DeBERTNeXT: A Multimodal Fake News Detection Framework

Kamonashish Saha
University of Windsor

Follow this and additional works at: https://scholar.uwindsor.ca/etd

Part of the Computer Sciences Commons

Recommended Citation
https://scholar.uwindsor.ca/etd/9073

This online database contains the full-text of PhD dissertations and Masters' theses of University of Windsor students from 1954 forward. These documents are made available for personal study and research purposes only, in accordance with the Canadian Copyright Act and the Creative Commons license—CC BY-NC-ND (Attribution, Non-Commercial, No Derivative Works). Under this license, works must always be attributed to the copyright holder (original author), cannot be used for any commercial purposes, and may not be altered. Any other use would require the permission of the copyright holder. Students may inquire about withdrawing their dissertation and/or thesis from this database. For additional inquiries, please contact the repository administrator via email (scholarship@uwindsor.ca) or by telephone at 519-253-3000 ext. 3208.
DeBERTNeXT: A Multimodal Fake News Detection Framework

By

Kamonashish Saha

A Thesis
Submitted to the Faculty of Graduate Studies through the School of Computer Science in Partial Fulfillment of the Requirements for the Degree of Master of Science at the University of Windsor

Windsor, Ontario, Canada

2023

©2023 Kamonashish Saha
DeBERTNeXT: A Multimodal Fake News Detection Framework

by

Kamonashish Saha

APPROVED BY:

________________________________________
N. Zhang
Department of Electrical and Computer Engineering

________________________________________
P. M. Zadeh
School of Computer Science

________________________________________
Z. Kobti, Advisor
School of Computer Science

April 12, 2023
DECLARATION OF CO-AUTHORSHIP AND PREVIOUS PUBLICATION

I. Co-Authorship

I hereby declare that this thesis incorporates material that is a result of research conducted under the supervision of Dr. Ziad Kobti (Advisor). In all cases, the key ideas, primary contributions, experimental designs, data analysis and interpretation, were performed by the author, and the contribution of coauthors was primarily through the proofreading of the published manuscripts.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis.

I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

II. Previous Publication

This thesis includes one original paper that has been submitted for publication in peer-reviewed conference/journal, as follows:

<table>
<thead>
<tr>
<th>Section</th>
<th>Full Citation</th>
<th>Publication Status</th>
</tr>
</thead>
</table>

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor.
II. General

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone’s copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.
ABSTRACT

There is a rapid influx of fake news nowadays, which poses an immense threat to our society. Fake news has been impacting us in several ways which include changing our thoughts, manipulating opinions, and also causing chaos due to misinformation. With the ease of access and sharing information on social media platforms, such fake news or misinformation has been spreading in different modalities which include text, image, audio, and video. Although there have been a lot of approaches to detecting fake news in textual format only, however, multimodal approaches are less frequent as it is difficult to fully use the information derived from different modalities to achieve high accuracy in a combined format. To tackle these issues, we introduce DeBertNeXT which is a multimodal fake news detection model that utilizes both textual and visual information from an article for fake news classification. We perform experiments on the immense Fakeddit dataset and two other smaller benchmark datasets named Politifact and Gossipcop. Our model outperforms the existing models on the Fakeddit dataset by about 3.80%, Politifact by 2.10% and Gossipcop by 1.00%. 
ACKNOWLEDGEMENTS

I would like to take this moment to show my sincere appreciation to everyone who played a role in not only helping me finish my research thesis but also making my time at the University of Windsor an unforgettable experience. Firstly, I would like to extend my heartfelt gratitude to my supervisor, Dr. Ziad Kobti, for his unwavering support, encouragement, and exceptional expertise. He provided me with the freedom to choose my research area and discover this domain. His encouragement has been a great source of support throughout my studies and while writing this thesis. In addition to my supervisor, I would like to thank my thesis committee, Dr. Ning Zhang and Dr. Pooya Moradian Zadeh, for their valuable feedback and inspiration that helped me complete this thesis.

I sincerely thank my parents for their unwavering love and unconditional care. Their support throughout these years has been a great source of motivation for me. They have always believed in me more than I believed in myself. I thank my younger brother for consistently being there and encouraging me to keep going. I am grateful and give special thanks to Jennifer, who has inspired me, pushed me and is always there to help and support me.

Last but not least, my experience at Windsor University would not have been the same without the support of my friends. I want to express my sincere gratitude to all my friends who have been essential pillars of support throughout my academic journey.
# TABLE OF CONTENTS

**DECLARATION OF CO-AUTHORSHIP AND PREVIOUS PUBLICATION** iii

**ABSTRACT** v

**ACKNOWLEDGEMENTS** vi

**LIST OF TABLES** ix

**LIST OF FIGURES** x

1 Introduction 1
   1.1 Background ......................................................... 1
   1.2 Problem Definition ................................................ 3
   1.3 Motivation ......................................................... 4
   1.4 Thesis Statement ................................................... 5
   1.5 Thesis Contribution ................................................ 5
   1.6 Thesis Organization ................................................ 6

2 Literature Review 8
   2.1 Defining Fake News .................................................. 8
      2.1.1 Traditional News Media for Fake News ....................... 9
      2.1.2 Online Social media for Fake News ......................... 11
   2.2 Feature Extraction .................................................. 12
      2.2.0.1 User-Based Features ...................................... 13
      2.2.0.2 News content-based features ............................ 13
      2.2.0.3 Social Context-based Features ........................ 14
      2.2.1 Textual content based .................................... 15
      2.2.2 Visual content based ..................................... 16
      2.2.3 Multimodal based content based ......................... 17
   2.3 Detection of Fake News .............................................. 18
      2.3.1 Knowledge base ............................................... 18
      2.3.2 Style Based .................................................. 20
      2.3.3 Social context based ....................................... 21
      2.3.4 Text Based Content ....................................... 21
      2.3.5 Visual Based Content .................................... 22
      2.3.6 Multimodal Based Content ................................ 22
   2.4 Background of Models used in DeBERTNeXT ....................... 26
      2.4.1 Model for text: DeBERTa ................................... 26
      2.4.2 Model for image: ConvNeXT ................................ 28
   2.5 Reasoning of the models used in DeBERTNeXT .................... 29
      2.5.1 Model for text: DeBERTa ................................... 30
      2.5.2 Model for image: ConvNeXT ................................ 31
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1</td>
<td>Advantages and disadvantages of some machine learning-based fake news detection models</td>
<td>23</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Comparison with other models on Fakeddit dataset. (+) indicates that the model does not classify using text and image; (-) indicates that results were not published</td>
<td>48</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Comparison with other models on Politifact and Gossipcop Dataset. (-) indicates that the results were not published</td>
<td>51</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Comparison of the accuracy of models ran on the same environment for Politifact and Gossipcop Datasets</td>
<td>54</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Comparison of the accuracy of the models ran on the same environment for Fakeddit Dataset</td>
<td>55</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

1.1.1 Analysis of the engagement of Fakebook users indicate that Fake news outperformed mainstream news during the US presidential election [45] ................. 2

1.1.2 News regarding 5G technology tied to COVID-19 spread was discarded as fake by a fact-checking website ........................................... 3

2.1.1 Components and major factors of fake news [96] ........................................... 10

2.2.1 Features of Fake News Detection [96] ......................................................... 13

2.2.2 Fake Post on social media [96] ................................................................. 17

2.3.1 Shows the architecture Spotfake [75] ......................................................... 25

2.4.1 Shows the architecture of DeBERTa [24] ................................................. 27

2.4.2 Architecture of Vision Transformer [15] ................................................... 29

2.5.1 Difference between Swin Transformer and ViT [48] ................................... 31

2.5.2 Comparison of the block design of a Swin Transformer, ResNet and ConvNeXt [49] ................................................................. 33

3.2.1 Model Architecture ................................................................................. 37

4.1.1 FakeNewsNet data integration process [72] ......................................... 41

5.2.1 Accuracy plot on Fakeddit ................................................................... 50

5.2.2 Loss plot on Fakeddit ......................................................................... 50

5.2.3 Accuracy plot on Politifact ................................................................. 52

5.2.4 Loss plot on Politifact ...................................................................... 52

5.2.5 Accuracy plot on Gossipcop ............................................................. 53

5.2.6 Loss plot on Gossipcop ................................................................. 53
CHAPTER 1

Introduction

1.1 Background

The rapid development of technology has increased the spread of news globally. Due to the advancement and ease of accessibility of social media platforms, an increasing amount of people are referring to social media for consuming news rather than traditional news organizations. According to a research in 2016, about 62% of American citizens have stated social media platforms as their source of consuming news [18], as compared to only 49% in 2012 [71]. This shows that in addition to the traditional news media, much of the news is viewed through online social media platforms which include Facebook, Instagram, Reddit, etc. This increasing consumption of news apart from traditional sources is mainly due to the shift in user habits as the use of social media is more frequent and it is easier to access and less costly comparatively. Social media platform plays a vital role in overcoming geographical barriers and lead to an eruption of virtual presence and online communication.

There are an ample number of advantages of the use of social media which includes information sharing, publicity, opinion and feedback sharing, online marketing, product, and brand awareness, etc. [16][6]. In addition, it also plays a significant role in some major events such as situational awareness, crisis, fundraising during disasters, support etc. [84]. Despite the immense benefit of the advancement of technology and the increasing use of social media, the creation of fake information and spreading it through social media is also increasing and harming society.
Fake news has been a well-discussed topic for a significant amount of time. It has recently started to get into the highlights after the US presidential election in 2016 as online false news has been spreading mostly through social media. It severely affected the citizen’s voting opinions on political choices which lead to unfairness and bias to some extent in the election results. Moreover, due to the ease of accessibility of technology and the ease of sharing content over social media platforms without adequate authentication and verification, there has been an eruption of fake news in almost all important aspects which greatly impacts and causes significant damage to the community. There can be different motives for the spread of such false information over social media platforms which includes defamation, political influence, increasing group polarization, etc. In the year 2020, fake news related to the conspiracy theories of COVID-19 tied with the 5G technology was very widely spread and shared across several social media platforms. This caused the whole world to panic and negatively impacted many people’s views regarding 5G technology. As per some research, some of the reasons behind the spread of fake news include 1) the reader’s lack of adequate knowledge about the topic 2) the authenticity of the news.
3) oblivious about the credibility of the news source. All these can cause direct or indirect reactions and chaos among the readers.

![Fake News Example](image)

Fig. 1.1.2: News regarding 5G technology tied to COVID-19 spread was discarded as fake by a fact-checking website

### 1.2 Problem Definition

Identifying fake news on social media entails a number of significant research problems. Fake news is purposefully created to confuse readers, making it difficult to spot based just on the textual substance. Apart from the textual content, features from different modalities are added along with the textual content to make it even more believable to the readers.

Similar to [68], we define our problem statement as follows:

Given a set of $n$ news articles which includes text and image content (multimodal),
the data as a collection of a text-image tuple can be represented as:

\[ A = (A^T_i, A^I_i)^n \]

(1)

Where, \( A^T_i \) represents textual content, \( A^I_i \) represents the image content and \( n \) represents the number of news articles. Since we are considering fake news detection as a binary classification problem, we represent labels as \( Y = \{0, 1\} \) where 0 represents fake and 1 represents real or true news. For a given set of news records \( A \), a set of features can be extracted from the textual and image information as represented by \( X^T_i \) and \( X^I_i \). The objective of multimodal fake news detection is to create a model \( F : \{X^T_i, X^I_i\} \epsilon X \rightarrow Y \) inorder to deduce the potential labels in news articles \( A \). Hence, the task of the model is to detect whether the news article \( A \) is either fake or real such that:

\[
    f(A) = \begin{cases} 
    0, & \text{if } a \text{ is fake news} \\
    1, & \text{otherwise} 
  \end{cases}
\]

(2)

### 1.3 Motivation

There has been a significant amount of research in the detection of fake news using textual content only with high accuracy. Recently, fake news can be more engaging to the readers and provides a better experience as there is an addition of visually-supported reporting, which includes images and videos that are distorted, forged and intentionally designed to be misinterpreted and mislead the readers. As per one of the research studies [33], it has been found that the average number of retweets is 11 times higher for articles consisting of images compared to the ones without images and with only text. Therefore, the classification of fake news using machine learning and deep learning methods that uses a combination of data from different modalities, which includes audio, video, image, and text [59] has become very crucial. Hence, the motivation for this is to utilize the multimodal information in a fake news detection model due to the following reasons:
1. INTRODUCTION

- Different modalities depict different parts of the news;
- Different sources may exploit different modalities compared to others. E.g. some sources may tamper with the image more compared to the text, while others can change the text to a large extent but not tamper with the image by much;
- When determining the veracity of news, information from different modalities complements one another.

1.4 Thesis Statement

There has been a significant amount of research in the detection of fake news using unimodal information. A majority of the research focuses on the usage of textual information as previously that was considered as the main source of news. Although such unimodal detection models are capable of detecting the fake news which utilizes textual content and attains a very considerable accuracy, however, they do not consider the significance of the other modalities which has recently got into spotlight. In addition, as per our literature survey, the existing multimodal detection models are heavily dependent on subtasks and are event oriented. While some of the models are only suited for a large dataset while others are suited for small ones only. Therefore, those models leave a lot of scope for improvement in terms of accuracy and other performance matrices. Hence, to solve these issues, we aim to design a fake news detection model utilizing the power of the transformer and the convolutional network architecture to classify a news record consisting of information from two different modalities as either real or fake. We hypothesize that our proposed method will outperform other currently existing models for similar task in terms of accuracy and other evaluation methods.

1.5 Thesis Contribution

Overall, the major contributions of our work are as follows:
- We propose DeBERTNeXT, which is based on transfer learning that utilizes the
textual features from the DeBERTa V3 model and the visual features learned from the ConvNeXT model. The DeBERTa model was built to overcome some of the limitations of the famous BERT model and the ConvNeXT was pre-trained on the large ImageNet dataset.

- The proposed architecture can take the images and texts of the news article as input and classifies it as either real or fake. Moreover, it is not dependent on any sub-task and is not domain-specific during the classification process. The representations from the two modalities (text and image) are concatenated to finally get to the classification result.

- Lastly, we have trained and tested the DeBertNeXT model on the immense Fakeddit dataset consisting of about 1 million records. Moreover, we have also experimented on two other benchmark datasets named Politifact and Gossipcop which are much smaller comparatively. To the best of our knowledge, we have surpassed the accuracy of other baseline models on the Fakeddit dataset by 3.80%, Politifact by 2.10% and Gossipcop by 1.00% and attained very good results on Precision, Recall, and F1 scores as compared to other models.

1.6 Thesis Organization

The thesis/research work is structured in the following way:

In chapter II, we give background knowledge about Fake News Detection by defining it and explaining the different ways it is spread. We will also give an explanation of its characterization and feature extraction. We will also discuss the related works/literature review in this field, which solves a common problem using different techniques.

In chapter II, we will explain the proposed model in detail and introduce the architecture we used for the multimodal fake news detection problem.

In chapter IV, we will discuss the experimental setup for the different types of datasets used for the experiment.

In chapter V, we compare the other methods and approaches that aim to solve
1. INTRODUCTION

the common problem and explain the insights received during the work. We will also
give a detailed analysis and the observations made.

In chapter VI, we will conclude the research paper, where we will provide various
opportunities for future work and solve unsolved problems.
CHAPTER 2

Literature Review

2.1 Defining Fake News

Fake news has been a well-discussed topic for a significant amount of time. Nowadays, false news is frequently obtrusive and varied in terms of topics, formats, and platforms \cite{71}, and so it is difficult to generate formally accepted definitions for the term \cite{71}. Fake news can be defined in several ways. In the earlier definitions, content such as satire, hoaxes, news propaganda, and click baits was defined as fake news. Another widely accepted definition of fake news \cite{3} focuses on the two key features which include:

1. The authenticity of the news: Where the news contains some information that is verifiably false

2. The intent of the news: Where the news has been populated with false information to mislead the readers and meet the creator’s dishonest intentions.

However, recently fake news has been defined as news that is intentionally and variably false, and that could mislead readers \cite{3}. Due to the advancement of technology, more fake news is being generated and distributed across social media platforms. According to \cite{96}, fake news has three basic characteristics:

1. Volume: Due to the ease of access to the internet, anyone is capable of creating fake news and sharing it on the Internet \cite{2}. This creates the scope of even creating websites such as denverguardian.com, wtoe5news.com that are indented for publishing such false information while keeping it to resemble legit news organizations.
2. LITERATURE REVIEW

Distributions of false information in this way makes the news more credible to the readers while fulfilling the creator’s intentions which can often be financial or political gain.

2. Variety: Fake news impacts people’s choices, opinions, interests and decisions. With the increasing use of social media platforms, false information is being created in almost all aspects that impact all types of people. For example, some news is created just to confuse and mislead young people as they lack adequate consciousness, while others impact older adults.

3. Velocity: Fake news creators tend to be short-lived \[^3\]. For example, during the US presidential election in 2016, there were several active fake news websites that no longer exist. Fake news usually propagates a lot in current affairs as they receive more attention and are typically short lasting. Hence, identifying them becomes more difficult.

The authors in \[^96\] also discussed four major components of fake news. These are as followed:

- Creator/Spreader: This can be either human or bots where some creators create the news unintentionally while other users intentionally create the fake news.

- Target Victims: These are usually the people for whom that specific false content was created. This can include youth who are primary users of social media platforms, voters, parents, etc.

- News Content: This is the actual news which consists of the information itself and includes content such as title, body text etc., along with additional multimedia such as images and video.

- Social Context: This resembles the way the news will be spread through the internet.

2.1.1 Traditional News Media for Fake News

Fake news has been spreading for a long time in several ways. “Traditional Fake News” is considered as the news that has been spreading for a long time in ways before social media. Such methods of spread include the following:
2. LITERATURE REVIEW

Fig. 2.1.1: Components and major factors of fake news [96]

Tabloids [11]: Tabloids are a type of newspaper that tends to prioritize sensational stories and eye-catching headlines over accuracy and truthfulness. They often feature stories about celebrities, crime, and other sensationalist topics that are designed to attract readers’ attention. Tabloids are known for their flashy headlines and provocative front pages, which can sometimes mislead readers into believing that the stories they are reading are true [30]. One of the ways in which tabloids can spread fake news is by publishing stories that are not based on verified facts or credible sources. In some cases, they may even knowingly publish false information in order to attract more readers or generate controversy. For example, a tabloid might publish a story claiming that a celebrity has been caught in a scandal without verifying whether the story is true or not.

Broadcast media: Television news programs and radio news shows are sometimes guilty of spreading fake news when they fail to properly fact-check stories or repeat information without verifying its accuracy [78].

Yellow journalism [87]: This type of journalism prioritizes sensational headlines and stories over accuracy and often uses misleading or false information to sell
Propaganda: Propaganda is a type of media that is designed to manipulate people’s beliefs and opinions. It often relies on false or misleading information to push a particular political agenda [78].

Rumor mills: These are informal networks of people who spread false information through word of mouth or social media. While they may not be traditional news sources, they can still have a significant impact on public opinion [62].

2.1.2 Online Social media for Fake News

In modern times, the use of social media has increased due to the ease of access and availability to the internet. Some notable social media platforms include Facebook, Instagram, Reddit etc., and these are some of the biggest platforms for the spread of fake news. Fake news can spread in different ways in social media platforms and some of them are discussed below:

1- Impersonation [63]: This involves creating fake accounts or pages that impersonate legitimate news outlets, journalists, or public figures. The goal is to make the fake account or page appear as credible as possible, so that people are more likely to believe and share the false information being posted. This can be done by using the same profile picture and username as the real account, as well as copying the layout and design of the real page.

2- Clickbait headlines [71]: Clickbait headlines use sensational language or make outrageous claims to grab people’s attention and encourage them to click on an article. Once people click on the article, they may be presented with false or misleading information. Clickbait headlines can be used to spread fake news quickly and easily on social media, as people are more likely to share an article if the headline is eye-catching.

3- Memes and manipulated images [78]: Memes and manipulated images are popular on social media, and they can be used to spread false information quickly and easily. For example, a fake quote from a politician can be paired with an image of them to create a meme that is then shared widely on social media. Similarly,
images can be manipulated to make it look like a celebrity or public figure said or
did something they did not.

4- Hashtag hijacking [58]: Hashtag hijacking involves creating false or misleading
hashtags to spread false information or to divert attention away from a topic. For
example, during a protest or demonstration, someone may create a false hashtag to
make it appear as though the protesters are advocating for something that they are
not.

5- Conspiracy theories [71]: Conspiracy theories are often spread on social media
to create distrust in legitimate sources of news and information. Conspiracy theories
may involve false claims about a particular event or person, or they may involve a
broader conspiracy to control the population or withhold information from the public.

6- Automated accounts [44]: Bots or automated accounts can be used to share false
information quickly and widely on social media. These accounts can be programmed
to post and share content at specific times or in response to specific keywords or
hashtags. Automated accounts can be used to spread false information or to amplify
the reach of a particular message or agenda.

7- Echo chambers [71]: Echo chambers are closed networks of like-minded individuals
who reinforce each other’s beliefs and ignore evidence that contradicts their views. In
these networks, false information can be circulated and believed without question, as
people are more likely to believe information that confirms their existing beliefs and
biases.

Figure 2.1.1 depicts the components and major factors of the spread of Fake news

2.2 Feature Extraction

There are several features that are specifically used for fake news representation,
which are divided into three main categories namely Created-based features, News
Content-based features and Social Context-based features 2.2.1.
2. LITERATURE REVIEW

2.2.0.1 User-Based Features

The user-based features are designed to identify specific characteristics of suspicious users or non-human accounts and are categorized into user profiling, user credibility, and user behaviour features.

User profiling features gather basic information about the user, such as their account name, geolocation data, and registration date. User credibility features evaluate the impact and reliability of the online account, taking into account features like the user’s credibility score, number of friends and followers, and total posts or tweets. User behaviour features are intended to identify patterns in the online behaviour of both legitimate and deceptive users. These features are part of a larger set of social context features. They include the user anomaly score, which is calculated by dividing the number of the user’s interactions within a specific time frame by their monthly average.

2.2.0.2 News content-based features

News content-based features can also be used to detect fake news. These features can be categorized into three types: linguistic and syntactic-based features, style-based features, and visual-based features. These features provide explicit clues for fake news detection and are commonly used for fake news representation and analysis.

Linguistic and syntactic-based features are fundamental aspects of natural language,
including components like structure and semantics. These features are valuable for analyzing suspicious news content, even though fake news is intentionally generated to mislead online users. The linguistic and syntactic-based features can be categorized into three types: word-level, sentence-level, and content-level features. Word-level features include bag-of-words, n-gram, term frequency (TF), Term Frequency-Inverted Document Frequency (TF-IDF), etc. Sentence-level features include Parts of Speech (POS), average sentence length [29], frequency of punctuations etc. Lastly, the content-level features include the raw information of the meta news content [71].

The goal of style-based features is to distinguish the unique characteristics of writing styles between fake news authors and true news authors. Despite fake news authors attempting to imitate the writing style of legitimate news authors, there are still detectable differences that can be used to identify fake news creators.

Visual-based features are one of the most important aspects of fake news detection, as news with visual features tends to be more believable and spread faster. Visual-based features such as the number of images or videos, clarity and coherence scores, similarity distribution histogram, diversity score, clustering scores, and image ratios are crucial in identifying suspicious or deceptive information in new content.

2.2.0.3 Social Context-based Features

Social context-based features are intended to show how online news is distributed and how users interact with it. These features can be categorized into three types: network-based features, distribution-based features, and temporal-based features.

Network-based analysis: Network-based analysis is concerned with a specific group of online users with similar characteristics such as location, educational background, and habits. Network-based features are selected and extracted from these groups to study their unique characteristics and the similarities and differences between different online accounts.

Impact-based features: Impact-based features are used to capture the unique diffusion pattern of online news, typically by building a propagation tree to describe the distribution nature of a news item. Features related to the propagation tree
include the degree of the root, the maximum number of subtrees, and the maximum/average degree and depth of the tree.

Temporal-based features: Temporal-based features are used to describe the posting behaviour of online news creators in a time series, which can be useful for detecting suspicious posting activities and indicating the level of the falseness of online news. Commonly used temporal-based features include the interval between two posts, the frequency of posting, replying, and commenting for a certain account, the time of day when the original information is posted/shared/commented, and the day of the week on which the post is published.

In this thesis, we will specifically be working on Textual based content and Visual based content. Hence, we will give a more in-depth explanation about the textual and visual features that help in the characterization of fake news.

2.2.1 Textual content based

News content is a combination of various elements, and it can be broadly categorized into four main components. The first component is the source, which is the origin of the news. It provides information on who created the news, where it came from, and whether the source is trustworthy or not. The credibility of the source is essential in determining the reliability of the news.

The second component is the headline, which is a brief summary of the news content. The headline’s purpose is to attract readers’ attention and give them an idea of the news story’s main topic or theme. It should be a concise and accurate description of the news story.

The third component is the body text, which contains the actual substance of the news article. It provides detailed information on the news topic, including background information, quotes from sources, and other relevant details. The body text should be informative, engaging, and objective.

The fourth and final component is any attachments such as visual content, which include images, videos, or audio, which is often added to textual content to enhance the news story’s visual appeal. Visual content can help illustrate the news story,
2. LITERATURE REVIEW

making it more engaging and easier to understand.

To detect fake news, the content of the news article is crucial. The material of a news report can be classified into both textual and visual modalities, and textual modality is particularly useful in identifying fake news. Analyzing the language used, fact-checking sources, and checking for bias are some of the methods used to determine the accuracy and reliability of news content.

2.2.2 Visual content based

Multimedia technology refers to the use of multiple forms of media such as audio, video, images, and text to convey information. With the rapid advancement of multimedia technology, self-media news has evolved from text-based posts to picture or video multimedia posts. This evolution has provided better storytelling capabilities and attracted more readers’ attention. In the past, self-media news was primarily text-based. While this approach allowed for in-depth reporting and analysis, it lacked the visual elements that can help bring a story to life. With the advancement of multimedia technology, self-media news can now be enhanced with pictures, videos, and audio that can help readers to better understand and connect with the content. Picture and video multimedia posts have several advantages over text-based posts as they are more engaging, capture attention faster, and are more likely to be shared on social media platforms. Multimedia content is also more likely to be remembered than text-based content, making it a powerful tool for self-media news creators. Moreover, the use of multimedia content in self-media news has enabled creators to tell more compelling stories. By using a mix of different media formats, creators can create a more immersive experience that helps readers to connect with the content on a deeper level. Fake news takes advantage of multimedia technology by using misrepresented or tampered images to attract and mislead readers, resulting in visual content becoming an essential part of fake news that cannot be ignored [60].
2.2.3 Multimodal based content based

Figure 2.2.2 which has been collected from [96], shows an example of fake news shared on Facebook and how the different elements of the news can be identified. The sample post shows the four key elements of the news: the creator/spreader (A), the news content (B), the social context (C), and the target (D). In this example, the Facebook user named Bob is identified as the news spreader, which is part of the creator/spreader element (A). The news content (B) includes the title of the news, the body of the news, any available multimedia, and the comment from Bob. The social context (C) includes all the interactions between other users and this news, such as comments, likes, dislikes, timestamps, and so on. These interactions help to create a social context that can influence how the news is perceived and shared. Finally, the target (D) includes all the potential users who may come across the news, interact with it, and be influenced by it. These users can be considered the target audience for
fake news. Overall, this example highlights how fake news can spread through social media platforms like Facebook where both text and image play a vital role, and how the different elements of the news can be identified and analyzed to understand the impact it may have on the target audience.

2.3 Detection of Fake News

In this part, we will describe the detection mechanisms for fake news and will focus on the textual and visual based detections.

2.3.1 Knowledge base

Testing the truthfulness of major statements in a news article is the best way to assess the truthfulness of the news. To do this, approaches based on knowledge, which involves using external sources to verify reported statements in news reporting, can be used. Fact-checking is a method that was initially developed in journalism, and it seeks to test the validity of news by matching information derived from news information to be checked (e.g., its claims or statements) with known facts (e.g., true information). There has been increasing focus on fact-checking, and several attempts have been made to build practical automated fact-checking programs and manual fact-checking can be categorized into three types:

(I) Expert-based fact-checking: Expert-based fact-checking is a widely used method of verifying the accuracy of news reporting. This approach involves using domain experts or fact-checkers to evaluate the truthfulness of news claims [31]. These experts are typically individuals who have specialized knowledge in a particular field, and they are able to assess the validity of claims made in news articles based on their expertise. Expert-based fact-checking is often done by a small group of highly reliable fact-checkers. This makes the fact-checking process easier to handle and can lead to reasonably accurate outcomes. However, this approach can be expensive, as it requires hiring and training a group of domain experts to conduct the fact-checking process. Hence, Expert-based fact-checking may not be scalable with the
rise in the amount of news content to be reviewed. As the volume of news articles and reports continues to increase, it may become more challenging to hire enough domain experts to handle the workload. However, while expert-based fact-checking has some limitations, it remains an important and valuable approach to verifying the accuracy of news reporting. By leveraging the expertise of domain experts, it is possible to obtain a deeper understanding of complex issues and identify false or misleading claims in news articles.

(II) Crowd-sourced fact-checking is a technique to determine the accuracy of news content by using a large group of ordinary individuals to serve as fact-checkers. Unlike expert-based fact-checking, which is typically carried out by a small group of experts, crowd-sourced fact-checking relies on the collective intelligence of the crowd. However, the crowd-sourced approach is less reliable and precise due to the fact-checker’s political bias and contradictory annotations. Still, this method can scale to handle a large volume of news content. The method works by encouraging average people to annotate news content and then aggregating these annotations to create an overall evaluation of the news veracity. Crowd-sourced fact-checking websites are still in their early stages, much like expert-based fact-checking. One example of a crowd-sourced fact-checking website is www.fiskkit.com, which allows users to upload posts and provide ratings for individual sentences within those posts, as well as select tags to describe the content. The sources cited in the articles help differentiate between news and non-news content and assess their credibility. Tags categorized into multiple dimensions enable trend analysis across fake and actual news articles.

(III) Computer-oriented fact-checking is a method of verifying the accuracy of news reporting that uses automation to detect true and false statements. It addresses two major issues, namely recognizing credible statements and distinguishing the veracity of statements of fact. This method extracts factual arguments from news material to facilitate the fact-checking process. To assess the truthfulness of a particular argument, computer-oriented fact-checking relies on outside resources, such as open web sources, to contrast against the particular statements being analyzed.
2.3.2 Style Based

Fake news publishers often have the intention to spread misleading and false information in order to manipulate a large number of people. To achieve this goal, they use different types of writing strategies to convince their audience. These strategies are different from those used by real news publishers because they are focused on deceiving readers rather than providing factual and unbiased information. One way to identify fake news is through style-based strategies that focus on the writing format. This approach involves analyzing the writing style and identifying patterns that are commonly used by fake news publishers to manipulate their readers. There are two main categories of style-based approaches: deception-oriented and objectivity-oriented.

**Deception-oriented** stylometric methods use forensic techniques to identify tricky statements in news articles. These methods draw inspiration from forensic psychology and include Criteria-based content analysis and Scientific-based content analysis. More advanced models, such as those based on natural language analysis, are also used to identify different stages of deception, including deep syntax and rhetorical structure.

**Objectivity-oriented** approaches aim to identify style cues that suggest a lack of objectivity in news content. This could indicate a potential for misleading audiences through hyper-partisan types and yellow journalism. Hyper-partisan types refer to actions that disproportionately target a single political group, often with the intent of producing false news. Linguistic-based features can be used to identify hyper-partisan articles, while yellow journalism refers to articles that rely on eye-catching headlines and tend to exaggerate, sensationalize, or scaremonger rather than presenting well-researched news. Deceptive clickbait titles can serve as a good indicator of inaccurate and deceptive articles, as titles often summarize the main points of view the author wishes to express.
2.3.3 Social context based

The nature of social media provides researchers with additional resources to improve news content models. Social context models involve collecting specific user social commitments from different perspectives to supplement the study. There are two categories of social context modeling methods: stance-based and propagation-based. Stance-based methods focus on identifying the attitude or opinion of users towards a particular topic, while propagation-based methods focus on analyzing how information spreads through social media networks.

2.3.4 Text Based Content

Textual features refer to statistical or semantic characteristics that can be extracted from the text content of messages, such as word frequency, sentence structure, sentiment analysis, and other linguistic features. These features are often used in the identification of fake news as they can provide insight into the writing style, tone, and emotional sentiment of the content. Most of the previous research on fake news detection has relied heavily on these textual features, along with user metadata, to identify and distinguish fake news from genuine content. By analyzing these features, it is possible to identify specific patterns and commonalities that are commonly found in fake news content, allowing for more accurate and reliable detection.

Moreover, Natural Language Processing (NLP) based processes involve a range of tasks, including pre-processing, word embedding, and feature extraction techniques. In the context of fake news detection, several models utilize data pre-processing as the initial step to represent obscure attributes, manage lost words, binarize attributes, and deal with complicated structures. Data pre-processing can also help to visualize the data and solve noisy data problems, while saving space and computational time. Word vectorizing is another key process in NLP-based fake news detection models. It involves mapping text or words to a list of vectors, which can then be used for analysis and classification. Common techniques for word vectorization include the bag of words model and TF-IDF (term frequency-inverse document frequency). However,
in recent years, pre-trained word embedding models like word2vec and GloVe have become increasingly popular due to their ability to handle larger datasets and improve accuracy.

2.3.5 Visual Based Content

Visual features extracted from visual elements are effective in detecting fake news. However, limited studies have verified the value of multimedia content on social media [33 71]. Recently, various visual and statistical features have been extracted for news prediction, including the use of a classification framework to recognize fake images based on user-level and tweet-level features [20]. Another research [52] has looked deeper into the efficiency of several fake image detectors using GANs for image-to-image conversion, but these models are still hand-crafted and complex to represent visual content.

2.3.6 Multimodal Based Content

Several unimodal approaches have been developed and have shown promising results for detecting certain types of information. However, in real-world scenarios, news and other forms of information are often presented in multimodal formats, which means that they contain information in multiple modalities, such as text, images, and videos. Therefore, relying solely on unimodal approaches may not be efficient or effective for detecting information in these scenarios. To address this issue, several new architectures have been proposed. These architectures leverage the strengths of different modalities and combine them to improve the accuracy and efficiency of detecting information in multimodal formats. By using these architectures, it is possible to extract and combine relevant information from different modalities, which can provide more comprehensive and accurate insights into the information being analyzed. Therefore, in this section, we will discuss the multimodal-based detection frameworks.

One baseline multimodal architecture includes SpotFake [75], which uses BERT
Table 2.3.1: Advantages and disadvantages of some machine learning-based fake news detection models

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>The TF-IDF model includes information on both the more significant and less important words</td>
<td>Slow for large vocabularies. Does not capture position in text, semantics and co-occurrences in different documents</td>
<td>[34][51]</td>
</tr>
<tr>
<td>Bag-of-Words</td>
<td>It is easy to implement</td>
<td>It ignores the ordering of the words in a document and also ignores the semantic relations among words</td>
<td>[5][79]</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>Maintains the semantic meaning of various words in the text and the context information is also preserved.</td>
<td>Unable to deal with unfamiliar words. There are no common representations at the sub-word level</td>
<td>[37][80]</td>
</tr>
<tr>
<td>Doc2Vec</td>
<td>Faster than Word2Vec and a numeric representation of a document regardless of its length</td>
<td>Benefit of Doc2Vec is diminished for shorter documents</td>
<td>[65][27]</td>
</tr>
<tr>
<td>GloVe</td>
<td>Unlike Word2Vec, it does not reply on local statistics (Words local context information) only</td>
<td>In order to obtain word vectors, global statistics (word co-occurrence) are used</td>
<td>[37][1]</td>
</tr>
<tr>
<td>BERT</td>
<td>Identify and capture contextual meaning in a sentence or text, especially from large-scale data.</td>
<td>Compute-intensive at inference time. Also, it does not incorporate data preprocessing which means that performance may be impacted by noise.</td>
<td>[83][43][55]</td>
</tr>
<tr>
<td>CNN</td>
<td>It has a lower tendency to overfit</td>
<td>Training requires a longer amount of time</td>
<td>[35][43][53]</td>
</tr>
<tr>
<td>k-means</td>
<td>This algorithm can effectively handle large-scale data, and offers robust scalability, fast training, and straightforward implementation.</td>
<td>Performance is impacted by the value of parameter K and its initialization, and it is ineffective when dealing with non-convex data.</td>
<td>[94]</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>It has the ability to learn incrementally and can handle noise effectively</td>
<td>It is not well-suited for handling attribute-related data.</td>
<td>[74]</td>
</tr>
<tr>
<td>SVM</td>
<td>It can generate results that are dependable, and it can extract meaningful insights from small-scale datasets.</td>
<td>Performance is affected by the kernel function parameters, and it is not suitable for handling large-scale data or multiple classification tasks</td>
<td>[17][40][57]</td>
</tr>
<tr>
<td>RNN</td>
<td>It is able to extract more meaningful latent representations</td>
<td>It demands significant computational resources</td>
<td>[70][1]</td>
</tr>
<tr>
<td>CNN-LSTM</td>
<td>It exhibits superior performance in comparison to individual networks</td>
<td>The feature extraction technique used in this algorithm may result in the loss of important information.</td>
<td>[80]</td>
</tr>
<tr>
<td>Gated Recurrent Unit (GRU)</td>
<td>This approach provides a novel method for leveraging both user comments and news content to predict instances of false news</td>
<td>More training time is required</td>
<td>[82]</td>
</tr>
<tr>
<td>Graph Neural Network</td>
<td>It retrieves user information with a high degree of accuracy</td>
<td>Nonetheless, it introduces some uncertainty when detecting fake news.</td>
<td>[77]</td>
</tr>
<tr>
<td>LSTM</td>
<td>It extracts a novel set of features from false news</td>
<td>Training and testing is time consuming</td>
<td>[80][7][54]</td>
</tr>
</tbody>
</table>
for learning text features and pre-trained VGG-19 for the image features. The authors use two modalities, text and image, to extract features from the news articles. The extracted features are then fed into a multimodal fusion model that combines the features to make the final prediction of whether the news articles are fake or not. The authors use a Support Vector Machine (SVM) classifier as the final classification model. Figure 2.3.1 shows the architecture of Spotfake and this is a commonly used structure of multimodal fake news detection model where different authors use different models for textual and image content. SpotFake+ [76] was also introduced later which is an improved and modified version of SpotFake. The authors utilize pre-trained XLNet and VGG-19 for the combined image and text classification. Similar to our approach, the authors train their model on the Politifact and Gossipcop dataset, hence we have used this model for comparison later on in the result and analysis section. Another architecture was named Event Adversarial Neural Networks for Multi-Modal Fake News Detection (EANN), which was an end-to-end model for the event discriminator and false news detection [88]. The text representation was attained using CNN, and VGG-19 was used for image representation, where both were later concatenated for classification. Similarly, another model using VGG-19 and bi-directional LSTMs for textual features was introduced, which the authors named Multimodal Variational Autoencoder for Fake News Detection (MVAE) [41]. In addition, FakeNED [66] was introduced, which utilizes finetuned BERT and VGG-19 for binary classification. SceneFND [95] is another system which combines textual, contextual scene and visual representations to classify a news as either fake or real. The model uses scene recognition which incorporates Wilson’s method [90] and ResNet-50 to extract visual features from the images associated with the news article. The model also uses a pre-trained bidirectional long short-term memory (BiLSTM) network to encode the article’s textual content. Finally, the visual and textual features are combined using a multimodal fusion layer to build a joint representation of the scene. The joint representation is fed into fully connected (FC) dense layers followed by a softmax layer to predict the probability that the article is fake news. During training, the model is optimized using cross-entropy loss.
As per our literature review, the baseline multimodal models give good accuracy, but they mainly concentrate on using BERT, XLNet, VGG architecture, ResNet and LSTM architecture, which has some shortcomings of its own. Moreover, the discussed models are trained and tested on small datasets, some depend on events or require additional preprocessing steps. Hence, to tackle these shortcomings, we propose our framework that uses DeBERTa and ConvNeXT. The DeBERTa model was built to overcome some of the limitations of BERT and RoBERTa models, and the ConvNeXT was pre-trained on the large ImageNet dataset. We later fine-tune them to train and test on the three different datasets as described in chapter three.
2.4 Background of Models used in DeBERTNeXT

In this section, we will give an idea about the evolution and background of the DeBERTa and the ConvNeXT model, both of which we are using in the buildup of DeBERTNeXT.

2.4.1 Model for text: DeBERTa

For attaining the textual features, the DEBERTa (Decoding-enhanced BERT with disentangled attention) model [26] is used in our architecture, which is based on the famous BERT [14] and RoBERTa [47] models. DeBERTa improves these state-of-the-art models by incorporating two novel techniques:

1- Disentangled attention mechanism is used where each word is characterized using two vectors which encode their content and position.
2- Enhanced mask decoder enables training with half of the data used in RoBERTa.

DeBERTa is a transformer-based language model, which is a collection of stacked transformer blocks where before a completely connected positional feed-forward network, each block has a multi-head self-attention layer [81] 2.4.1. There is no natural way to encode word location information in the conventional self-attention process. In order to represent each input word as a vector whose value relies on its content and position, existing techniques add a positional bias to each input word embedding. Absolute position embedding or relative position embedding can be used to implement the positional bias. According to [14] [26], relative position representations are more successful for natural language understanding and generation tasks. Hence in DeBERTA disentangled attention mechanism is used which differs from all other methods in that the content and location of each input word are encoded using two independent vectors, and compute attention weights between words using disentangled matrices on their contents and relative positions.

The word (content) embedding and position embedding of each word in the input layer are combined to form a vector in BERT. However, the word representation in DeBERTa was motivated by the fact that the attention weight of a word pair depends
2. LITERATURE REVIEW

Fig. 2.4.1: Shows the architecture of DeBERTa \[24\]

not only on their contents but also by their relative positions, as by the stated example that when the terms "deep" and "learning" are used together, the dependency on one another is much stronger than when they are used separately. Hence in DeBERTa, the attention weights for the words are calculated using disentangled matrices based on their contents and relative placements, and DeBERTa is represented using two vectors that encode its content and position, respectively. The attention weight of a word pair is calculated as a sum of four attention scores using disentangled matrices upon their position and content as content-to-content, position-to-content, content-to-position and position-to-position.

One important aspect of transformer models is they incorporate the Masked Language Model (MLM) \[64\], which is similar to a fill-in-the-blank task where the architecture is trained to utilize the surrounding words of the mask token to predict the mask token correctly. DeBERTa incorporates the content, and relative position information of the context words and the absolute word position is incorporated before the softmax layer. Typically, transformer-based are pre-trained on a lot of text to learn contextual word representations using a self-supervision objective called the Masked Language Model \[14\]. Here, some tokens are masked randomly, and the language model is trained to reconstruct the sequence by predicting those masked tokens. According to the authors in \[26\], in the DeBERTa model, they corrupt the sequence by masking 15% of the tokens at random and then training the model, whereas the authors of BERT state to keep 10% of the masked token unchanged and
replace another 10% with randomly picked tokens and the remaining replaced with the [MASK] token.

2.4.2 Model for image: ConvNeXT

Transformers were first developed for the purpose of Natural Language Processing and have led to several immensely successive algorithms in this field, which include the GPT-2 [61], GPT-3 [8], BERT, RoBERTa etc. All these eventually led to the development of Transformers in the field of Computer Vision. Vision Transformers (ViT) was developed, which gained immense popularity in the Computer Vision sector over Convolutional Neural Networks (ConvNets), which are trained using backpropagation. ConvNets have been one of the most important aspects of neural networks and deep learning in computer vision for a very long time, as they have several built-in inductive biases, making them significant in a wide range of applications [13].

Visual Transformer (ViT) had an immense impact on the network architecture of the traditionally used ConvNets. ViTs had an added ‘patchify’ layer in its architecture which splits the image into a sequence of flattened patches. Moreover, ViT did not have many differences compared to the NLP-based transformers and also had no inductive bias, which was image-specific. All these, combined with the fact that such models are very scalable, make the transformer-based models stand out from the traditional ones. The following figure 2.4.2 shows the Vision Transformer Architecture.

However, despite its immense success, the ViT model faces a challenge as, with respect to input size, the attention design has a quadratic complexity that makes classification unmanageable for datasets with high-resolution inputs. Hence a hybrid approach was introduced by the Swin Transformer, which incorporates the “sliding window” approach to the Transformer model that makes it outperform the shortcomings of the ViT. Figure 2.5.1 shows the main difference between ViT and Swin Transformer’s sliding Window properties. This makes the Swin Transformers more efficient and reduces the complexity. ViT outperformed the traditional ConvNets on image classification.
but was lacking behind in Object Detection and Semantic Segmentation when compared to accuracy, and much higher resolutions cannot be reached by scaling up the Vision Transformers, and they lacked the simplicity as compared to ConvNets. In addition, even after Swin Transformers overcame the shortcoming of ViTs, however, ConvNets still had most of the properties in a simplistic way as compared to the Swin Transformer. One of the main reasons ConvNets was falling behind was due to the Transformer’s capacity of being very scalable with the multi-head self-attention, and they can perform significantly with large datasets and large models. All these have been pointed out in the paper[49], where ConvNeXt was first proposed to overcome all these shortcomings.

2.5 Reasoning of the models used in DeBERTNeXT

The multimodal model DeBERTNeXT utilizes the power of a Transformer and Convolutional models for text and images for a particular record. This section gives an idea of the
advantages of the models we have used and hence the reason behind utilizing them.

2.5.1 Model for text: DeBERTa

In our model, we use DeBERTa V3 [25], which is an improvement in the DeBERTa model by substituting the mask language modelling (MLM) with a more sample-efficient pre-training task called Replaced Token Detection (RTD), which was proposed by ELECTRA [12]. The efficiency of the DeBERTa model was increased using the ELECTRA-Style pre-training with Gradient Disentangled Embedding Sharing, which ultimately increased the model performance.

Some of the advantages of DeBERTa V3 over the DeBERTa model include the following:

- Deeper Transformer layers: DeBERTa V3 uses deeper Transformer layers, which allows the model to capture more complex relationships between words in a sentence. This leads to improved performance on NLP tasks, such as sentiment analysis and named entity recognition.

- Larger corpus of training data: DeBERTa V3 has been trained on a larger corpus of text data, which allows it to capture the nuances of the language better. This leads to improved performance on NLP tasks, as the model has a better understanding of the language.

- More advanced training techniques: DeBERTa V3 uses more advanced training techniques, such as dynamic masking and enhanced data augmentation, which improves the model’s ability to capture the relationships between words in a sentence.

- Better pre-training: DeBERTa V3 uses the BERT (Bidirectional Encoder Representations from Transformers) pre-training method, which has been shown to be effective in a wide range of NLP tasks. The pre-training stage helps the model better understand the relationships between words in a sentence, leading to improved performance on NLP tasks.
2.5.2 Model for image: ConvNeXT

For the image part, we have used ConvNeXT [49], which is a convolutional model (ConvNet) trained on the large ImageNet dataset, and it has been inspired by the Vision Transformers (ViT) [15]. The authors of ConvNeXT claim to “modernize” a standard ResNet model to the design of a ViT and ultimately outperforms it on the ImageNet dataset and the Swin Transformers [48] on the COCO dataset [46].

2.5.2.1 ConvNeXt

The author of ConvNext first designed using the ResNet 50 and also the much larger ResNet-200. ResNets are Residual Networks that contain smaller Residual Block, which enables to train the deep neural network model by skipping connections. In order to develop ConvNext, the authors change the architecture and training of the standard ResNet50 to match the construction of a Vision Transformer and Swin transformers. This makes the ConvNeXt attain better accuracy, performance and scalability as compared to transformer-based models while also holding the simplicity of the Convolutional Neural Network. The following gives a brief description of how
the traditional ConvNet was changed to the creation of ConvNeXt:

2.5.2.2 Modernization of ConvNets:

1.0 Training Technique: Macro Design

1.1 Compute Ratio: Training the model has a vital role in the performance of the Neural Network. ViTs incorporate new training techniques such as AdamW optimizer, Data Augmentation etc. However, in ConvNext, the number of training epochs was increased from 90 to 300, AdamW optimizer was used instead of the traditional Adam optimizer, and different Data Augmentation techniques were used, which include CutMix, randAugment, Mixup and RandomErasing. Lastly, Regularization Schemes, which as Label Smoothing and Stochastic Depth, were also incorporated, which increased the training accuracy of the traditional ResNet model.

1.2 Stem to Patchify: The “stem cell” signifies the initial network’s processing of the input images. Downsampling is a widely accepted practice for the stem cell in both Convolution Networks and Transformers due to the high level of redundancy in visual data. A standard ResNet’s stem cell has a 7x7 convolution layer with stride 2 and a max pool, which downscales the input images to 4x4. However, in transformers such as the Vision Transformer (ViT), a more aggressive ”patchify” strategy corresponding to a large kernel size, such as kernel size = 14 or 16 and non-overlapping convolution, is utilized as the stem cell. In order to suit the multi-stage architecture, the Swin Transformer utilizes a “patchify” layer with a small patch size of 4. Hence, to modify the ResNet architecture, the “patchify” layer is constructed using a 4x4 stride 4 convolutional layer and is used instead of the ResNet-style stem cell, which ultimately resulted in the improvement of the ConvNext architecture.

1.3 ResNetXt-ify: Similar to ResNet, ResNeXt [91] is a group of Convolutional Neural Network where the convolutional filters are split into different groups through which it can utilize more groups and expand on the width too. On the contrary, ConvNeXt employs depthwise convolution. In depthwise convolution, each filter operates only on a single channel of the input feature map rather than on all channels.
2. LITERATURE REVIEW

This allows for a more computationally efficient operation and can lead to improved performance in certain tasks. The use of grouped convolution, where the number of groups is equal to the number of channels, is a common technique for reducing the computational cost of convolutional neural networks [19].

1.4 Inverted Bottleneck: Transformer models can make use of an "inverted bottleneck" architecture, which refers to a design principle used to reduce the number of parameters in the model while increasing its capacity to handle complex input sequences. The idea behind an inverted bottleneck is to increase the number of hidden units in the lower layers of the network while reducing the number of hidden units in the higher layers. In the case of transformers, the hidden dimension of the Multi-Layer Perceptron (MLP) is four times wider compared to the input dimension.
This allows the model to learn low-level features in the input data while still being able to capture long-range dependencies and higher-level abstractions. The Transformer architecture, as introduced in the 2017 paper "Attention Is All You Need," \[81\] is a prime example of an inverted bottleneck design, as it uses multi-head self-attention mechanisms to allow the model to attend to different parts of the input sequence and build representations from the information it gathers. ConvNeXt also uses an inverted bottleneck design in its architecture.

2.0 Large Kernel Size: The benchmark architectures, such as VGGNet use a 3x3 small kernel size convolutional layer. However, the Swin Transformer uses 7x7 window size, which is larger than the 3x3 kernel size, which is that of a ResNeXt. Similar to the Swin Transformer, ConvNeXt also uses a large kernel size, and a 7x7 depthwise convolution is used in each block.

3.0 Macro Design: The figure 2.5.2 shows the difference between the architecture of the Swin Transformer, ResNet and ConvNeXt.

3.1 Replacing ReLU with GELU: Although ReLU is the most common activation function, however in ConvNeXt, Gaussian Error Linear Unit (GELU) \[28\] is used which is a smoother version of the ReLU is used which is commonly used in Transformer models such as BERT, ViT and GPT-2.

3.2 Fewer Activation Functions: Similar to the one used in the Transformer model and to keep it similar to the ConvNets, all the GELU layers are dropped from the residual block except for the one between two 1x1 layers.

3.3 Normalization Layers: In each residual block of the ResNet mode, batch normalization is used. However, in ConvNeXt, there is only one linear normalization layer before the 1x1 convolution layer, as the other two batch normalization layers are dropped.

3.4 Downsampling Layers: In ResNet, at the beginning of each stage, spatial downsampling is used by the residual block, which uses 3x3 convolution with stride 2. In contrast, in Swin Transformers, between stages, separate downsampling is
used. Combining these two concepts for spatial downsampling, ConvNeXt uses a 2x2 convolution layer with stride 2, which enables better training.

Combining all these features between ResNet and Transformer models, the “hybrid” ConvNeXt outperforms other architectures by several metrics.
CHAPTER 3

Methodology

In this chapter, we will describe the proposed architecture and the models explicitly used for the textual and visual features.

3.1 Proposed Architecture

This section will introduce our proposed framework for multimodal fake news detection. Our framework is used for the classification of news by taking considering two modalities into account: text and image. Since we consider the classification problem a distortion in bias, we consider it a binary problem \[71\]. This enables us to independently classify a piece of news as either real or fake from a given dataset without being domain specific.

3.2 Proposed Methodology:

In this section, we will describe our proposed methodology for this experiment. All three datasets used are in the format of a .csv file where the text is available in a column, and the image is available as URL to the actual news or the post. Hence, a crawler is required to collect and download the images from the .csv dataset. Once that has been done, we get a complete dataset with text, images, labels and the news ID, which is unique to a given news article and consists of both textual and image content. We then pass the image to the ConvXNet model after doing some
pre-processing steps, and similarly, we pass the textual input to the DeBERTa V3 model. Once the output from these two models is attained, we then carry out the concatenation and classification step before finally classifying the news as either true or false. The whole process is further described in the following paragraphs.

3.2.1 Crawler

A URL crawler, also known as a web crawler or spider, is a program that automatically navigates through web pages and retrieves information from the internet. The information retrieved can include text content of web pages, images, videos and other data that is publically accessible. The goal of the URL crawler is to gather data from a large number of web pages in an automated and efficient manner. We first built a crawler to crawl through the URLs provided in the .csv file of the dataset to crawl through the actual new records and to download their corresponding image to the text of that news article.

3.2.2 Input Image:

Once the images are collected, they are then resized and normalized. Normalization of the images refers to the process of scaling the pixel values of an image to a specific range. The goal of normalization is to change the range of the pixel values so that they are more suitable for processing and analysis. It is also an important step as it
reduces the risk of overfitting and helps ensure the input data is consistent. Once the resizing and batch normalization of the images are complete, it is then fed into the ConvXNet large, cased model [49].

The ConvNeXT model was configured with the proposed model for transfer learning. Transfer learning is a machine learning technique where a model trained on one task is fine-tuned or adapted to perform another related task. The idea is to leverage the knowledge learned from the original task to improve performance on the new task. This is especially useful when the new task has limited training data available, as the knowledge learned from the original task can be used as a starting point to reduce the amount of data required to train the new model. For example, consider a deep learning model that has been trained to recognize images of cats and dogs. This model has learned to extract features from images, such as edges, textures, and shapes, that are useful for classifying images into different categories. If we want to use this model to recognize images of birds, we can use transfer learning by fine-tuning the model on a small dataset of bird images. The model will start with the knowledge learned from the original task, such as how to extract features from images, and then adapt to the new task of recognizing birds by learning the unique features that distinguish birds from cats and dogs. In practice, transfer learning is often used in computer vision and natural language processing (NLP) to improve performance on tasks with limited training data.

All the weights of the ConvNeXT model are used except for the last classification layer from the ImageNet pre-trained version. Using weights from transfer learning makes training faster and helps in getting better performance. The last layer of the ConvNeXT model with 1536 dimension output is replaced with a fully connected linear dense layer with 1024 output nodes.

3.2.3 Textual Input:

The textual content is provided in the dataset as a separate column which is used for feature extraction. The textual data is pre-processed, and some pre-processing steps are used, which will be further explained in the experimentation section. Once the
text is pre-processed, the input IDs and Attention Masks are added and then finally fed into the DeBERTa V3 model.

The DeBERTa V3 base model consists of 12 layers and a hidden size of 768. Moreover, it consists of 86M backbone parameters with a vocabulary containing 128K tokens which introduce 98M parameters in the Embedding Layer. This was trained using 160 GB data, similar to the DeBERTa V2 model. The DeBERTa V3 base model takes Input IDs and attention masks as inputs and outputs a 768-dimension-long tensor. For both images and text, a batch size of 16 is used.

### 3.2.4 Concatenation and Classification:

Concatenation refers to the process of combining two or more tensors into a single tensor by joining them along a specific axis. The resulting tensor has a higher dimensionality than the original tensors, as it includes all the values from each of the original tensors. The output from the DeBERTa V3 model is then concatenated along with the ConvNeXT large output along the first axis. The concatenation layer takes 768 dimension output from the DeBERTa V3 model and 1024 dimension output from the ConvNeXT model, producing 1792 dimension output. Finally, a fully connected layer with a sigmoid activation function generates the final classification output. The last layer takes 1792 dimension output from the concatenation layer as input and outputs a value ranging from 0 to 1. Figure 3.2.1 shows our overall model architecture.
CHAPTER 4

Experiments and Dataset

In this chapter, we will first explain the three datasets that we have used during the experimentation step. We will then briefly discuss some of the state-of-the-art models in multimodal detection, which use the same dataset. Moreover, we will also discuss the hyperparameters and fine-tuning our model along with the matrices being used for a proper and fair evaluation of our proposed model, and we will compare those with the existing baseline approaches.

4.1 Datasets

There are several datasets available; however, most are for the unimodal approaches, which are present in either textual or visual content. Given the task of the problem we address, we use three standard benchmark datasets to evaluate our model. We train our model on three datasets that are publicly available, which are the Fakeddit dataset [56], Politifact and Gossipcop dataset. These datasets both contain textual and visual content as images, which is necessary for our model.

4.1.1 Fakeddit Dataset

The Fakeddit dataset [56] is a novel multimodal dataset which contains over one million samples of fake news records from various categories. It has been collected from the reddit portal from 22 different subreddit sections where people usually share
their views and comments about various topics. It contains information from multiple sources, such as text, images, and audio. The authors of the paper have collected and annotated a large number of examples of fake news and real news from various sources and combined them into this benchmark dataset. The dataset provides a comprehensive and diverse collection of examples, which is important for training machine learning models to detect fake news in a wide variety of contexts. Each of the samples was labelled by a 2-way, 3-way and 6-way classification, which was done by distant supervision. Hence, it is one of the most extensive publicly available datasets, which consists of both text and image content for multimodal detection.

4.1.2 FakeNewsNet

FakeNewsNet [72] is a multi-dimensional data repository consisting of two benchmark datasets called Politifact and Gossipcop. In order to build the dataset, the authors first collected the tweets using a crawler and then labelled them from the two fact-checking websites: politifact.com and gossipcop.com.

Fig. 4.1.1: FakeNewsNet data integration process [72]
4. INTERPRETING CNN BASED MALWARE DETECTORS

4.1.2.1 Politifact Dataset

Politifact is a fact-checking website where domain experts and journalists mainly determine if the news pieces are fake or real by evaluating the political news and producing fact-checked evaluation results. Such claims were later labelled by comparing them with the ground truth as either fake or real. The URLs of the web pages which published the news were later used to crawl and collect the news article. However, at times the web pages of the news records were removed, and this problem was tackled by looking if the news content was archived and retrieving it at the Wayback Machine.

4.1.2.2 Gossipcop Dataset

Similarly to Politifact, Gossipcop is a fact-checking website focused on entertainment stories from several media platforms. However, in order to classify the degree of the news from fake to real, the GossipCop provides the output as rating scores which are set to a scale of 1 -10, where scores less than five are classified as fake and more than five are classified as real news. Similar to Politifact, when news reports of some records did not have the URL of the news stories. It was later manually added from either Google surfing or through the Wayback Machine archive.

4.2 Data Processing

All the images are extracted using the crawler with the help of the python libraries: Beautiful Soup and urllib. Once the filtered and refined images were attained, the selected images are reshaped and normalized later during the training phase.

We had to filter through these three datasets from their raw format to remove any unusable or corrupt news records. Some of the dataset did not have textual content, or the textual content was in a different language such was Spanish and other languages. We had to discard those news records completely instead of using online translation platforms as the sentence loses its meaning to some extent in such cases. Additionally, some of the records had corrupt visual URLs provided, or the links were no longer available. In such scenarios, we also discarded those records too.
Finally, when downloading the images from the URLs of the news records, some of the images were corrupt or were GIFs and videos for which the whole news record was also discarded as we require a clean dataset where both the image and the textual content was present.

The URLs are removed for the text since different transformer architectures accept tokens instead of plain English words. For the textual content, the dataset has a clean text column which was already pre-processed to some extent from the raw text. However, for a much cleaner text, we removed the punctuations which included the comma (,) and the full-stop (.). Moreover, we perform stemming, where the prefixes and suffixes are eliminated from a word ending with the stem. For example, words such as follows, following, and followed are stemmed to the word follow. This enables us to make a much more refined and clean text which will help better utilization during training.

For image per-processing, once we filter through the corrupt and unusable images, we only resize it to 224*224 and normalize the image. This ensures that all the images were of the same size for better training.

4.2.1 Dataset Pipeline

In the proposed model, the input and output data are read from the hard drive and passed on to GPU in batches which are done with the help of a data loader pipeline. The data loader pipeline increases loading time significantly as compared to directly loading data to RAM but is necessary as required in case of large datasets when there are RAM constraints.

The dataset pipeline accepts the image file path, the text and the label for a particular record in the dataset and processes it to tensors which are accepted by the model. All the images are first loaded from their file path and then reshaped to the shape of (224, 224, 3). The reshaping is followed by normalizing the images. Once the images are normalized, they are converted to tensors.

All the text inputs are passed through a model-specific tokenizer. The tokenizer is obtained from the HuggingFace library. All the textual content in the record
4. INTERPRETING CNN BASED MALWARE DETECTORS

is tokenized to a max length of 48 words or tokens for Fakeddit Dataset and 32 for Politifact and Gossipcop dataset, and the text beyond the maximum limit is truncated. Special tokens such as [CLS], [SEP] and [PAD] are used. During tokenization, the [CLS] token is the first predefined token added at the beginning. The CLS token output is further used in the model architecture as the output of the DeBERTa V3 base model because the proposed model is responsible for the classification task. The [SEP] token is added at the end, which signifies the end of a sentence. The [PAD] token is added only when the sequence is less than the maximum length. For all the [PAD] tokens, the corresponding attention mask tokens are set to 0.

The tokenizer output is in the form of tensors as input_ids and attention_masks. The input_ids are the tokenizer-processed tokens extracted from text, whereas attention_masks are tensors of 0 and 1 that serve the purpose of allowing the model to use attention on selected tokens. The input_ids and attention_masks are given as output for the text inputs of the DeBerta V3 model. At last, the 2-way_label is also converted to a tensor and given as output to the dataset pipeline.

4.2.2 Setting Hyperparameters

Our model is trained on the Fakeddit, Politifact and Gossipcop datasets. Various experiments are conducted to get the best set of hyperparameters to produce the best performance. The suitable hyperparameters were the same for the Poiltifact and Gossipcop datasets, but slightly differed from the Fakeddit Dataset. The image size are chosen to be 224 pixels in width and 224 pixels in height. The max length for text inputs in the Fakeddit dataset is set to 48 words after analyzing the mean and max sequence length in all the text inputs. For Politifact and Gossipcop, the max length is set to 32 after analyzing in a similar pattern. This difference is due to the size of the text in the different datasets as Fakeddit dataset had a longer sequence of text as compared to the Politifact and Gossipcop where the length of the text was similar. The training data is further split into a train and a validation set by an 80:20 ratio, and the test set created separately is untouched until the training was completed.

The AdamW [50] optimizer, and a linear learning rate scheduler is used for the
training phase. Only Normalization layers and Bias are excluded from weight decay while setting up the AdamW optimizer. For the Fakeddit dataset, a batch size of 16 and the maximum learning rate chosen is $3e^{-6}$ and is scheduled with the help of the scheduler. The scheduler is also customized by adding warm-up steps. The warm-up steps help constrain the pre-trained weights from not exploding during fine-tuning. For our experiment, the warm-up step used is 0.2. The model is trained for 4 epochs and evaluated on the validation set after every epoch. However, for the Politifact and the Gossipcop dataset, the maximum learning rate is chosen to $2.5e^{-6}$ and is trained for 6 epochs. This is found to be the optimum based on several experiments. Finally, once the training is complete, it is tested on the test set, which is 20% of the respective datasets. The experimentation is carried out in the Google Colaboratory platform using Tesla T4 GPU.
CHAPTER 5

Results and Discussion

In this section, we will discuss the results we have obtained after following the steps in the previous section. We will provide the evaluation functions used for a fair comparison. Moreover, we will also discuss and analyze our results by comparing them with the baseline models using the same dataset and also provide the assumptions and limitations of our model.

5.1 Evaluation Functions

There are several assessment criteria in order to assess the efficiency of the algorithm used for the multimodal fake news detection problem, which we will discuss in this section.

After training was completed, the model was evaluated on the test set on various classification and evaluation metrics, including accuracy, precision, recall and f1 score.

The following were noted along with the ROC AUC score in order to calculate the Accuracy, Recall, Precision and F1-Score:

- True Positive (TP): when predicted fake news records are actually annotated as fake news,

- True Negative (TN): when predicted true news records are actually annotated as true news,
5. RESULTS AND DISCUSSION

- False Negative (FN): when predicted true news records are actually annotated as fake news,

- False Positive (FP): when predicted fake news records are actually annotated as true news.

These were used to define the calculation of the following matrices, which were used to compare with other models based on our literature review. The Accuracy (1) calculated is the percentage of correctly predicted news across all samples. The Recall value (2) displays the proportion of false news that is successfully predicted over all false news. The fake news that is accurately predicted from the total predicted news in the fake class is measured using the Precision metric (3). The harmonic average of the recall value and precision value produced for the identification of fake news is known as the "F-measure" (4).

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{2}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{3}
\]

\[
F1 - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}
\]

5.2 Results

With all three of the datasets, we report and highlight the Accuracy, Precision, Recall and F1 score of our proposed model as compared to the benchmarks and the state-of-the-art models out there working on the multimodal issue. For our experimentation, accuracy and F1 score are the important metrics for evaluation and hence we focus on that.
5. RESULTS AND DISCUSSION

Table 5.2.1: Comparison with other models on Fakeddit dataset. (+) indicates that the model does not classify using text and image; (-) indicates that results were not published.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy</th>
<th>Fake News</th>
<th></th>
<th></th>
<th></th>
<th>Real News</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VGG 19 + Text-CNN [69]</td>
<td>0.804</td>
<td>0.838</td>
<td>0.749</td>
<td>0.791</td>
<td>0.704</td>
<td>0.728</td>
<td>0.716</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VQA* [4]</td>
<td>0.631</td>
<td>0.712</td>
<td>0.512</td>
<td>0.596</td>
<td>0.590</td>
<td>0.693</td>
<td>0.637</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NeuralTalk [93]</td>
<td>0.612</td>
<td>0.698</td>
<td>0.610</td>
<td>0.651</td>
<td>0.612</td>
<td>0.712</td>
<td>0.658</td>
<td></td>
<td></td>
</tr>
<tr>
<td>att-RNN [32]</td>
<td>0.745</td>
<td>0.798</td>
<td>0.637</td>
<td>0.708</td>
<td>0.627</td>
<td>0.713</td>
<td>0.667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EANN [88]</td>
<td>0.699</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MVAE [41]</td>
<td>0.784</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNN for text and image [67]</td>
<td>0.870</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeepNet [36]</td>
<td>0.864</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DistilBERT + VGG 16 [38]</td>
<td>0.604</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDGCN+ [39]</td>
<td>0.906</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeBERTNeXT</td>
<td>0.912</td>
<td>0.910</td>
<td>0.950</td>
<td>0.930</td>
<td>0.917</td>
<td>0.854</td>
<td>0.884</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.2.1 Comparison and Discussion: Fakeddit Dataset

Table 5.2.1 depicts the results of the baseline and state-of-the-art models that were trained and tested on the Fakeddit Dataset and compared to the results obtained from our proposed model. We have collected the results depicted in this table from different research papers. We measure and compare the Accuracy, Precision, Recall and F1 score from both the fake and real news as compared with other models for a fair comparison. The authors from Fake News Detection Based on Multi-Modal Classifier Ensemble \[69\] fused the outputs from Text-CNN with VGG-19 to attain an overall accuracy of 0.804. Moreover, they modified and trained the VQA \[4\], NeuralTalk \[93\] and att-RNN \[32\] on the Fakeddit dataset for binary classification. We had trained and tested the EANN \[88\] and MVAE \[41\] on the fakeddit dataset using the same set of hyperparameters as DeBERTNeXT. Hence, we are incorporating the results represented there for a fair comparison. Visual Question Answering VQA\[4\] is a multiclass classification model created to provide natural language answers about a
5. RESULTS AND DISCUSSION

given image. A binary-class layer was put instead of the multi-class layer as the final layer. The LSTM layer was changed to only one layer to enable binary classification. The authors \cite{69} named it as VQA*.

In the paper \cite{32}, the author’s proposed model consists of three main components: a text model, an image model, and a multimodal fusion model.

1. The Text Model: The text model is a recurrent neural network (RNN) that processes the textual content of the microblogs. Specifically, the authors use a Long Short-Term Memory (LSTM) architecture to encode the sequential information in the text. They also use a convolutional neural network (CNN) to extract the textual features. These two models are combined to generate a feature vector that represents the text content.

2. The Image Model: The image model is a convolutional neural network (CNN) that processes the images associated with the microblogs. The authors use a pre-trained VGGNet model to extract the visual features of the images. The VGGNet model is fine-tuned on the rumor detection dataset to learn the features that are relevant for the task.

3. The Multimodal Fusion Model: The multimodal fusion model combines the features extracted by the text and image models to generate a final prediction about the rumor status of the microblog. The authors use a gated multimodal unit (GMU) to fuse the textual and visual features. The GMU is a variant of the gated recurrent unit (GRU) that is designed to fuse information from different modalities. The fused features are then fed into a feedforward neural network to make a binary prediction about the rumor status of the microblog.

This common approach and the algorithm have been used to test in different dataset and modified by different authors accordingly as it is considered as one of the benchmark algorithms for multimodal detection. For our results, we will be using this for comparison with all three of the dataset.

In FakeNED \cite{66}, after the textual and visual features were extracted, a step was added where a single one-dimensional tensor was passed to fully connected layers for binary classification, where it attained an accuracy of 0.878 and an F1 score of 0.910.
5. RESULTS AND DISCUSSION

Compared with FakeNED, our model outperforms the accuracy by 3.80% and achieves an F1 score of 0.912 by weighted average. CNN was used for both text and images in [67] where both the feature vectors attained after the convolutional layer was passed through two dense layers with ReLU non-linear activation before being concatenated. In the end, the logsoftmax function is applied to attain a micro-average accuracy of 0.870, precision of 0.880, recall of 0.870 and F1 score of 0.870. Similarly, for DeepNet [36], which has ReLU as an activation function and softmax function for the final output layer, attains a precision of 0.894, recall of 0.850, an F1 score of 0.872 and an accuracy of 0.864. As compared to the macro-average, our model also outperforms them as we attain an accuracy of 0.912, precision of 0.913, recall of 0.902 and F1 score of 0.917, respectively. Heterogenous Deep Convolutional Network HDGCN [39] is a graph-based neighbour sampling strategy and a hierarchical attention mechanism in order to learn better node embedding. Although it is not a multimodal model, as it only focuses on textual content, we have still used it for comparison as it is a novel approach that includes deep and hierarchical attention strategies. It achieves an accuracy of 0.906 and an F1 score of 0.885, in which our model also outperforms it.

The figure 5.2.1 shows the train and validation accuracy obtained at each epoch for four epochs. Moreover, figure 5.2.1 depicts the loss plot at each epoch of our model for four epochs.
5. RESULTS AND DISCUSSION

Table 5.2.2: Comparison with other models on Politifact and Gossipcop Dataset. (-) indicates that the results were not published

<table>
<thead>
<tr>
<th>Models</th>
<th>Dataset: Politifact</th>
<th>Dataset: Gossipcop</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Precision</td>
</tr>
<tr>
<td>SpotFake+ [76]</td>
<td>0.846</td>
<td>-</td>
</tr>
<tr>
<td>SAFE [98]</td>
<td>0.874</td>
<td>0.889</td>
</tr>
<tr>
<td>Cross-Domain Detection [73]</td>
<td>0.840</td>
<td>0.836</td>
</tr>
<tr>
<td>att-RNN [32]</td>
<td>0.769</td>
<td>0.735</td>
</tr>
<tr>
<td>CMC [89]</td>
<td>0.894</td>
<td>-</td>
</tr>
<tr>
<td>EANN [88]</td>
<td>0.740</td>
<td>-</td>
</tr>
<tr>
<td>MVAE [41]</td>
<td>0.673</td>
<td>-</td>
</tr>
<tr>
<td>SpotFake [75]</td>
<td>0.721</td>
<td>-</td>
</tr>
<tr>
<td>SceneFND [95]</td>
<td>0.832</td>
<td>-</td>
</tr>
<tr>
<td>DeBERTNeXT</td>
<td>0.913</td>
<td>0.921</td>
</tr>
</tbody>
</table>

5.2.2 Comparison and Discussion: FakeNewsNet Dataset

We have also experimented on two much smaller but widely used benchmark datasets, namely Politifact and Gossipcop. For a fair evaluation, we have also incorporated the matrices commonly used by other authors for a fair comparison. In this scenario, we have used the weighted average of our results for comparison. Table 5.2.2 shows the results that are collected from different research papers, it can be seen that our model surpasses other baseline and state-of-the-art models by a good margin.

Most of the multimodal models use transfer learning to improve the accuracy of detection. SpotFake+ [76], SpotFake [75], CMC [89] uses VGG19 for image content and a transformer-based model such as XLNet [92] and BERT for the textual content. SpotFake+ consists of the Pre-trained XLNet model for the textual features and VGG19 for the visual features. After these are passed through the model, they then pass through several dense layers, batch normalization and dropout layers before being combined together. The combined multimodal feature vector then passes through a series of dense layer before coming to the final classification layer. XLNet has a potential for bias [42] and other disadvantages such as more computational cost and longer training time. It can be seen that SpotFake+ have attained an accuracy of
5. RESULTS AND DISCUSSION

Fig. 5.2.3: Accuracy plot on Politifact

![Accuracy Plot](image)

Fig. 5.2.4: Loss plot on Politifact

0.881 for the politifact dataset and 0.836 on the gossipcop dataset which is less as compared to our model. This can be because DeBERTa has some notable advantages over the XLNet model for the textual representations:

1. Higher Efficiency: DeBERTa has been designed to be more efficient than XLNet, which makes it faster to train and use in practice. DeBERTa uses dynamic masking to reduce the number of calculations required during training, which leads to faster convergence and higher accuracy.

2. Better Performance on Downstream Tasks: DeBERTa has been shown to outperform XLNet on several benchmark NLP tasks, including the General Language Understanding Evaluation (GLUE) [85] benchmark and the SuperGLUE benchmark [86]. In particular, DeBERTa has achieved better results on several of these tasks, including sentiment analysis, natural language inference, and question answering.

3. Improved Masking Strategy: DeBERTa uses a more sophisticated masking strategy than BERT and XLNet. DeBERTa uses a combination of static masking and dynamic masking to expose the model to a wider range of training examples, which improves the model’s ability to handle different types of input data.

4. Pre-training Techniques: DeBERTa uses a number of pre-training techniques that help the model to learn more effectively from the available data. For example, DeBERTa uses a training technique called ”span masking,” which involves masking entire spans of text rather than just individual tokens. This enables the model to
5. RESULTS AND DISCUSSION

5. Flexibility in Model Size: DeBERTa allows for more flexibility in model size, which makes it easier to fine-tune for specific tasks. By varying the number of layers in the model, DeBERTa can be adapted to different computational constraints or task requirements.

Similar to the textual aspect, our model has some notable advantages as it uses ConvXNet over the commonly incorporated VGG 19 used in SpotFake, CMC and Spotfake+ for the image content, which includes:

1. Improved Accuracy: ConvXNet was designed to improve on the accuracy of traditional models. In fact, it has been shown to achieve higher accuracy on certain image classification tasks, such as recognizing small and medium-sized objects in complex scenes.

2. Faster Inference Time: ConvXNet has a smaller number of parameters than VGG19, making it more efficient in terms of computation time and memory usage. This means that ConvXNet can perform inference on images faster than VGG19.

3. More Scalable: ConvXNet is more scalable than VGG19, meaning that it can be easily modified to handle different image sizes or types. This makes ConvXNet more versatile and useful for a wider range of image classification tasks.

4. Improved Regularization: ConvXNet uses a combination of L2 and L1 regularization techniques to prevent overfitting, which can result in better generalization performance and more robust models.
5. RESULTS AND DISCUSSION

Table 5.2.3: Comparison of the accuracy of models ran on the same environment for Politifact and Gossipcop Datasets

<table>
<thead>
<tr>
<th>Models</th>
<th>Politifact</th>
<th>Gossipcop</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpotFake</td>
<td>0.697</td>
<td>0.783</td>
</tr>
<tr>
<td>SpotFake+</td>
<td>0.881</td>
<td>0.836</td>
</tr>
<tr>
<td>EANN</td>
<td>0.712</td>
<td>0.857</td>
</tr>
<tr>
<td>MVAE</td>
<td>0.701</td>
<td>0.815</td>
</tr>
<tr>
<td>DeBERTNeXT</td>
<td><strong>0.913</strong></td>
<td><strong>0.902</strong></td>
</tr>
</tbody>
</table>

The authors of SpotFake+ have trained the SpotFake, MVAE and EANN models too on the politifact and gossipcop dataset. We have used those results for a comparison with our model and it can be seen that our model outperforms all of them in terms of accuracy.

SAFE (Similarity-Aware Multi-Modal Fake News Detection) [98] model uses Text-CNN architecture for both image and textual content. In order to attain the visual features, an additional fully connected layer is used to process the visual features. In SAFE methodology, the text and images in each article are preprocessed to extract the features and are then passed to the Modality-specific classification, where each modality is classified separately using a classification model. Then the similarity between the text and images in each article is calculated using cosine similarity. Finally, multi-modal classification is done where the modality-specific classification and the similarity score are combined using a multi-modal classification model that assigns a final score to each news article labelled as either real or fake. Although Text-CNN is a powerful algorithm for textual content, however, transformer models like DeBERTa use a bidirectional attention mechanism, which allows it to capture contextual information from both the left and right contexts of a given word in a sentence. This helps DeBERTa to better understand the relationships between words in a sentence and to make more accurate predictions. SAFE’s methodology surpasses our model’s recall result for the Gossipcop dataset; however, our model
5. RESULTS AND DISCUSSION

Table 5.2.4: Comparison of the accuracy of the models ran on the same environment for Fakeddit Dataset

<table>
<thead>
<tr>
<th>Models</th>
<th>Fakeddit</th>
</tr>
</thead>
<tbody>
<tr>
<td>EANN</td>
<td>0.637</td>
</tr>
<tr>
<td>MVAE</td>
<td>0.772</td>
</tr>
<tr>
<td>DeBERTNeXT</td>
<td>0.913</td>
</tr>
</tbody>
</table>

still outperforms it in terms of all other matrices. The authors of SAFE trained and tested the att-RNN model on the same dataset which we also incorporated for the comparison. Figure 5.2.3, 5.2.4, 5.2.5 and 5.2.6 shows the accuracy and loss plot on the Politifact and Gossipcop dataset.

5.2.3 Validation

In order to validate our result, we ran some of the algorithms in the same environment where we used the same hyperparameters as DeBERTNeXT during experimentation with the respective datasets. SpotFake, SpotFake+, EANN and MVAE were the codes of the algorithms which were publicly available. Hence, we have used those to validate our model in those datasets. It can be seen from table 5.2.3 and 5.2.4 that our model still outperforms other models by a good margin. SpotFake+ has higher accuracy than other models being compared as it uses XLNet and has no max length limitations which are required to be set as in SpotFake, which uses BERT. In BERT, we require to set the max length of the textual input, which was changed to 32 as per our experimental setup. It can also be seen that the EANN performs better in Politifact and Gossipcop datasets but its performance decreases in Fakeddit dataset. One of the reasons can be that EANN was originally trained and tested on Twitter and Weibo datasets by the authors which have a similar textual length as the PolitiFact and Gossipcop dataset. However, the Fakeddit dataset is composed of longer sentences in each news article, which suggests that EANN may perform better in smaller text-length data. In comparison, our model performs better both in the
smaller and bigger datasets.
CHAPTER 6

Conclusion and Future Work

6.1 Conclusions

Social media is becoming increasingly popular with the advancement and ease of access to technology. The number of people who are consuming news from online sources rather than traditional news media is also increasing due to the same reason, which creates a big scope for the spread of fake news. Consequently, social media is one of the biggest platforms for the spread of fake news, which has significant short and long-term negative impacts on individual users and also the whole society. There are several unimodal approaches in the detection of such fake news; however, statistics have shown that people are more likely to believe in false information that has some visual content associated with it or is multimodal. This makes it easier for fake news to spread on social media, as fake news creators can add multimedia elements to their content to make it appear more legitimate and appealing. While there are several multimodal approaches to preventing fake news, these methods either lack accuracy, are domain-specific or are trained on either very large or small datasets. This means that they are ineffective in detecting all types of fake news, especially those that are highly sophisticated with visual content and targeted toward a specific group of people.

Hence, to tackle all these problems, we introduce our framework, which is a fusion of DeBERTa model and the ConvNeXT model, and both of these are state-of-the-art in their field as they overcome the obstacles of the previous benchmark models. Our
model was trained and tested on both the immense benchmark dataset and also on two other smaller datasets where it outperformed the current models by about 3.80% on Fakeddit dataset, 2.10% on Politifact and 1.00% on Gossipcop dataset.

6.2 Limitations and Future Work

We intend to broaden our investigation into fake news detection by expanding our research to include various modalities, such as video and audio, which are emerging sources of fake news. Currently, our model is only suited for images and textual content and is unable to process any other modalities such as video, audio, etc. By studying the characteristics of video and audio-based news content, we hope to develop advanced multi-modal models to detect fake news from such sources. We also aim to study the type of news that attracts users based on age, sex, demographics, and other factors. This analysis will involve identifying patterns in the way different users engage with news content and developing strategies to detect fake news based on these patterns. This could lead to the development of more efficient and targeted algorithms to identify fake news and reduce its spread. Moreover, we plan to develop an application to classify news stories from social media platforms in real-time instantly. This could significantly improve the detection of fake news as it can be identified before it reaches a larger audience. Additionally, the application could help to segment news content that is attractive to different user categories and contribute to the development of a more diverse detection framework. We aim to explore these research problems to help users identify fake news and prevent its spread. By combining these different research areas, we hope to contribute to developing a more advanced detection framework that can identify fake news from other modalities and in real time, reducing the spread of fake news and limiting its impact.
REFERENCES


VITA AUCTORIS

NAME: Kamonashish Saha
PLACE OF BIRTH: Dhaka, Bangladesh
YEAR OF BIRTH: 1995
EDUCATION:

BRAC University, B.Sc in Computer Science, Dhaka, Bangladesh, 2019

University of Windsor, M.Sc in Computer Science, Windsor, Ontario, 2023