance scores for the training and testing splits for Gossipcop (clean) and Politifact, respectively. We hypothesize that the average distance between the components of a real (fake) news sample will be small (significant). The mean distance between real and fake news components is 0.476 and 1.39, and 0.06 and 2.03, respectively, for GossipCop (clean) and Politifact datasets, as shown in figure 4.3. We observe that the distributions in our density plots are narrow, which shows that the variance in the output of a class is low. Additionally, the intersection of the area under the curve for real and fake news is minimal, indicating a significant separation between the classes. All the observations mentioned above is consistent across datasets and training-validation splits, as depicted in figure 4.3.

In another attempt to understand the importance of inter-modality discordance loss, we performed an ablation study to examine the performance of each component of our model. The results are depicted in Table 4.5. Here $L2$ and $L3$ signify the performance of the proposed model when using only text and visual uni-modal loss, respectively. On the other hand, $L2 + L3 + L4$ represents the multimodal framework without the inter-modality discordance loss. The general performance of the different variants on both datasets is: $ProposedMethod > L2 + L3 + L4 > L2 > L3$. From Table 4.5, we observe an improvement in the $L2 + L3 + L4$ variant on the



(a) GossipCop (Train)

(b) GossipCop (Test)

(c) Politifact (Train)

(d) Politifact (Test)

Figure 4.3: Measuring modality discordance score on train and test set of GossipCop (clean) and Politifact.

addition of L1, by increasing the model's generalizability. Thus, showing that the addition of inter-modality discordance loss in the multimodal detection method aids in better fake news detection.

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

Table 4.5: Comparison of the proposed model with its different variants.

### 4.1.5 Conclusion

In this section, we present an inter-modality discordance based fake news detection system. The system leverages information from the all the multimodal components present in a news article, and investigates the relationship between them via a modified version of contrastive loss. In addition, cross-entropy loss enforces the model to learn unimodal and multimodal discriminative features both independently and jointly. Extensive experiments on two real-world datasets demonstrate the strong performance of our proposed method.

## 4.2 Cross-Attention Based Model for Explainable Fake News Detection

Multi-modal Fake News Detection systems focus on extracting meaningful shared representations of data. Most previous studies incorporate them separately and then use concatenation with a single fully-connected layer to find multi-modal representation. However, this involves very little interaction between the two modalities. The SAFE [65] paper tried to address similar concerns. They used an explicit similarity measuring function (Cosine similarity) to quantify this relationship.

Our proposed model uses a cross-attention module to design more robust multi-modal features fused to perform better classification. The attention maps generated in the process could be utilized to explain the prediction. Attention weights define the importance of each image pixel and text token to assist in the classification.

### 4.2.1 Methodology

This section introduces the Cross-Attention based model for fake news detection. The model consists of four main components figure 4.4.

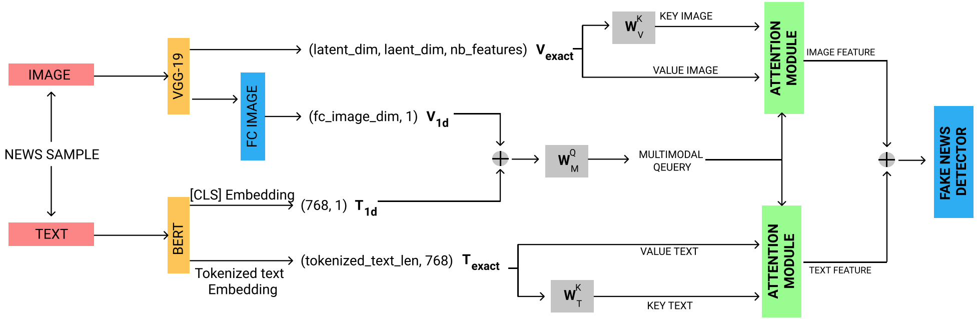


Figure 4.4: Model architecture of the proposed Cross Attention model.

### (1) Visual Encoder:

We employ Pre-trained Deep Convolutional Neural Networks to extract latent representation for the visual modality. These are used extensively as visual feature extractors in many transfer learning tasks. We adopted a pre-trained VGG-19 architecture [52] on the ImageNet dataset. This visual encoder has been used in other multimodal classifiers [23, 53, 54]. After passing the image through the encoder, we get two types of outputs: $V_{\mathrm{1d}}$ $(shape = [latent\_size, latent\_size,$ $nb\_channels])$ and $\mathrm{V}_{\mathrm{exact}}$: $(shape = [fc\_img\_dim, 1])$ list of Images through the visual encoder.

### (2) Text Encoder :

There has been in-depth research surrounding explicit (hand-crafted features based on text statistics) and latent representations learned through deep-learning models for Fake News detection models. Previous methods have incorporated Text-CNN [24] , Bi-LSTM network, and Transformer [56] based encoders such as BERT [10] and Transformer-XL [8] as textual feature extractors. Based on the current success of pre-trained Transformer based models on downstream NLP tasks, we use BERT-Base with 12 encoder layers as the text encoder. We add the special [CLS] token at the start of the text and use the output state corresponding to it as the feature representation of the news article's textual information. Along with this, we extract the output states corresponding to the rest of the input news tokens. The text information transforms into two vectors: $T_{\mathrm{1d}}(shape = [768, 1])$ dimensional vector and $T_{\mathrm{exact}}$ $(shape = [tokenized\_text\_len, 768])$ vector through the encoder.

### (3) Cross-Attention module:

Following the same notation as the authors of [56] we use key, query, and value to define the vectors' set involved in the attention mechanism. We used the concatenation of the 1-dimensional vectors from both the encoders to create a query vector and keys corresponding to news' image and text from the exact vectors. Where $\mathbf{Q}_M$ is the Multimodal Qeuery of the news sample, $\mathbf{K}_V$ and $\mathbf{K}_T$ are the keys for visual and text content. The matrices $\mathbf{W}_M^Q$, $\mathbf{W}_V^K$ and $\mathbf{W}_T^K$ are learned during the training phase of the model

$$\mathbf{Q}_M = \mathbf{W}_M^Q * Concat(\mathbf{V}_{1d}, \mathbf{T}_{1d}) \tag{4.5}$$

|  | **PolitiFact** | **Gossipcop** |
|---|---|---|
| Real | 321 (624) | 10259 (16817) |
| Fake | 164 (432) | 2581 (5323) |

Table 4.6: Numer of samples after pre-procesiing in Politifact and Gossipcop respectively. The values in the brackets indicate the total number of samples before any pre-processing.

$$\mathbf{K}_V = \mathbf{W}_V^K * \mathbf{V}_{exact} \tag{4.6}$$

$$\mathbf{K}_T = \mathbf{W}_T^K * \mathbf{T}_{exact} \tag{4.7}$$

We then perform Scaled Dot-Product Attention for the calculated image key and text keys separately using the multi-modal query, along with $\mathrm{V_{exact}}$ and $\mathrm{T_{exact}}$ as values, respectively.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}})\mathbf{V} \tag{4.8}$$

**Fake News Detector :**
The weighted sum output of both the modalities' values is then concatenated and passed through a shallow one-layer feed-forward network to perform classification. The entire model is jointly trained using cross-entropy loss.

### 4.2.2 Dataset

To evaluate the performance of our proposed architectures, we use two standard public benchmark multimodal fake news dataset, the FakeNewsNet Dataset [51]. This repository contain two sub-datasets collected from Politifact[6] and Gossipcop. [7] Politifact is a US based fact-checking website that debunks statements regarding politics and recently Covid-19 as well. GossipCop fact-checks information related to entertainment published in various magazines.

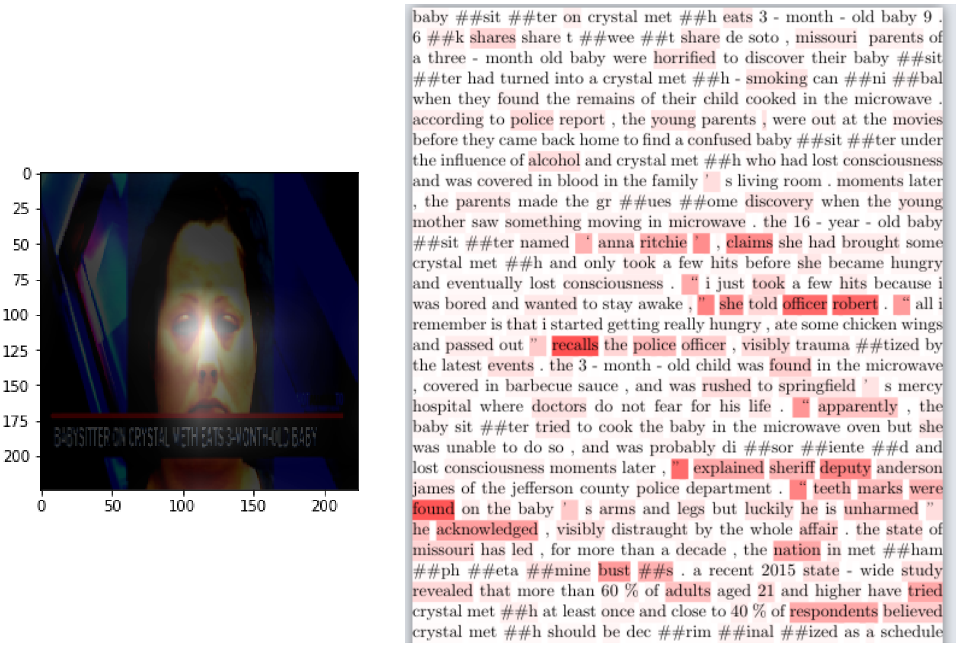
We used the pre-processed version presented in [53]. This data was manually cleaned by removing non-useful images such as logos from the the samples and dropped the samples which lacked images or contained GIFs. See Table 4.6 for the exact distribution.

### 4.2.3 Results

We compare the results of the proposed models with baselines ranging from text-only, visual-only and multimodal frameworks. Table 4.7 shows the performance of selected model on the FakeNewsNet Dataset, the proposed models achieve comparable results without any hyperparameter-

[6] https://www.politifact.com/

[7] https://www.gossipcop.com/



baby ##sit ##ter on crystal met ##h eats 3 - month - old baby 9 . 6 ##k shares share t ##wee ##t share de soto , missouri parents of a three - month old baby were horrified to discover their baby ##sit ##ter had turned into a crystal met ##h - smoking can ##ni ##bal when they found the remains of their child cooked in the microwave . according to police report , the young parents , were out at the movies before they came back home to find a confused baby ##sit ##ter under the influence of alcohol and crystal met ##h who had lost consciousness and was covered in blood in the family ' s living room . moments later , the parents made the gr ##ues ##ome discovery when the young mother saw something moving in microwave . the 16 - year - old baby ##sit ##ter named ' anna ritchie ' , claims she had brought some crystal met ##h and only took a few hits before she became hungry and eventually lost consciousness . " i just took a few hits because i was bored and wanted to stay awake , " she told officer robert . " all i remember is that i started getting really hungry , ate some chicken wings and passed out " recalls the police officer , visibly trauma ##tized by the latest events . the 3 - month - old child was found in the microwave , covered in barbecue sauce , and was rushed to springfield ' s mercy hospital where doctors do not fear for his life . " apparently , the baby sit ##ter tried to cook the baby in the microwave oven but she was unable to do so , and was probably di ##sor ##iente ##d and lost consciousness moments later , " explained sheriff deputy anderson james of the jefferson county police department . " teeth marks were found on the baby ' s arms and legs but luckily he is unharmed " he acknowledged , visibly distraught by the whole affair . the state of missouri has led , for more than a decade , the nation in met ##ham ##ph ##eta ##mine bust ##s . a recent 2015 state - wide study revealed that more than 60 % of adults aged 21 and higher have tried crystal met ##h at least once and close to 40 % of respondents believed crystal met ##h should be dec ##rim ##inal ##ized as a schedule

Figure 4.5: A visualization of generated Attention maps for a false news article with Title: *"Babysitter on crystal meth eats 3-month-old baby."*. The image (left) contains a bright spot on the face present in the picture, indicating higher attention weight. For the text present on the right, the attention weight is denoted by the intensity of the red colour [61]. It gives more attention to named-entities (officer's name, suspect's name, hospital name), and other words like unharmed.

tuning.

| **Modality** | **Model** | **PolitiFact** | **Gossipcop** |
|---|---|---|---|
| **Text** | XLNet+FC [53] | 0.74 | 0.836 |
| | LIWC [65] | 0.822 | 0.836 |
| **Image** | VGG-19 [65] | 0.649 | 0.775 |
| **Multimodal** | MVAE [23] | 0.673 | 0.775 |
| | SpotFake [54] | 0.721 | 0.807 |
| | SAFE [65] | *0.874* | 0.838 |
| | SpotFake+ [53] | 0.846 | 0.856 |
| | EANN [58] | 0.74 | 0.86 |
| | *Cross-Attention Network* | **0.8558** | **0.869** |

Table 4.7: Accuracy comparison of different Fake News Detection methods on FakeNewsNet Dataset.

### 4.2.4 Conclusion

In this section, we present an explainable multimodal fake news detection system. The system takes into account information from both the textual and visual components present in a news article, and leverages cross-attention module to calculate the most discrimnative words in text and patches in images. In the age of deep learning, and big black box models for high impact problems such as fake news detection, explainable predictions is a crucial step towards understanding the inner working of these models. Extensive experiments on two real-world datasets demonstrate the strong performance of our proposed method.

# Chapter 5

# Future Works

- Building upon the recent works of the multi-task learning paradigm in fake news detection, combine the different tasks and study the performance gain using each. Use these non-primary tasks to augment the small fake news datasets to be able to train deep learning models fully without the fear of overfitting.

- In multimodal fake news detection models, the visual modality features are extracted using non-specialized encoders primarily trained on a classification task with a limited number of classes. The images in these large training sets are also such that there is one primary class and not representative of the diversity and complexity of the images encountered by the model in a fake news detection task.

- Recently, researchers have looked into machine-generated news articles; one possible direction could be using the same adversarial training method to generate authentic news articles conditioned on fake news articles and vice versa to augment the existing datasets and add regularization.

- Content-based fake news detection systems try to learn patterns in the writing style and topics present in the news articles to help in early-stage misinformation classification where the factual information about the event is still missing. Going beyond the early detection framework or for scenarios with the correct factually information available, incorporating such a database into the pipeline would be crucial in combating this problem. Therefore, it would be necessary to explore methods that make use of credible sources to check facts or data for fake news detection (REALM [17] or KG approach).

- Add other popular fake news detection systems to the public repository, which currently contains MVAE and SpotFake, to help advance the fake news detection community.

# Chapter 6

# Limitations

In this chapter, we discuss the limitations and challenges encountered during the course of the thesis.

- To segregate fact-checked articles written in Indian regional languages other than Hindi, we utilized Google Translate[1] to first convert text to English and then manually find the boundary key words. This approach was limited by the efficacy of the translation tool used.

- To find fake-modality and first platform for the fact-checked articles we employed a keyword based pipeline, which could sometimes fail if a specific keyword was absent or the ordering of words changed, leading to small errors in the analysis.

- To clean the multiple images dataset that we had collected over the FakeNewsNet repository, we resorted to manual checking for repeating icons, and other placeholders, along with previous thresholds which caused degradation of the quality of the news sample, because of a loose pre-processing pipeline.

- To prove the efficacy of the explanations provided by our proposed cross-attention multimodal fake news detection system, we resort to manual inspection of few random samples and couldn't quantitatively prove its effectiveness using some pre-defined metric.

[1] https://translate.google.co.in/

# Bibliography

[1] Agarwal, P., Garimella, K., Joglekar, S., Sastry, N., and Tyson, G. Characterising user content on a multi-lingual social network. In *Proceedings of the International AAAI Conference on Web and Social Media* (2020), vol. 14, pp. 2–11.

[2] Allcott, H., and Gentzkow, M. Social media and fake news in the 2016 election. *Journal of economic perspectives 31*, 2 (2017), 211–36.

[3] Bartholomé, T., and Bromme, R. Coherence formation when learning from text and pictures: What kind of support for whom? *Journal of Educational Psychology 101*, 2 (2009), 282.

[4] Boehm, L. E. The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin 20*, 3 (1994), 285–293.

[5] Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.-T., Boato, G., Riegler, M., Kompatsiaris, Y., et al. Verifying multimedia use at mediaeval 2015. *MediaEval 3*, 3 (2015), 7.

[6] Chugh, K., Gupta, P., Dhall, A., and Subramanian, R. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 439–447.

[7] Cui, L., Wang, S., and Lee, D. Same: sentiment-aware multi-modal embedding for detecting fake news. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining* (2019), pp. 41–48.

[8] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860* (2019).

[9] De Koning, B. B., and van der Schoot, M. Becoming part of the story! refueling the interest in visualization strategies for reading comprehension. *Educational Psychology Review 25*, 2 (2013), 261–287.

[10] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Eitel, A., and Scheiter, K. Picture or text first? explaining sequence effects when learning with pictures and text. *Educational psychology review 27*, 1 (2015), 153–180.

[12] Feng, S., Banerjee, R., and Choi, Y. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2012), pp. 171–175.

[13] Giachanou, A., Zhang, G., and Rosso, P. Multimodal multi-image fake news detection. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (2020), IEEE, pp. 647–654.

[14] Gunning, D. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web 2*, 2 (2017).

[15] Gupta, A., Kumaraguru, P., Castillo, C., and Meier, P. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics* (2014), Springer, pp. 228–243.

[16] Gupta, A., Lamba, H., and Kumaraguru, P. $1.00 perrt\#bostonmarathon\#prayforboston : Analyzingfakecontentontwitter.In 2013 APWGeCrimeresear$
$-12.$

[17] Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909* (2020).

[18] Han, Y., Karunasekera, S., and Leckie, C. Graph neural networks with continual learning for fake news detection from social media. *arXiv preprint arXiv:2007.03316* (2020).

[19] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[20] Huh, M., Liu, A., Owens, A., and Efros, A. A. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 101–117.

[21] Jamieson, K. H., and Cappella, J. N. *Echo chamber: Rush Limbaugh and the conservative media establishment.* Oxford University Press, 2008.

[22] Jin, Z., Cao, J., Zhang, Y., Zhou, J., and Tian, Q. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia 19*, 3 (2016), 598–608.

[23] Khattar, D., Goud, J. S., Gupta, M., and Varma, V. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference* (2019), pp. 2915–2921.

[24] Kim, Y. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[25] Kintsch, W. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological review 95*, 2 (1988), 163.

[26] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences 114*, 13 (2017), 3521–3526.

[27] Kumaran, D., Hassabis, D., and McClelland, J. L. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences 20*, 7 (2016), 512–534.

[28] Li, Q., Zhang, Q., and Si, L. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pp. 1173–1179.

[29] Li, Z., and Hoiem, D. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence 40*, 12 (2017), 2935–2947.

[30] Liao, Q., Chai, H., Han, H., Zhang, X., Wang, X., Xia, W., and Ding, Y. An integrated multi-task model for fake news detection. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[31] Liu, P., Qiu, X., and Huang, X. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vancouver, Canada, July 2017), Association for Computational Linguistics, pp. 1–10.

[32] Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., and Cha, M. Detecting rumors from microblogs with recurrent neural networks.

[33] Ma, J., Gao, W., and Wong, K.-F. Detect rumor and stance jointly by neural multi-task learning. In *Companion proceedings of the the web conference 2018* (2018), pp. 585–593.

[34] Ma, J., Gao, W., and Wong, K.-F. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference* (2019), pp. 3049–3055.

[35] Mar, R. A. The neuropsychology of narrative: story comprehension, story production and their interrelation. *Neuropsychologia 42*, 10 (2004), 1414–1434.

[36] Mayer, R. E. Multimedia learning. In *Psychology of learning and motivation*, vol. 41. Elsevier, 2002, pp. 85–139.

[37] Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Networks 113* (2019), 54–71.

[38] Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 1532–1543.

[39] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[40] Qi, P., Cao, J., Yang, T., Guo, J., and Li, J. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM)* (2019), IEEE, pp. 518–527.

[41] Qian, F., Gong, C., Sharma, K., and Liu, Y. Neural user response generator: Fake news detection with collective user intelligence. In *IJCAI* (2018), vol. 18, pp. 3834–3840.

[42] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog 1*, 8 (2019), 9.

[43] Rubin, V. L., and Lukoianova, T. Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology 66*, 5 (2015), 905–917.

[44] Ruchansky, N., Seo, S., and Liu, Y. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (2017), pp. 797–806.

[45] Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. Progressive neural networks. *arXiv preprint arXiv:1606.04671* (2016).

[46] Schnotz, W. Commentary: Towards an integrated view of learning from text and visual displays. *Educational psychology review 14*, 1 (2002), 101–120.

[47] Schnotz, W., and Bannert, M. Construction and interference in learning from multiple representation. *Learning and instruction 13*, 2 (2003), 141–156.

[48] Schuster, M., and Nakajima, K. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2012), IEEE, pp. 5149–5152.

[49] Schuster, M., and Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing 45*, 11 (1997), 2673–2681.

[50] Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019), pp. 395–405.

[51] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286 8* (2018).

[52] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[53] Singhal, S., Kabra, A., Sharma, M., Shah, R. R., Chakraborty, T., and Kumaraguru, P. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *AAAI* (2020), pp. 13915–13916.

[54] Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., and Satoh, S. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)* (2019), IEEE, pp. 39–47.

[55] Singhal, S., Shah, R. R., and Kumaraguru, P. Factorization of fact-checks for low resource indian languages. *arXiv preprint arXiv:2102.11276* (2021).

[56] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems 30* (2017), 5998–6008.

[57] Vosoughi, S., Roy, D., and Aral, S. The spread of true and false news online. *Science 359*, 6380 (2018), 1146–1151.

[58] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (2018), pp. 849–857.

[59] Wu, L., Rao, Y., Jin, H., Nazir, A., and Sun, L. Different absorption from the same sharing: Sifted multi-task learning for fake news detection. *arXiv preprint arXiv:1909.01720* (2019).

[60] Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., and Hu, X. Xfake: explainable fake news detector with visualizations. In *The World Wide Web Conference* (2019), pp. 3600–3604.

[61] Yang, J., and Zhang, Y. Ncrf++: An open-source neural sequence labeling toolkit. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (2018).

[62] Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., and Yu, P. S. Ti-cnn: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749* (2018).

[63] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237* (2019).

[64] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. Defending against neural fake news. In *Advances in Neural Information Processing Systems* (2019), pp. 9054–9065.

[65] Zhou, X., Wu, J., and Zafarani, R. Safe: Similarity-aware multi-modal fake news detection. *arXiv preprint arXiv:2003.04981* (2020).

[66] Zhou, X., and Zafarani, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR) 53*, 5 (2020), 1–40.

[67] Zwaan, R., and Radvansky, G. Situation models in language comprehension and memory. *Psychological bulletin 123*, 2 (March 1998), 162—185.