

AI Against Hate: Multimodal Detection of Islamophobic Content with Deep Learning



Author

Niha Latif

Registration No: 00000400606

Supervisor

Prof Dr. Naima Iltaf

A thesis submitted to the faculty of Computer Software Engineering Department,
Military College of Signals, National University of Sciences and Technology, Islamabad
Pakistan in partial fulfillment of the requirements for the degree of MS in
Software Engineering

(September 2024)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Ms Niha Latif, Registration No. 00000400606, of Military College of Signals has been vetted by undersigned, found complete in all respect as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial, fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the student have been also incorporated in the said thesis.

Signature: Naima

Name of Supervisor: Prof Dr. Naima Iltaf

Date: 01/10/24

Signature (HoD): [Signature]
~~Head of Dept of CSE
MCS College of Sigs (NUST)~~

Date: 02/10/24

Signature (Dean/Principal): [Signature]

Date: 26/9/24

**Brig
Dean, MCS (NUST)
(Asst Masood, PhD)**

NATIONAL UNIVERSITY OF SCIENCES & TECHNOLOGY
MASTER THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by Niha Latif, Regn No 00000400606 Titled: "AI Against Hate: Multimodal Detection of Islamophobic Content with Deep Learning" be accepted in partial fulfillment of the requirements for the award of MS Software Engineering degree.

Examination Committee Members

1. Name: Assoc Prof Dr. Ihtesham Ul Islam

Signature: 

2. Name: Lt Col Khawir Mahmood

Signature: 

Supervisor's Name: Prof Dr. Naima Iltaf

Signature: 


Date: 01/10/24


 Head of Dept of CSE
 Coll College of Eng (NUST)
 Head of Department

21/10/24
 Date

COUNTERSIGNED

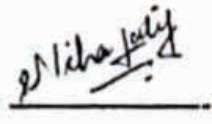
Date: 26/9/24


 Brfg
 Dean, MCS (NUST)
 (Asif Masood, PhD)
 Dean

CERTIFICATE OF APPROVAL

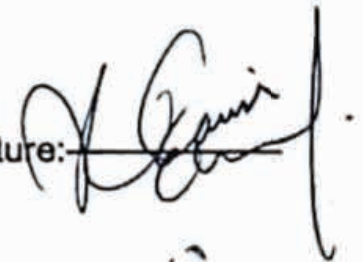
This is to certify that the research work presented in this thesis, entitled "AI Against Hate: Multimodal Detection of Islamophobic Content with Deep Learning." was conducted by Mr. Niha Latif under the supervision of Prof Dr. Naima Iltaf. No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the Department of Computer Software Engineering in partial fulfillment of the requirements for the degree of Master of Science in Field of Computer Software Engineering Department of Military College of Signals, National University of Sciences and Technology, Islamabad.

Student Name: Niha Latif

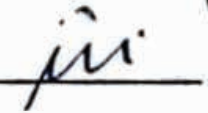
Signature: 

Examination Committee:


a) External Examiner 1: Lt Col Khawir Mehmood
(Department of Computer Software Engineering)

Signature: 

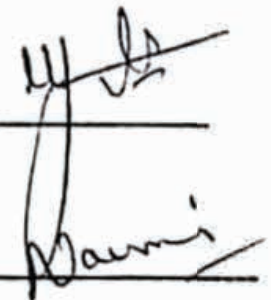
b) External Examiner 2: Assoc Prof Dr. Ihtesham Ul Islam
(Department of Computer Software Engineering)

Signature: 

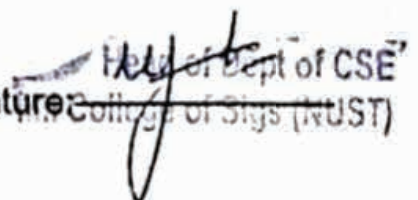
Name of Co-Supervisor: Brig Adnan Ahmed Khan, PhD

Signature: 

Name of Supervisor: Prof Dr. Naima Iltaf

Signature: 

Name of Dean/HOD: Brig Adnan Ahmed Khan, PhD

Signature:  Head of Dept of CSE
Military College of Signals (MUST)

AUTHOR'S DECLARATION

I Niha Latif hereby state that my MS thesis titled “**AI Against Hate: Multimodal Detection of Islamophobic Content with Deep Learning**” is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Student Signature: _____



Name: _____ Niha Latif _____

Date: _____ 05/08/2024 _____

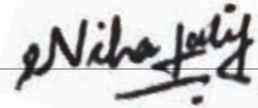
PLAGIARISM UNDERTAKING

I solemnly declare that the research work presented in the thesis titled “**AI Against Hate: Multimodal Detection of Islamophobic Content with Deep Learning**” is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and National University of Sciences and Technology (NUST), Islamabad towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the University reserves the rights to withdraw/revoke my MS degree and that HEC and NUST, Islamabad has the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized thesis.

Student Signature: _____



Name: _____ Niha Latif _____

Date: _____ 05/08/2024 _____

DEDICATION

“In the name of Allah, the most Beneficent, the most Merciful”

Glory be to Allah Almighty, the Creator and Sustainer of the Universe, the Omnipotent and the Omnipresent. There is nothing I could have accomplished without His guidance and blessings. I dedicate this thesis to my family, friends, and teachers, particularly my parents, who supported me each step of the way.

ACKNOWLEDGEMENTS

In the name of Allah (S.W.A), the Creator and Sustainer of the Universe, to whom belongs all glory and power. He alone has the authority to elevate and humble individuals as He pleases. Truly, nothing can be accomplished without His will. From the moment I stepped foot into NUST until the day of my departure, it was by His divine blessings and guidance that I was able to navigate the path of success. His unwavering support and the opportunities He bestowed upon me were instrumental in completing my research journey.

I would also like to express my heartfelt appreciation to my thesis Supervisors and Co-Supervisor, Prof Dr.Naima Iltaf, PHD, Prof Hammad Afzal, and Brig Adnan Ahmed Khan, PhD for their unwavering support and guidance throughout my thesis. Their knowledge, expertise, and dedication to their field have been a source of inspiration to me, and I am grateful for the time and effort they invested in my success. Whenever I encountered any difficulties, they were always available to offer their assistance and provide me with insightful feedback.

In addition, I extend my gratitude to my GEC members, Assoc Prof Dr. Ihtesham Ul Islam and Lt Col Khawir Mahmood, for their continuous availability for assistance and support throughout my degree, both in coursework and thesis. His expertise and knowledge have been invaluable to me, and I am grateful for his unwavering support and guidance.

Lastly, all praises and thanks be to Allah (S.W.A), the Most Merciful and the Most Gracious.

Abstract

Living in a world of constant connectivity with others through electronic devices, Islamophobia has become a very critical issue, especially in social network sites. In contrast to regular hate speech, which is most often expressed in words, Islamophobia in the internet world can be expressed in pictures, text, videos, and audio and therefore, is much more complex to trace. Conventional machine learning techniques cannot be used for their classification since they tend to lack context in detecting hate speech. Therefore, researchers have shifted to deep learning techniques. Previously developed deep learning approaches are based on unimodal architectures that classify either textual or visual data, thereby not considering the overall context of data including both visual and textual. This research aims to fill the gap that currently exists in the identification and categorization of Islamophobic memes. We have proposed a multimodal technique that integrates deep learning models for the classification of Islamophobic content from both, textual and visual information. BERT and ResNet-50 models are used for text and image classification respectively. The evaluation results demonstrate that the proposed multimodal approach accurately identifies Islamophobic content with an overall accuracy score of 95% and cross-entropy loss of 15%.

Keywords: Islamophobic detection, Islamophobia, Non-Islamophobia, deep learning, multimodal, hate.

Contents

ABSTRACT	IX
LIST OF TABLES	XIII
LIST OF FIGURES	XIV
LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS	XV
1 Introduction	1
1.1 Motivation And Problem Statement	5
1.2 Research Objectives	7
1.2.1 Research Contribution	7
1.3 Thesis Outline	8
2 Related Work	10
2.1 Introduction	10
2.2 Summary	24
3 Preliminaries	25
3.1 Overview:	25
3.2 Overview of Models Applied in Our Methodology	25
3.2.1 ResNet-50 Model:	26

3.2.2	BERT Model:	28
3.2.2.1	General Architecture of BERT:	29
3.3	Multimodal Data and Proposed Models	31
3.3.1	ResNet-50 Model:	31
3.3.1.1	Benefits of ResNet-50 for Multimodal Image Classification	35
3.3.2	BERT Model:	36
3.3.2.1	Advantages of BERT for Text Classification for Multi- modal Data:	39
3.4	Summary	42
4	Proposed Framework	44
4.1	Overview	44
4.1.1	Data Splitting and Feature Extraction:	45
4.1.2	Model Training:	46
4.1.2.1	Training Parameters	47
4.2	Framework Architecture	51
4.2.1	<i>Visual Feature Extraction</i>	52
4.2.2	<i>Textual Feature Extraction</i>	53
4.2.3	<i>Unified Visual and Textual Framework</i>	53
4.2.4	<i>Mathematical Illusion</i>	54
4.3	Summary:	56
5	Dataset and Model Selection	58
5.1	Data Collection	58
5.2	Data Augmentation:	59
5.3	Data Annotation:	61
5.4	Data Extraction and Labeling:	63
5.4.1	Summary:	63

6	Evaluation and Results	65
6.1	Overview	65
6.2	Evaluation Metrics	65
6.3	Results and Discussion:	67
6.4	Summary	71
7	Conclusion and Future Work	74
7.1	Overview	74
7.2	Limitations:	75
7.3	Future Work:	76
7.4	Summary	77

List of Tables

4.1	Data Splitting	46
5.1	Dataset Collection	61
5.2	Percentage Label of Data	62
6.1	Model Performance Comparison	70
6.2	Multimodel Accuracy	70

List of Figures

1.1	Islamophobic Content	5
1.2	Non-Islamophobic Content	6
3.1	Structure Diagram by ResNet-50	28
3.2	General BERT Model Architecture	30
3.3	General ResNet-50 Architecture	31
4.1	Proposed Architecture	52
5.1	Dataset Search on different Platforms Using Specific Keywords	60
5.2	Dataset Collection	63
6.1	Comparison Chart	71
6.2	Cross-Entropy Loss	72
6.3	Accuracy	72

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

ML Machine Learning

AI Artificial Intelligence

NLP Natural Language Processing

LSTM Long Short Term Memory

BERT Bidirectional Encoder Representation from Transformers

RNN Recurrent Neural Network

CNN Convolutional Neural Network

Bi-LSTM Bidirectional Long Short Term Memory

ResNet Residual Network

LLMs Large Language Models

GCN Graph Convolutional Network

LMMs Large Multimodal Models

GRU Gated Recurrent Unit

GCR-NN Graph Convolutional Recurrent Neural Network

SVM Support Vector Machine

LR Logistic Regression

RoBERTa Robustly Optimized Bidirectional Encoder Representation from Transformers

GPT Generative Pre-trained Transformer

PLMs Pre-trained Language Models

EfficientNet Efficient Network

CLIP Contrastive Language Image Pre-training

AUROC Area Under the Receiver Operating Characteristic Curve

SupCon ResNet Supervised Contrastive ResNet

IDS Intrusion Detection System

CC Conceptual Captions

ResNeXt Residual Network with aggregated Transformations

API Application Programming interface

AUC Area Under the Curve

HMC Hateful Memes Challenge

CROM-AE Cross Modality Auto Encoder

C-BLSTM Contrastive-Bidirectional Long Short Term Memory

R-CNN Region based Convolutional Neural Network

SOTA State-Of-The-Art

HatRed Hateful Memes with Reasons Dataset

Chapter 1

Introduction

Fear, prejudice, or discrimination against Islam and Muslims, generally known as Islamophobia is gradually emerging as a serious issue in the increasingly connected global village. Since most societal discussions occur on social media platforms, it becomes important to research how hate speech is exhibited on online platforms [1]. Prejudice is present in most societies in the modern world today. It has always been a social issue in most societies starting with the societies in the United States of America [2]. Fluency and growth in the Internet through portable devices and other low-priced gadgets have recently led to social and electronic media users' emergence. A negativity that came with this growth is the existence of clashes, hate speech, cyberbullying, and trolling. In this domain, a lot of various research has been conducted on factors of hate speech which include; gender, racism, religion, color, disability, and citizenship [3]. Earlier, any of these services could only be attained through assistance from friends, colleagues, or relatives, and today, we rely on digital media platforms to connect with them. How-

ever, as has been realized, it is getting hard to avoid the influence of these channels' negativity and at the same time deny ourselves the services of these various threatening channels [4]. Among these, prejudice and discrimination against Islam and Muslims popularly known as Islamophobia has recently become a more pronounced problem being evidenced in various discursive formations of cyberspace, racy or otherwise. It has also given the ground for the emergence of negative information such as racism, fake news, and discriminating information.

Previously, The 9/11 terrorist attacks brought about a significant change in the Western imagination and perception of Islam: They regard Muslim men as the lowest form of human beings and since 9/11, they are even less than zero, "Animal-like, stripped of all legal rights allowed under US domestic and International laws and force feed like animals" [5, 6]. Most recently, the UK has recorded an increased number of Islamophobic attacks in the course of far-right riots that have affected many cities including London, Liverpool, and Manchester. The violence erupted in the wake of fake news circulating on social media, that blamed the Muslims for a stabbing. This fake news made people furious and incited them to attack Muslim neighbors; even mosques were not spared by the rioters.

The riots established that Britain has an inherent problem of Islamophobia, which extremists and some media personalities have worsened. They have made the Muslims in the UK experience fear and this has made them tighten up on security matters in

mosques and other places of worship. These cases illustrate the contemporary social problem of countering Islamophobia and the risks related to the uncontrolled dissemination of hatred on the Web.

This problem can only be solved with solid and complex elaborate instruments designed to analyze the heterogeneity of ways in which Islamophobic content is disseminated. The present study deals with the recognition of both, Islamophobic and non-Islamophobic tweets and memes, as it is critically important to take into account the fact that hate speech is often multimedia. As models capable of distinguishing between toxic and non-toxic content for various media types are proposed in this investigation, this research in this study intends to make a modest contribution to the global initiatives to combat the dissemination of Islamophobia.

As deep learning progresses in identifying hate speech on social media, the research in the direction of identifying Islamophobic hate speech is limited. It is worth mentioning that recently a lot of attention has been paid to the frequency and dissemination of Islamophobic content on social networks. This material can also be presented in words, pictures, video, and music and could include; messages encouraging racial or religious violence, direct websites with photos, videos, or descriptions describing violence to anyone based on the color of their skin, their views on God, their mental or physical disability, their sex or that they are a trans-sexual and chat rooms in which people are asking others to engage in hate crimes [7]. More and more far-right organi-

zations have been using disinformation and other techniques to demonize Muslims and their religion. The purpose of this study is to design an effective and efficient means of identifying and classifying Hate memes against Islam and memes that propagate extremism towards Muslims, also this research contributes to improving the conditions of the hashtag environment and its impact on the promotion of diversity and reduction of hazards of hate speech.

To identify between Islamophobic and non-Islamophobic content using deep learning, we can therefore use many models to analyze both text and images since these are the two main ways. For decision-making and especially for text classification, one can fine-tune the Bidirectional Encoder Representation from Transformers (BERT) or Generative Pre-trained Transformer (GPT) models for the detection of the so-called Hate Speech, Discriminatory Phrases, and Context that represents Islamophobia. These models have been trained on voluminous data about text, making them capable of perceiving tender discrimination forms that can be masked by ordinary language. In image classification, models such as Residual Network (ResNet) or Efficient Network (EfficientNet) can be used to find visual signs of Islamophobia in images, including discriminatory signs, memes, or hate-inciting images. These models may be trained with labeled imagery datasets to identify Islamophobia through the patterns that it exhibits. In this way, the proposed multimodal deep learning system is capable of recognizing Islamophobic and non-Islamophobic content coming from the web. It is



Figure 1.1: Islamophobic Content

double protection because apart from employing text and image data the model helps to detect dangerous content for all kinds of media.

Here are some example images to help illustrate the concept of Islamophobic and non-Islamophobic images as shown in Figure 1.1 and 1.2.

1.1 Motivation And Problem Statement

The main reason that led to the motivation of developing AI for identifying the multimodal Islamophobic content through deep learning is the inherent realization of the need to fight the hate speech in social media that leads to violence and social fragmentation. In that sense, this research intends to propose more efficient tools that analyze not only the text information but also photos to prevent the spreading of Islamophobic



Figure 1.2: Non-Islamophobic Content

content which could remain unnoticed in a commonly used approach. This not only increases the chance of preserving endangered populations but also increases the creation of a safe environment on the internet. The study represents the ethical imperative to use AI technology to make positive social changes so that everyone is well-represented within the digital world.

Recent unimodal approaches for detecting Islamophobic content are very limited, the main focus is only on text and visuals separately. This insufficient scope ignores the significance of considering both textual and visual data. A more detailed approach is needed. Recently available schemes will not classify content accurately, ignoring the combination of text and images. To mark this issue, a multimodal framework is needed. This framework should work on both textual and visual features from the input

dataset and extract and analyze these features. As a result, it will create classifications, improving accuracy to detect Islamophobic content. By using a unified approach, we can achieve a more refined understanding of Islamophobic content. However, a unified framework will help identify all patterns and features that may be missed by unimodal approaches. It can lead to more efficient content moderation and a safer online environment. Moreover, the enhancement of a multimodal framework can give advanced deep learning models in our field. By surveying the intersection of textual and visual data, we can make a modified model better for online content's complexities.

1.2 Research Objectives

- Introduce a multimodal approach using new and in-trend models to achieve state-of-the-art results.
- Generate a corpus of extremely negative comments on Islam and Muslims in social media with a focus on memes and specific #hashtags and anti-Muslim slogans.

1.2.1 Research Contribution

The following objectives were achieved during the research process:

- Implemented Multimodal approach using modern and efficient deep learning models and achieved state-of-the-art results.

- Generated our memes dataset by implementing Hashtags and anti-Muslim keywords collected from social media including Facebook, Instagram, and Twitter.

1.3 Thesis Outline

Chapter 01: Introduction – This chapter discusses the phenomenon of social memes with a focus on their definition, effects on society, and the challenges due to the object’s multimodality. It ends by highlighting the need to build a framework to address such challenges.

The research objectives are presented as follows;

Chapter 02: Literature Review – This section discusses the research work done in the field of multimodal content classification.

Chapter 03: Preliminaries– This chapter focuses on the application of classifiers from the deep learning approach on multi-modal data and also, gives an insight into the classifiers used in this approach.

Chapter 04: Proposed Framework – This part of the research provides a clear description of the deep learning classifiers that are used in the study together with a description of how each of the classifiers is implemented in the study.

Chapter 05: Dataset Creation – This chapter emphasizes the collection and augmentation of data that are most suitable for the detection of Islamophobic memes. For this reason, it dwells upon the criteria for dataset collection and annotation and

introduces the concerned study hypotheses.

Chapter 06: Evaluation and Results – This chapter provides the details of the experiments performed using deep learning classifiers and makes a comparison of different models to achieve the results.

Chapter 07: Conclusion and Future Work – This chapter concludes our research, lists its limitations, and draws light on future work.

Chapter 2

Related Work

2.1 Introduction

Especially in the last few years, the availability of social platforms caused an even larger boost in the use of internet memes. These memes as a rule are often humorous or sarcastic but can also be a way of disseminating stereotypes and hate speech. Islamophobic memes, usually, are one of the most trending forms of material that contribute to the proliferation of hate speech in social media. In response, researchers have come up with the various ways through which one has to identify and categorize the content that has to be banned. Prior methods have chiefly relied on literary analysis with the use of Natural Language Processing (NLP) tools to categorize the discovered items as hate speech. However, due to the combination of the graphical and textual components of internet memes as well as sometimes even the integration of the video aspect, these traditional approaches are not without their difficulties.

In response to these issues, recent research has analyzed the application of deep

learning models that can handle such data inputs. These models combine the graphics with textual data to enhance efficiency in conveying the information. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are the most popular types of deep learning that are used for image and textual data respectively. Some of the approaches have even mixed these architectures to arrive at the true hybrid architectures capable of handling the complexity of the multimodal content. Nevertheless, the advancement of methods from Natural Language Processing (NLP) has not yet been applied to fill this gap in the literature of developing a method to detect Islamophobic content within such a setting. Current conventional approaches also failed to provide a means of distinguishing Islamic and non-Islamic hate speech, thus a stronger approach is needed. It further extends from these endeavors by introducing this new hybrid deep learning model to flag this gap and improve the detection and classification of Islamophobic content on internet memes.

In the course of the present study, we engaged in a critical appraisal of the literature where we reviewed several papers relevant to the development of a strong concept for this research. Such sources have been very helpful in informing viewpoints and strategies required to identify and categorize content using deep learning models.

The authors discuss a novel approach ‘Multimodal Hate Speech Detection in Memes Using Contrastive Language-Image Pre-Training’ which aims to examine the pressing problem of hate speech on social media platforms, that uses both text and image to

spread hatred. The study also presents a new method using the ‘Facebook Hateful Meme Dataset’ fine-tuning the Contrastive Language-Image Pre-Training (CLIP) model through prompt engineering to improve the detection of hateful memes using both vision and language processing. The study also focuses on CLIP’s utility in distilling, the interaction of textual and visual components of memes, and further the versatility of the model to do zero-shot forecast that lets it make good predictions on implicit data. The proposed method achieved an accuracy rate of 87.42% and the AUROC (Area Under the Receiver Operating Curve) was 88.35%, which emphasizes the detection of multi-modal hate speech [8]. Another study deals with the problem of recognizing hateful memes [9], which develop through the blending of undesirable meanings in different cultural references. To collect the data, the authors randomly sampled 1 million image/text pairs from 4chan board/pol/. The authors put forward a framework based on multimodal contrastive learning models and specifically OpenAI’s Contrastive Language Image Pre-training (CLIP) for recognizing targets of hateful content and even studying the dynamics of such memes, if any, systematically. Although the framework can identify new variants of hateful memes for manual verification due to proper semantic relationship extraction within and across the text and image modalities, it helps in modulating the modulators to prevent the spread of evil content in the virtual space.

In [10], authors have surveyed online hate speech moderation employing a versatile

set of publicly available datasets from which cross-sectional representation across text, image, video, and audio-based hate speeches is searched. Hate speech detection and moderation are seen as crucial areas of application for Large Language Models (LLMs) and Large Multimodal Models (LMMs), given the multi-modal nature of digital content. Precisely, three machine learning models: Adaboost, Naive Bayes, and Random Forest were employed for the classification of hate speech from audio, and the Deep Learning approach that makes use of Convolutional Neural Networks (CNNs) and self-attentive CNNs for extracting the features from the audio data. The use of the above models shows that they can adequately handle the complexity of contemporary hate speech, especially in multimodal hate speech.

A new deep learning approach was introduced for the in-vehicle Intrusion Detection System (IDS) [11], in which the supervised contrast (SupCon) ResNet was used and uses transfer learning. The study seeks to eliminate the issues that have been mentioned with the current IDS approaches, mainly, the IDS approach works in binary classification and requires a large data set that can be specific to the vehicle. The proposed Supervised Contrastive (SupCon) ResNet model is evaluated using two real car datasets: The first dataset is the Car Hacking dataset and the second one is the Survival dataset. The methodology uses contrastive learning to better classify an attack and transfer learning to apply the model to different car models. The results show that the proposed SupCon ResNet model decreases false negative rates and has

high F1 measures on different vehicle models; the F1 measure of 0.9989 and 0.9979. It achieves 0.9979 on the KIA Soul and 0.9975 on the Chevrolet Spark datasets, which makes it suitable for practical Intrusion Detection System (IDS) implementation.

Another study focuses on the problems of classifying social media memes containing religious hatred using a multimodal approach. The authors generated a completely new dataset which was determined to contain more than 2000 religiously hateful memes and then it was merged with the Facebook Hateful Memes dataset. They used Visual-BERT, a BERT-based model fine-tuned in the Conceptual Captions (CC) dataset, and fine-tuned for the identification of religiously hateful memes. For this study, the visual features were obtained through ResNeXT-152 Aggregated Residual Transformations-based Masked (R-CNN) and the textual features through Bidirectional Encoder Representation from Transformer (BERT). A separate evaluation of the model showed the classification of multimodal hateful content using an implementation of the Area Under the Operator Characteristic Curve (AUROC) curve with a score of 78 and the accuracy of the model at 70% [12].

An important problem of antisemitic discourse was discussed in this paper [13], potentially igniting hatred and violence. The study employs large language models such as the Bidirectional Encoder Representative Transformer (BERT) to track the contextual resemblance of posts on the extremity of social media using a dataset assembled and annotated within the Unmasking Antisemitism project. The method comprises an

online unsupervised machine learning method where clustering is employed for clustering the like posts and subsequent formation of new clusters as subcategories or whole new categories of antisemitic discourse surface. As with most papers of this type, there is little concern with the kinds of accuracy that lend themselves to quantification; instead, coverage measures are used to track the development of anti-Semitic terminology within this that has been identified. Also in another study [14], authors target to define and detect antisemitism and islamophobia on the 4chan/pol/ board, working with both text and image data. Using a collection of 66 million posts and 5.8 million images shared in 18 months, the study estimates 173000 Islamophobic memes (21000 images and 246000 posts) and 420 hateful phrases and the method uses OpenAI's CLIP (Contrastive Language-Image Pre-training) for detecting images that match semantically with islamophobic phrases. Contextual image understanding is addressed with the use of the CLIP model, whereas the CLIP model is combined with Google's Perspective API (Application Programming Interface) as well as manual annotation for text analysis. This form of sentiment analysis for recognizing Islamophobic memes successfully receives 81% accuracy and 54% f1-score.

In another research [15], the authors investigate the distribution of Antisemitic and Islamophobic materials on the 4chan Politically Incorrect board (/pol/) from July 2016 to February 2021 with the help of the dataset with 204 million posts. The authors used 48 terms to define the Antisemitic content and 135 terms for the Islamophobic content

to investigate the hate speech existence and distribution. The approach incorporates a word embedding method to measure semantic similarity and independently annotated lexicons to determine the correctness of the methods. The paper explains how such ideologies have been embraced in the /pol/ making it a haven for the hatred groups.

Prompting for Multimodal Hateful Meme Classification [16], deals with the diffusion of distasteful memes which is rather a complicated task, especially if it assumes the necessity of reasoning and background knowledge, and understanding. To this end, the authors propose what they refer to as “PromptHate,” which the authors use to refer to the use of prompt-based models for this task. In this regard, the study uses Pre-trained Language Models (PLMs) such as Roberta (Robustly Optimized BERT Approach), as well as fine-tuning to enhance the understanding of the kind of hateful memes by using the contextual information built in the models. Two publicly available datasets were used to conduct the experiments and the remarkable result provoked by PromptHate reflects an AUC(Area Under the Curve) score of 90%. The Machine Learning system is very efficient in identifying and categorizing hate speech memes with an accuracy level of 96%.

ISSUES – a new approach for classifying hate memes [17] was introduced by authors. Multimodal representations learned on text and vision domains are derived using the vision-language model CLIP and textual inversion. The researchers use two datasets: the Hateful Memes Challenge (HMC) dataset and the HarMeme dataset. The stated

strategy includes trainable linear projections for the post-training adaptation of the employed pre-trained models and comprises two stages of training. One of them is applying textual inversion to transfer images from memes to pseudo-word tokens that improve textual information with visuals. It also uses a Combiner network for multi-modal feature fusion in the manner mentioned above. Some of the techniques that have been utilized are separate image and text representations, change of embedding space, and increasing the interchangeability of the fusion function. ISSUES performs almost at most with the state of the art by scoring 77 % on both the datasets. An AUC (Area Under the Curve) of 70% and an AUROC (Area Under the Character Curve) of 85. On the (Hateful Meme Challenge) HMC test set, the accuracy of correlation reached 51% and the accuracy obtained was 81%. Recall 64% and AUROC of 92.83% of the test set in the HarMeme dataset thereby outperforming other approaches.

A semi-supervised learning approach in [18], building on insights from CLIP, utilizing two datasets: including the Multimedia Automatic Misogyny Identification and the Hateful Memes. This approach was divided into two sub-sections. The first step was an unbiased pre-training step, where the Cross Modality Auto Encoder (CROM-AE) was pre-trained using the unlabeled data only. In the second stage, the supervised fine-tuning was done and here a new model was trained to learn the classification task from the labeled data. As in the former stage, in CROM-AE, the encoders were frozen to get new representations from the initial CLIP features. These new representations,

sometimes called “cooked features”, were then concatenated with the CLIP features, or “raw features”, to predict the classification target. The proposed model is called the Raw and Cooked Features Classification Model or RAW-N-COOK and it was shown to achieve better results on the Hateful Memes dataset than the Hate Clipper.

In [19], researchers focus on a Deep learning approach for identifying online tweets containing Islamophobic hate speech. The researchers developed a dataset of 1,290 manually classified tweets into Islamophobia (N= 724) and Non-Islamophobia (N=566,) because the authors could not find any ready-made datasets for performing this particular task. The methodology involves several key steps: data pre-processing of case folding, tokenization, cleaning, stemming and stop word removal, word embedding using Word2Vec, feature extraction using 1D-CNN (One-Dimensional Convolutional Neural Network), and final classification using Bi-directional LSTM Bi-directional Long-Short Term Memory). The model that is used in the current work and is called C-BLSTM, contains seven convolutional layers and two bidirectional LSTM layers. To this aim, the authors tried out different types of recurrent neural network architectures with and without convolutional components. The techniques applied in this paper include Word2Vec for Word embeddings, 1D CNNs for feature mapping, and bi-directional LSTMs for sequence modeling and classification. The C-BLSTM proposed in this paper received the highest accuracy of 90.13% on the test set, better than other herein-described variants, and proving the efficiency of utilizing both convolutional and re-

current layers for this purpose.

In [1], the authors focus on the detection of Islamophobic hate speech from tweets using a deep learning approach. The study revealed that bidirectional layers enhance classification performance than unidirectional layers, but LSTMs outcompeted GRUs, and standard RNNs. In the subsequent experiment, they added the convolutional neural network (CNN) along with the bi-directional LSTM (Bi-LSTM) which gave better results. The last model used CNN for feature extraction and Bi-LSTM for classification in which the training accuracy was 92.39% and test accuracy of 90.13%. In another paper, the authors focus on sentiment analysis to identify the given racist content on the social media platform Twitter. The authors work with a dataset of 31,962 tweets, which includes 2,242 racist ones and 29,720 non-racist ones. The study also proposes LSTM + GCN with BERT as an ensemble model; in this model, LSTM and GCN will act together to improve racist tweet detection. The results illustrate good levels of accuracy with the RNN model being at 0.95 accuracy and LSTM, GRU(Gated Recurrent Unit), and CNN models the Accuracy was quite captivating with 99% proving the efficiency of the proposed method in identifying racism in social networks.

The authors proposed a methodology that works through the detection of racist tweets detected by sentiment analysis [20]. Examining racism as a mediating variable, the authors explain the negative consequences of racism, including direct and indirect racism in the manifestation of memes and comments for social, political, and cultural

order. To overcome this, they suggest the utilization of deep learning methods of GRU, CNN, and RNN and arrive at a final stacked ensemble model of Gated Convolutional Recurrent Neural Networks (GCR-NN). Twitter data was used for training as well as testing the model with a good level of accuracy at 0.98. The GCR-NN model performs better than the traditional machine learning models like the Support Vector Machine and the Logistic Regression, the program accurately classified racist tweets 97% of the time with a misclassification of 3% in comparison to SVM (Support Vector Machine) that was 96% accurate while the LR with 95% accuracy.

Using the multimodal approach, Another paper[21] discusses how to detect hate speech in memes through the internet. The work employs the Hateful Memes dataset of more than 10,000 samples of memes containing text and images. To support the authors' claim, more memes are gathered and merged into the dataset; to provide the analysis, the VisualBERT model, which is a vision-and-language transformer pre-trained on the new dataset, is used. It involves image encoding using a ResNeXT-152-based Mask-RCNN (Recurrent Convolutional Neural Network) model, and ensemble learning in which 27 models are employed, and the final decision is given by a majority voting system. The approach reaches 0% of AUROC on protein level and an accuracy of 0.765 on the challenge test set, and ranked third out of 3,173 competitors in the Hateful Memes Challenge.

In another paper[22], authors developed the Hateful Meme Challenge (HMC) dataset

which was used in this evaluation. The dataset includes 8500 memes in the training set, 500 memes in the development seen split, 540 memes in the development unseen split, 1000 memes in the test seen split, and 2000 memes in the test unseen split. Apart from that, two other assessments were made. On assessment, one evaluation used the Propaganda Memes data set which was developed by authors in [23]. This dataset is unique by its nature; it is multiple-label and multiple-modal, and it includes 22 propaganda categories. In the end, to measure multilingual generalization, the performance was checked with the help of the Tamil Memes dataset which was introduced by authors in [24]. This is a new dataset for the classification of memes in the Tamil language as troll or non-troll.

In a related work by [25], the authors proposed a modular and explainable architecture that incorporated multimodal classification techniques using example and prototype-based learning over training instances. This architectural design uses both the textual as well as the visual State-of-the-Art (SOTA) models to comprehend Internet memes. The model was tested on two existing datasets: The Hate Speech Detection dataset which was developed by authors and the Multimedia Automatic Misogyny Identification (MAMI) dataset. Textual and visual data were used, and BERT was used for extraction of features from text while Contrastive Language Image Pre-training (CLIP) was used for the features from image recognition and clustering parts, features from BERTweet and CLIP were concatenated and employed for downstream prediction. In

each of the meme datasets and methods, the CLIP-based image model had a higher classification accuracy than the BERT-based text models. Also, the fused model, that is, the one that includes the features of CLIP and BERTweet, provided the highest accuracy and is superior to the models based on CLIP only.

The authors examine the features of contrastive learning strategies for servicing the difficult task of detecting misogynistic memes discussed in [25]. Although the experiments that they conducted were not so successful in getting the first rank, these sources are important and worthwhile to provide some exploratory approaches to addressing this particular task. Other pre-trained image encoders were the ones that were provided in the initial presentation to try to get better results. But the best performance came from ResNet-50, a deep CNN trained on over a million training images from the ImageNet database. The text encoder used was DISTILBERT.

In this context [26], they presented the Hateful Meme with Reasons Dataset(HatRed), a new multimodal dataset that automatically explains the referential hate motive of these memes. It is therefore fundamental to unravel the meaning of the hateful memes in the multiformity to fully appreciate the trends in hate speech. To this end, the following framework was created to annotate Facebook’s Fine-Grained Hateful Meme dataset in [27], for the effect of underlying hate: PLMs for Text Generation were trained on HatReD and several tests were performed to assess their strengths and weaknesses. In the case of text-only PLMs, the T5 model and the GPT-2 were employed. Be-

cause GPT-2 is a decoder-only model, RoBERTa was used as the encoder model in this work. For VL-PLMs, VisualBERT was used for benchmarking. Since VisualBERT is an encoder-only model, RoBERTa and GPT2 were employed as decoders in two different ways. This limitation is a weakness of this work because the HatReD dataset only has the annotated reasons for the Facebook memes; the actual memes are located somewhere else; users have to download the memes from the Facebook Hateful Meme Challenge (HMC).

A convolutional GRU-based deep neural network was applied to a public dataset of hate speech [28], which was created by collecting thousands of tweets about religion and refugees using the Twitter Streaming API and then annotated into two classes: love and its opposite hate and non-hate. To enhance learning accuracy, the CNN+GRU neural network model was applied with better functional enhancement such as dropout as well as pooling layer and elastic net. As a regularization technique a dropout layer was inserted and a global max pooling layer and elastic net regularization were used to fine-tune the output of the Softmax layer extracted from the GRU. The model was tested against several baselines and previously reported results on all preceding open datasets, beating 6 out of 7 datasets by as much as 13 in terms of F1 score.

The authors have considered these models for this research; Stacked LSTM, BiLSTM, and CNN. Since there are no existing datasets available for this research, Roshan et al in [29] used memes that were created in the 2016 U.S. presidential election to

develop the MultiOff multimodal meme dataset for offensive meme analysis. A classifier was then designed for this purpose. Image-text fusion was done using an early fusion method and the performance was compared to text-only and image-only modalities. For images, we employed transformer-based models while for the text, we used BiLSTM, and for the embeddings, they were either trained or borrowed from the pre-trained DistilBert model.

2.2 Summary

This chapter presents earlier literature that focuses on the classification of Islamophobic memes. This is a clear gap in the available literature on the subject. Although there is a considerable number of studies on hate speech detection and online content classification, the special focus on identifying and analyzing the Islamophobic and non-Islamophobic content within memes has been paid insufficient attention. That is why, as the results of the review indicate, the majority of works have focused more on textual Hate speech, for instance, on Islamophobic tweets, and non-Islamophobic tweets, than on Islamophobic memes. This is rather an area in research that has not been well developed. Thus, by specifying this problem the review clears the ground for a new study dedicated to the absence of this direction and for the development of the rather uncovered area of Islamophobic memes classification.

Chapter 3

Preliminaries

3.1 Overview:

In the previous chapter, we listed the literature review related to our work. Now, coming to our work, Multimodal in deep learning classification model means the information from different modalities like image, text, audio, and video data are incorporated. Such integration can be done in several techniques that are considered in the methods section for handling multimodal data.

3.2 Overview of Models Applied in Our Methodology

In this study, we have used two hybrid models by creating the theory that is engraved into text and visual data categories in an unimodal manner. These features are then integrated within the multimodal framework and the resulting integrated representation is passed through a further stage of processing.

The first model is a CNN architecture-based ResNet-50 (Residual Network) for im-

age classification, its goal is to differentiate between Islamophobic and Non-Islamophobic images, the other model is a BERT (Bi-directional Encoder Receiver from Transformers) based model for text classification with the same two labels. Below, is a brief overview of both proposed approaches, which are described in detail in the following chapter.

3.2.1 ResNet-50 Model:

ResNet50 is a deep Convolutional Neural Network (CNN) architecture and it was introduced in Microsoft Research in 2015 [30]. It is derived from another network architecture which is called ResNet or Residual Network and the “50” in the name signifies that it is 50 layers deep.

ResNet50 is a fairly strong image classification model that uses Residual blocks allowing it to be trained on massive data sets and accurately classify images. Some of the important features of the proposed network are residual connections which enable the learning of a set of residual functions to realize the input-to-output mapping. The back connections help the network learn much deeper architectures than would have been possible in the past, and without having to face the issue of vanishing gradients.

The architecture of ResNet50 is divided into four main parts as: the Convolutional layers, the Identity block, the Convolutional block, and the fully connected layers. The features in the input image are extracted by the convolutional layers whereas the

identity and convolutional blocks process and transform these features. Also, the fully connected layers are used to classify themselves in the end based on the available data.

The layers in ResNet50 are convolutional layers which are accompanied by several other layers such as batch normalization and ReLU activation. These layers come out with feature extraction from the input image, for instance, edges, texture patterns, and shapes. The convolutional layers are followed by the max-pooling layers which decrease the size of feature maps while extracting the maximum values from those maps.

There are two widely used components in the ResNet-50 model, namely the identity block and the convolutional block. We use the identity block which means passing the input through convolution layers and adding the input back to the output. This permits the network to learn residual functions that map the input into the desired output. The convolutional block is very similar to the identity block but with one feature-specific difference which is a convolution of (1×1) that will be used to down-sample the number of filters before the convolution of (3×3) .

The last component of ResNet50 is the layers of fully connected neurons. These layers are hereby accountable for making the last classification. The output of the final fully connected layer is passed before the softmax activation function to generate the class probabilities.

The ResNet-50 model has a residual learning unit structure as shown in Figure 3.1, where the inputs are passed through the convolution layers and at the same time

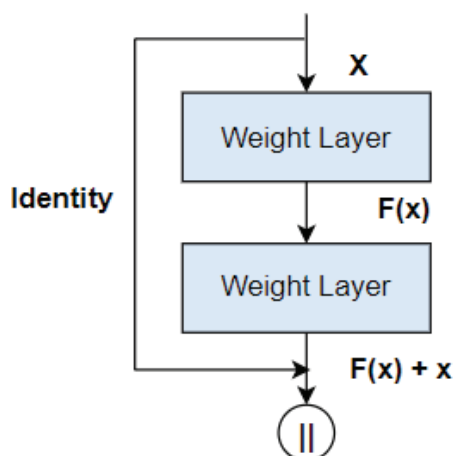


Figure 3.1: Structure Diagram by ResNet-50

passed through what is termed “shortcut connections”. This enables the information to flow through nets with ease thus overcoming some of the problems such as vanishing and exploding gradients that occur in usual deep networks. In this design, ResNet-50 can learn feature hierarchies which in turn aids in faster training of deep networks with fewer parameters. ResNet-50 is more effective than other models because of the extraction of features. It is best suited for image processing and where complicated feature extraction is needed as in the case of user sentiment analysis [31].

3.2.2 BERT Model:

BERT which stands for Bidirectional Encoder Representations from Transformers was developed on a transformer attention mechanism by researchers at Google that identifies the context relationship between words [32]. It consists of an encoder that feeds the text input and a decoder predicts the task to be performed. Contrary to direc-

tional models that go through text word by word, the transformer encoder takes all the words in the text at the same time making it non-directional. This works well because, with such an approach, the model can take into consideration all the other words that surround the particular word of interest.

3.2.2.1 General Architecture of BERT:

BERT is transformer-based, it is essentially an “encoder only” architecture. At a high level, BERT consists of 4 modules [32] as shown in Figure 3.2:

- **Tokenizer:** This module turns an English text into an array of integers, which will be called “tokens”.
- **Token ID Mapping:** This module maps the obtained token sequence into the array containing real numbers of the same length as the array of tokens. It can be described as the transition of discrete token types to space with smaller dimensions where the condition of the Euclidean metric is met.
- **Transformer Layer:** A stack of Transformer blocks with self-attention, however, without the restriction of the causal masking mechanism.
- **Feature Vector:** This module transforms the representations to the token space, and specifically the last representation vectors into one-hot encoded tokens by generating probability distribution over the token types. It can be seen as a

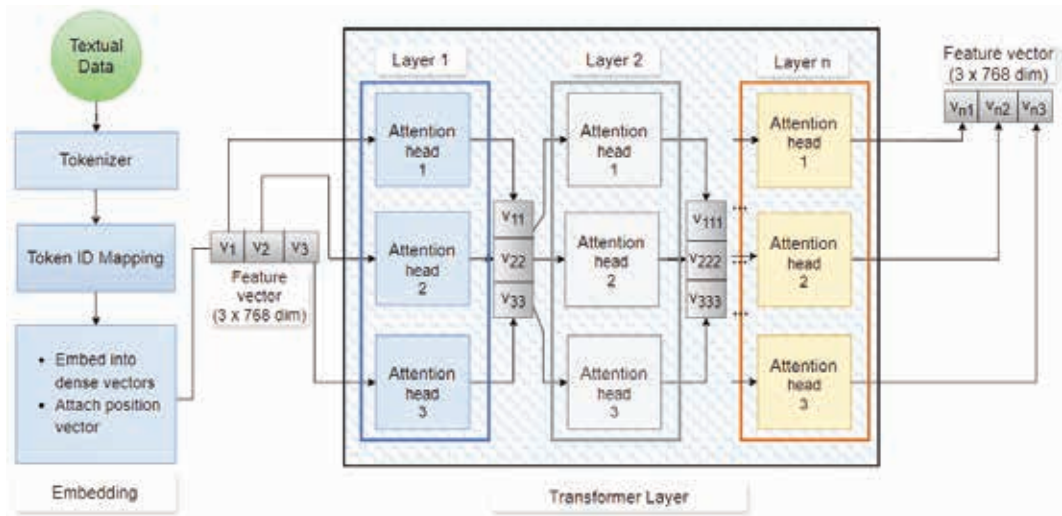


Figure 3.2: General BERT Model Architecture

simple decoder, which decodes the latent representation into token types or as an “un-embedding layer”.

This implies that although a feature vector is crucial in pre-training, it is relatively useful for what is referred to as ‘downstream tasks’ such as Question Answering or Sentiment Classification. However, one gets rid of the task head and puts a new one more appropriate for the task initialized and then, fine-tunes it. The latent vector representation of the model is directly fed to this new module so that we can achieve sample-efficient transfer learning.

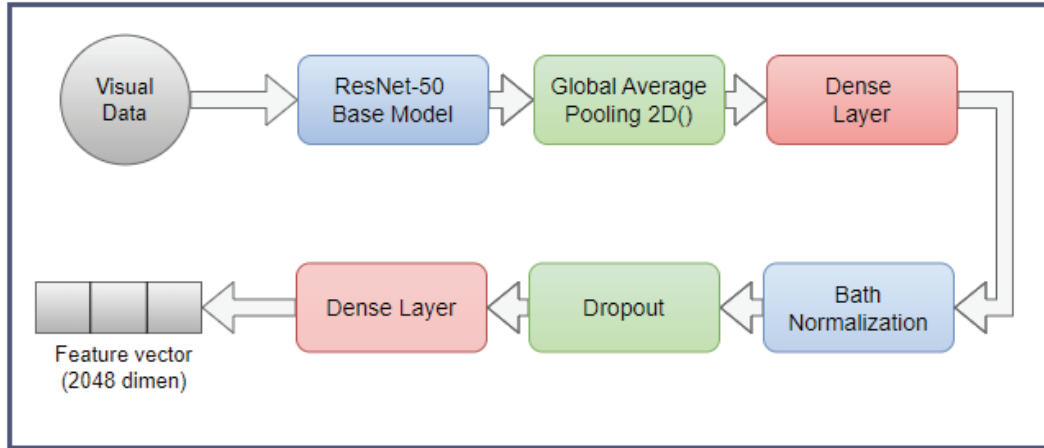


Figure 3.3: General ResNet-50 Architecture

3.3 Multimodal Data and Proposed Models

3.3.1 ResNet-50 Model:

ResNet-50, a deep Convolutional Neural Network (CNN), can also work with multimodal data, which means information from one or multiple sources, for instance, text, audio, images, or videos. Using multimodal data can provide additional information that can improve the performance of ResNet-50 in other tasks such as emotion detection, sentiment analysis, or stress level identification. However, it is not without its challenges; some of them are on how an integration of data from different modalities will be done and also how to handle cases of missing or noisy data.

The strategies of combining the multimodal data for ResNet-50 as shown in Figure 3.3 can be categorized based on the following methods, which are as follow:

- **Early Fusion:** The first pass fusion entails fusing data from the different modal-

ities and passing them into ResNet-50. For instance, if you are working with both text and image data, you are free to concatenate the text embeddings with image pixels to form a single input. The combined input from these two images is then fed to ResNet-50 where it learns a common representation.

- **Late Fusion:** Late fusion can use separate models as in the ResNet-50 for each modality and then fuse them which is normally done using fully connected layers. This approach is beneficial in a way that each of the modalities is processed separately followed by the optimal learning of the features. For instance, two different ResNet-50 models can take an image and text inputs and then their outputs can be combined using a voting method or weighted mean.
- **Hybrid Fusion:** Hybrid fusion process is to combine the features at one or more layers of the ResNet-50 model from different modalities. For example, a cross-channel reconstruction module is used to transfer information between modalities for each convolutional layer.

Since, the problems such as missing or noisy data in multimodal ResNet-50 models some techniques can be applied:

- **Imputation:** Imputation is used while dealing with the data that is missing. Imputation is a method of putting reasonable values to the missing values such

as the average, the middle, or even zero value. This ensures that the model has constant data input so that it does not stop during training or making inferences

- **Regularization:** The purpose of the regularization method is to reduce the noise of the data and avoid the train model that is a good fit for the type of data. The technique that involves adding more restrictions or imposing penalties on the process of learning through the model is called Regularization. Ordinary methods are:

1. **Dropout:** At times a few neurons it excluded throughout the training phase because they hamper the network estimate's ability to generalize and leave no trace of noise.
2. **Batch Normalization:** As mentioned above, stabilizes training making it faster, and contributes to the model's immunity to noise in the input layer.

- **Attention Mechanisms:** Its purpose is to use plots in situations where the data are textual and of varying length and or depth. The encoder and the decoder facilitate attention that allows the forwarding of concentration on specific features or specific modalities sometimes. This is especially advantageous if some of the data is contaminated with noise; the model is then able to rank some areas or parts of the data as being cleaner and therefore containing more information.

- **Data Augmentation:** Its purpose is to guarantee that the proposed model fulfills noise and missing data invariance criteria. Data augmentation means data generation processing which usually reproduces some additional data to the existing data; for example by rotating, scaling, adding some noise, etc. This also assists the model in overcoming issues arising from variations and imperfections within the data.
- **Cross-Modal Learning:** Its purpose is to utilize information of other modalities when one modality is either unavailable or has high noise content. In cross-modal learning, data from multiple modalities are incorporated to offer remediation for the gap in one of the modalities. For instance, when analyzing image data there may be much noise and during this time, the model can use the existing text data to arrive at a proper prediction.
- **Ensemble Methods:** Its purpose is to combine several models to enhance model robustness. In ensemble methods multiple models are trained with different sub-samples or transformed samples of the dataset and its final prediction is made by combining the predictions of all these models. It also means that if there are some noisy or incomplete data then the errors from some of the models are offset by other correct models.

Thus, the presented techniques enable achieving increased performance of ResNet-

50 models in multimodal scenarios with missing or noisy data.

3.3.1.1 Benefits of ResNet-50 for Multimodal Image Classification

The ResNet-50 model has a lot of benefits when applied to multimodal data:

- **Feature Learning:** ResNet-50 outright stands out due to its ability to extract hierarchical features from raw data, thus fits perfectly into the process of pattern recognition within each of the modalities, be it image, text, or any other form of data.
- **Shared Representations:** The results illustrated that ResNet-50 can capture and learn multimodal representation for different modality inputs. This ability is useful in modeling emergent relationships in multiple data sources, which is a great benefit, especially in multimodal applications.
- **Hierarchical Processing:** Some of the properties include; the residual connections means that ResNet-50 can perform hierarchical feature extraction asking to human vision. This kind of hierarchical approach brings advantages especially when sorting out and correlating information from different modalities.
- **Adaptation:** As observed earlier ResNet-50 is highly flexible and can hence be easily fine-tuned for specific multimodal tasks. For this reason, it can be applied to several multimodal scenarios including those involving facial expressions and

emotions as well as sentiment analysis.

- **Transfer Learning:** ResNet-50 models, bought with the large image dataset, can be applied as feature extraction in multimodal tasks. One of the benefits of such an approach is the ability to reuse some of the features and enhance performance without the necessity to expose the model to new multimodal datasets.
- **Handling Complex Data:** The nature of ResNet-50 architecture and design especially the depth allows ResNet-50 to properly process multiple types of data from different sources.

Thus, ResNet-50 demonstrates strengths in feature learning, shared representation, hierarchical processing, adaptability, and transfer learning; making it a model that is capable of effectively addressing the multimodal data challenges.

3.3.2 BERT Model:

Here we can present **BERT** (Bidirectional Encoder Representations from Transformers) as a very effective one focused on the context and attending to the bidirectional text analysis. In the case of multimodal data, the use of BERT can be integrated with other models, for instance, those handling images or audio to improve BERT comprehension and performance. When combining textual information with other modalities, one obtains a textual understanding of interrelations with strengths that are many

times higher than in BERT. It is therefore effective in tasks that require multimodal analysis including sentiment analysis where both image and text features need to be analyzed for correct understanding.

In the case where BERT is used to classify text in multimodal data, the text processing ability of BERT is fused with other models for the other data types such as image, audio, or video. Here's how you can use BERT for text classification in a multimodal setup:

- **Text Processing with BERT:**

1. **Tokenization:** Transform the text data to input that BERT can process, mostly by applying a BERT processing known as BERT tokenizer. This step involves the inclusion of some extra characters such as [CLS] for classification problems and [SEP], which is used to separate different inputs.
2. **Input Representation:** BERT needs three arrays of input and their names are input IDs, attention masks, and token type IDs. These inputs are then used to prepare the text data for feeding for the pre-trained BERT model.
3. **BERT Model:** Feed the input representations through the BERT. For the classification problem usually the [CLS] token representation of the final hidden layer is taken to make the overall text representation.

- **Feature Fusion:**

1. **Concatenation:** Merge the text embedding using BERT with feature embedding of other modalities. This can be done by merging the output arrays of BERT and the other modality-specific models.
2. **Attention Mechanisms:** In addition, other types allow to assignment of different weights to features from different modalities, so the model can ignore ‘noisy’ information.

- **Multimodal Classification:**

Finally, pass the combined features through fully connected layers, softmax function, or other classifiers to get the final classification output. This could be a single prediction based on all modalities i.e., whether the combined text and image or ‘meme’ is Islamophobic or non-Islamophobic.

- **Training and Fine-Tuning:**

Further, optimize the whole of the multimodal model including BERT and the other additional modality models on a labeled dataset containing samples of all the modalities. This training enables the model to combine the information that is obtained from the different sources coherently.

- **Inference:**

At inference, the trained multimodal model is capable of working on unseen data and producing the desired output by comprehensively fusing information from the textual modality through BERT as well as other modalities.

Thus, using BERT's ability to take into account rather complex textual context, along with other modalities, you can build a rather resilient multimodal classification system capable of processing virtually any type of data.

3.3.2.1 Advantages of BERT for Text Classification for Multimodal Data:

Among the challenges that may be encountered while using BERT for text classification on multimodal data, especially when dealing with such phenomena as noise and missing data, there are several approaches designed to maintain the model's stability and further enhance its performance, no matter how far from ideal the data is. Here are some approaches:

- **Data Imputation:**

1. **Handling Missing Data:** To guess the missing content in text, one may attribute it by stand-in for some special key such as [MASK], or by other modalities, such as; Filling in the missing text data based on the image content. This still allows BERT to take input although it will be in-comprehensive.

2. **Cross-Modal Imputation:** Use data from other modalities to supply the missing text data. For example, where text is absent but there is an image, a description of the image may be used to generate text input to BERT.

- **Noise Reduction:**

1. **Pre-processing:** Pre-process the text data before inputting the data to BERT. This can include item elimination, spelling check, and improper formatting elimination among others to minimize noise.
2. **Regularization:** Regularization methods such as dropping out during BERT training will increase robustness to noisy inputs. Dropout randomly omits some neurons during the training process and this aids the model to be less prone to noise.
3. **Data Augmentation:** Introduce noise during the training phase using techniques such as data augmentation; a process by which slight modifications are made to the text. This can help the model to cater for noisy data and enhance the generality of the model.

- **Attention Mechanisms:**

1. **Focus on Relevant Information:** By nature, BERT's attention mechanism assists the overall supposition of the model towards the parts of the text that are

more important. However, when it comes to noise, it can be boosted by modifying the relevance model to give noise-free portions of the text greater weight and push away noisy ones.

2. **Cross-Modal Attention:** In a multimodal system, a cross-modal attention mechanism can be employed for BERT to focus dynamically on information from other modalities. For instance, if the text is dense, then the model will have to go for the image or audio information to get context.

- **Ensemble Methods:**

Combine Multiple Models: Applying ensemble methods when several BERT models (or others) trained on different subsets or versions of the data are learned. This would reduce the effect of noise or missing data since all the models' averages would help even out the noise from all the models.

- **Transfer Learning:**

Use Pre-trained Models: To do this, use BERT models that are pre-trained with other models but have been trained on large datasets of different types of information. These model states are usually less sensitive to noise since the models themselves have been trained on different kinds of text. This is especially beneficial when your multimodal dataset has noisy or missing data as fine-tuning these models will assist with this.

- **Error Handling in Multimodal Fusion:**

Robust Fusion Techniques: When embellishing BERT’s textual output with other media types, employ fusing methods that are robust to lack of data or data noise. For instance, weighted fusion can even consider less important values that belong to the noisy modality, and in case the data obtained from a certain modality is either missing or not very reliable, the system may apply a fallback modality.

It is also beneficial to apply these approaches in the context of the treatment of BERT-based text classification in multimodal data since it allows to minimization of possible problems such as noise or missing data, thus improving the accuracy of classification and reliability of the predictions.

3.4 Summary

In this study, we employed two different models for the multimodal classification work which follows the hybrid approach. The first model is the ResNet-50 model, It is a deep convolutional neural network or deep learning model that has been designed for image classification. ResNet-50 takes the images and applies convolutional layers and fully connected layers to determine whether the image is Islamophobic or non-Islamophobic. The other model is BERT which is an improvement on the transformer model and is primarily used for text classification. From this, it can be seen that BERT works by first tokenizing, then embedding and encoding the text to generate contextual

representations for classification. In multimodal integration, we utilized feature fusion methods for merging the output of ResNet-50 and BERT where the approaches included concatenation and attention. These combined features are then passed through further layers of classification to again ascertain the classification result. Both models are equipped with efficient methods for dealing with noisy and missing data; ResNet-50 uses dropout and data augmentation while BERT uses cross-modal imputation and data augmentation. This cross-stage idea will be utilized to blend the positive aspects of the two models for improving the classification of complex multimodal tasks.

Chapter 4

Proposed Framework

4.1 Overview

In this chapter, the proposed framework detects Islamophobic content and non-Islamophobic content with the use of textual and visual data. It describes techniques for creating subsets of datasets used for training, testing, and validation to provide a consistent assessment of the optimal performance of the model. Some ideas concerning the methods of feature extraction are described, focusing on the analysis of the difference between features and labels. In the model training process the model used for textual data is BERT and for visual data is ResNet-50, where the main attention is given to optimizing the hyperparameters. This approach takes advantage of using both textual and visual data to improve the performance of the detection of Islamophobic content. The basic knowledge of these proposed models was discussed earlier in Chapter 3.

4.1.1 Data Splitting and Feature Extraction:

Both of the pre-processed datasets i.e. textual and visual are split into three subsets i.e. training, testing, and validation datasets. While pre-processing the data, the most crucial step is to divide the obtained sets into training, testing, and validation data sets necessary for the proper assessment of the quality of the created prediction models. The data split is then trained with the help of a training set that contains data for the model's learning. The validation set can also be used in the case of tuning the hyperparameters and selecting the best configuration of the model using it as feedback while training the model. Last of all, the testing set measures the model's robustness on data the Machine Learning (ML) algorithm has not seen before; it helps avoid overfitting. It is important for constructing reliable models to divide the dataset into three categories.

The training dataset plays a very vital role in feeding the model in an attempt to teach by fitting it on data and learning some patterns. Validation data also comes in handy during model development because they offer information halfway through learning, which is beneficial when adjusting hyperparameters and choosing the right configuration of the model. The test dataset which contains independent data is utilized to check the generality of a model and determine the true effectiveness of the model. Every subset serves a special function in making sure the model is trained adequately,

Our Dataset	Sample Count
Training Data	3200
Testing Data	400
Validation Data	400

Table 4.1. Data Splitting

well-adjusted, and works effectively and efficiently under real-world conditions. The split ratio is set to 80:10:10 for training, testing and validation respectively. The sample count of these subsets are shown in Table 4.1.

The next step is to extract features from each of the datasets. The feature and label from each sample is extracted from the original dataset.

- **Features**, known as predictor or independent variables, refer to all the attributes that will be used by the model to make a prediction.
- **Labels**, known as dependent variables, are based on which the model has to learn to forecast, and are actual or the final classification of a given set of data.

4.1.2 Model Training:

The features and labels extracted in the previous stage, for both visual and textual datasets, are fed to the Machine Learning (ML) models. In the initial, the features and labels from the training dataset are fed to the models. The textual data is fed to the BERT model, while the visual data is fed to the ResNet-50 model. Then the text data is fed to the models to evaluate their performance and the models are fine-tuned

to give optimum performance.

4.1.2.1 Training Parameters

Training parameters are another type of hyperparameter that determines how the model is trained in the data. The former must be chosen rather wisely to satisfy the learning process and to get the best performance of the model. In this framework, the following training parameters for text training using BERT are set:

- **batch size = 16:** It is the number of samples that go through the model before parameters are updated. A batch size of 16 means that for every sixteen samples, the model adjusts the internal parameters it has adopted.
- **learning_rate = 5e-5:** This specifies the amount of amplification to be made on the weights of a model for every input error encountered while training the model. This is a very small value and it reduces the amount of learning per step so that the model learns.
- **weight_decay = 0.01:** Weight decay is a form of regularization that allows for discouraging large weights in the model to reduce the risk of overfitting. This means adding a regularization term specifically a weight decay of 0.01 which indicates that during the update process, weights are slightly adjusted downwards which enables the model not to focus much on the training data, instead it is allowed to learn simpler or even general forms of the data patterns.

- **warmup_steps = 0:** So, warmup steps enable the gradual increase in learning rate from the initial small value to the above-specified value through a particular number of iterations. This is when warmup_steps is set to 0, there is no warmup phase and the learning rate starts from 5e-5. Warmup is the process of regulating the training especially in the initial stages when massive updates can jeopardize the training process.
- **num_train_epochs = 5:** This is the number of epochs that this model will pass through the training dataset fully. The constant 'num_train_epochs' equals 5, so the model will pass through each of the samples five times. This is because more epochs can lead to better learning however if it is taken to the extreme then the model may overfit to the training data and therefore be poor on unseen data.
- **max_seq_length = 128:** This is to indicate the maximum sequence length that the model will accept. For text data, this implies that each of the input sequences (for example, a sentence or a paragraph) will be clipped or extended to 128 tokens. Reducing the length of sequencing ensures that the consumption of memory is not abused, as well as invaluable in helping to maintain the standard input size for model processing.

In this framework, the following training parameters for visual training using ResNet-50 are set:

- **ModelCheckpoint:** `ModelCheckpoint('/kaggle/working/best_resnet_model.keras', monitor='val_loss', save_best_only=True, mode='min')`

This callback is used to save the model during training, whenever it achieves a better performance on the validation, that is when the validation loss is lower. The `logs = dict` specifying the following: `monitor = 'val_loss'`, this means that the validation loss is what is going to be being monitored. The `'save_best_only=True'` guarantees that only the model that has the lowest validation loss is saved in order not to overwrite the best model with the worst one. The mode is set to `'min'` which means that the callback is expecting to find the minimum of the monitored metric and in this case the minimum validation loss.

- **EarlyStopping:**

`EarlyStopping(monitor='val_loss', patience=10, restore_best_weights=True)`

This callback halts the training process the moment there is a consecutive epoch that does not record the desired metric, which in this case is the validation loss. The `'patience =10'` means that if the validation loss does not decrease for ten consecutive epochs the program will stop training. The argument, `'restore_best_weights=True'` will make sure that the model goes back to the weights that it had attained its best performance at, thus, preventing the final model from being one that actually overfit or performed poorly.

- **ReduceLROnPlateau:**

```
ReduceLROnPlateau(monitor='val_loss', factor = 0.2 , patience = 5, min_lr = 1e-06)
```

This callback enables the reduction of the learning rate every time validation loss stops increasing, implying that it is no longer decreasing at all. The 'factor=0. 2' means that the learning rate will be increased by adding 0. 2 (will be reduced by 80%) when triggered. The 'patience=5' informs us that the validation loss has to stop dropping in the next five epochs to reduce the learning rate. The 'min_lr= 1e-6' helps in setting to the minimum point to reduce the learning rate below which the learning rate will not be reduced.

- **Epochs:**

```
epochs=50
```

This parameter defines the maximum number of times the training data will be passed through the architecture of the training model. The output section of the model is set in such a way that will train the model for 50 epochs. However, the number of iterations which is represented by 'epochs' could be smaller due to the 'EarlyStopping' method.

- **Callbacks:**

```
callbacks=[checkpoint, early_stop, reduce_lr]
```

This parameter is a list of functions that are called on different phases of training this is true for the classifier. In this case, `checkpoint`, `early_stop`, and `reduce_lr` are the specific callbacks used here to ensure that the model saves the best weights, thus stopping training if the model does not improve and reducing the learning rate to improve the training of the model respectively.

4.2 Framework Architecture

Data preparation and model training are crucial in this case, where the goal is to generate a system for detecting Islamophobic content. This section presents the unified approach used to accomplish a definitive process by which texts and images are effectively classified as Islamophobic and non-Islamophobic. It starts with data segmentation into training, testing, and validation sets, where Resnet-50 and BERT models were used for visual and textual classification respectively.

Figure 4.1 presents a detailed summary of the overall architecture. It gives a unique structure for the combination of Deep Learning techniques in our proposed framework, to get aligned with multimodal detection. The main plan of its framework is to create a highly unified architecture of concatenated visual and textual features acquired from ResNet-50 and BERT respectively. The proposed framework includes four main steps:

1. Visual Feature Extraction.
2. Textual Feature Extraction.

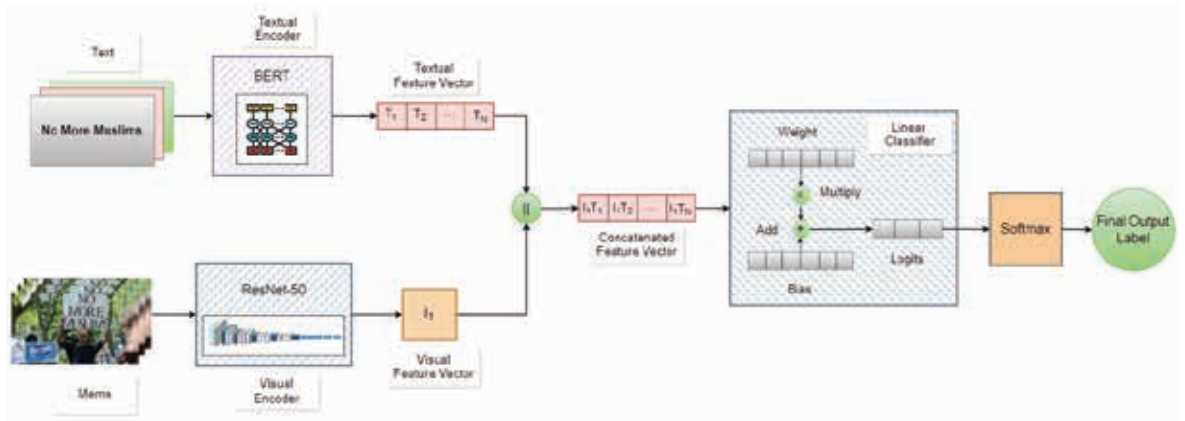


Figure 4.1: Proposed Architecture

3. Unified Visual and Textual Feature Vectors.
4. Mathematical Illusion.

4.2.1 *Visual Feature Extraction*

The proposed visual model, ResNet-50 used for extracting visual features from the dataset, was introduced in Microsoft research in 2015 [30]. It is derived from another network architecture called a Residual Network and contains 50 convolutional layers. It uses the ResNet-50 model for visual classification, as when the visual dataset is loaded into the model, it is then trained by passing through all the convolutional layers and including some of the parameters like; the callback parameter which is a list of functions that are called on different phases of training. In this case, checkpoints, early stopping, and reduce learning rate are the specific callbacks used to ensure that the models save the best weights and generate a unique feature vector.

4.2.2 *Textual Feature Extraction*

The proposed textual model, pre-trained BERT named “Bidirectional Encoder Representation from Transformers”, was developed on a transformer attention mechanism by researchers at Google that identifies the content relationship between words [?]. BERT, which is utilizable into two main versions; BERT-base and BERT-large. In this framework, it utilizes the BERT-base pre-trained model which has 768 network modules and 12 encoder layers [?].

For textual classification, when the textual labeled dataset is loaded into the BERT model, it then tokenizes the text to generate a textual feature vector.

4.2.3 *Unified Visual and Textual Framework*

After extraction of visual and textual feature vectors from our respective proposed models, concatenate them into a unified single feature vector. Convolutional techniques are used after the feature extraction which includes concatenation, linear classifier, softmax, and argmax functions.

First, the combined feature vector passes through the linear classifier, the final layer to transform the features learned from the images (and probably text embedding) into the class probabilities. The linear classifier would have a weight matrix and bias terms. The weight matrix in the linear classifier is multiplied by the input feature vector to obtain the weighted sum. The bias terms were added after performing the dot product

between the feature vectors and the weight matrix. The result of this linear classifier is logits, which is a final feature vector after adding bias terms. It is the unnormalized scores for each class before softmax normalization.

Softmax function is applied on the logits turning into a probability distribution for each of the classes. This dimension has the same value as the number of classes. The last stage is the selection of the class with the maximum probability value, as the result of the calculation. This is done using the argmax function that gives the index of the maximum value in the model probability vector. The index obtained from the argmax function is then converted into the class label. The index that the model assigns to images is one of the two classes named simply “Islamophobic” or “non-Islamophobic”. Classify the output label as Islamophobic or Non-Islamophobic.

4.2.4 *Mathematical Illusion*

Giving a detailed explanation of feature extraction for both modalities, this part provides a mathematical illusion of the proposed framework. It starts with explaining the expressions of individual features (i.e. visual and textual) and their unified features followed by different functions to get an output expression.

The textual features extracted by BERT, are mathematically expressed as X_T , and textual labels are connected with the textual features (i.e. Islamophobic and Non-Islamophobic). While the visual features extracted by ResNet-50, are mathematically

expressed as X_I .

The textual feature vector extracted from the BERT model is as follows;

$$X_T = \text{BERT}[T_1, T_2, \dots, T_n]$$

Where, $X_T \in \mathbb{R}^{d_T}, d_T$ is the dimensionality of the text vector. $[T_1, T_2, \dots, T_n]$ is the tokenizer, which is generated using BERT, containing n number of tokens. The image feature vector extracted from the ResNet-50 model is as follows;

$$X_I = \text{ResNet}[I_1]$$

Where, $X_I \in \mathbb{R}^{d_I}, d_I$ is the dimensionality of the visual vector. I_1 is the visual feature vector generated using ResNet-50. The extracted textual and visual feature vectors from both models are combined to create a multimodal feature vector;

$$X_C = (\mathbf{X}_T \parallel \mathbf{X}_I)$$

Where, $X_T \in \mathbb{R} \quad \mathbf{X}_T \parallel \mathbf{X}_I$, combine both models' feature vectors into a unified feature vector. The combined feature vector then passes through the last layer, a linear classifier, to predict the label O_L .

$$O_L = \text{softmax}(wX_C + b)$$

where, w and b are the learnable parameters in which w is weight and b is the bias term. It gives the output which is the probability distribution of classes containing Islamophobic and non-Islamophobic labels.

To classify the labels as Islamophobic or non-Islamophobic, the argmax function is used, which predicts the output label as;

$$O_{\text{pred}} = \arg \max(O_L)$$

Where, $O_{\text{pred}} = 0$ if $p_0 < p_1$ (Islamophobic), $O_{\text{pred}} = 1$ if $p_0 > p_1$ (*Non – Islamophobic*).

The cross-entropy loss is computed between the predicted labels O_{pred} and actual labels O .

$$L = - \sum_{i=1}^n O \log(O_i)$$

where, n is the number of samples. O is the actual label and O_i is the predicted probability with class i .

The accuracy Acc of the unified model is computed by dividing correct prediction t_c by total prediction t_p which is mathematically expressed as;

$$\text{Acc} = t_c/t_p$$

4.3 Summary:

In this chapter, we explored the critical stages of our proposed framework that are essential for creating effective deep-learning models. We discussed the importance of cleaning and splitting data to improve model accuracy and reliability. The feature extraction process, which involves extracting attributes and labels from datasets, was also highlighted. Model training involved using BERT for text and ResNet-50 for images, with detailed explanations of various parameters and hyperparameters. It further elaborated on the combination of feature vectors from both models into a unified

representation and the subsequent classification using a linear layer. The application of the softmax function to convert logits into class probabilities and the argmax function to select the predicted class were also covered. This comprehensive approach ensures the development of a model capable of making accurate and reliable predictions based on both textual and visual data.

Chapter 5

Dataset and Model Selection

In this chapter, we will discuss how we collect datasets from different platforms and augment them to be used in our dataset.

5.1 Data Collection

The paper with the title “The Hateful Memes Challenge” is devoted to a multimodal classification dataset that contains 12,140 samples aimed at the identification of hate speech in memes. The dataset is designed with a more complicated structure to pose the dilemma to the AI systems to identify both text and image information. It has some challenging cases referred to as ‘benign confounders’ which are designed to ensure that unimodal models (be they text or image) cannot excel, thus highlighting the importance of multimodal reasoning. The dataset entails memes, which are texts combined with images that can be harmless and harmful depending on how the text and image are

presented, making the recognition of AI quite difficult since it is a creative way of presenting hate speech[22].

Out of 12,140 samples of hateful memes, we extracted 2000 samples of Islamophobic and non-Islamophobic memes, combined.

5.2 Data Augmentation:

Augmentation of data is one of the central approaches to handling the problem of insufficient data in the dataset or dealing with datasets where certain classes are significantly rarer than others. Flipping, rotation, scaling, and adding noise are some of the most common transformations that enable the expansion of the dataset through data augmentation. This makes the models generalize better since they are trained to cater to a variety of variations beyond those that were initially trained on. In the text and visual tasks, similar augmentation techniques such as synonym replacement or shifting of pitch in a given voice further model resilience as it learns to handle variability in real-life applications. To sum up, data augmentation is critically important for enhancing the model's performance while it is impossible to collect more data. As shown in Figure 5.1, some specific keywords like; #islamophobia, #antimuslim, #uknews, etc., were used to augment the dataset and searched from social media platforms.

We use data augmentation for;

- **Reducing Overfitting:** One major problem that tends to arise when working

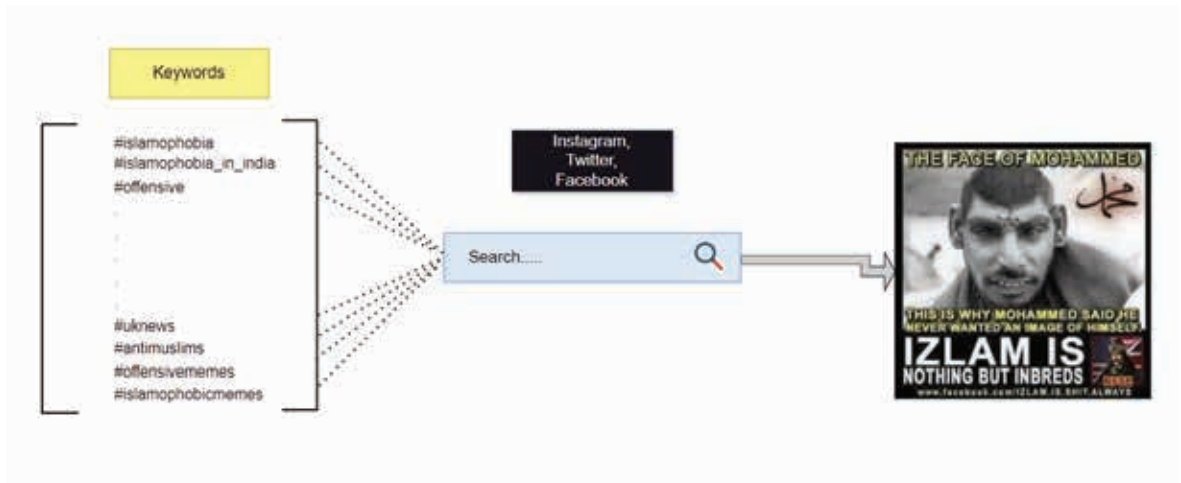


Figure 5.1: Dataset Search on different Platforms Using Specific Keywords

with small data sets is known as ‘overfitting’. Data augmentation can also perform random transformations (e.g., rotations, flips, color jittering) to an input so that the model does not over-rely on distinctive features of the training data. This makes the model make more general features and this is useful for any new data that the model encounters.

- **Simulating Real-World Variability:** Real-world images can be in any orientation and have different lighting conditions or even occlusions as compared to the images used in memes. Whereas augmentation can mimic such variations in the training data, thus training the model to be more robust to such changes at the time of inference.

The sample count for datasets is given in Table 5.1.

Dataset	Referenced Dataset	Extracted Dataset	Augmented Dataset
Sample Count	12,140	2000	4000

Table 5.1. Dataset Collection

5.3 Data Annotation:

Data annotation is the process of categorizing image, text, audio, or video data so that it can be of use in machine learning models. This may involve integrating tags, categories or metadata to be used in the specification of several data sets based on their content. For instance, as in (image annotation), objects within an image may be assigned names such as ‘cat,’ ‘car,’ ‘tree’ and so on so that the model can learn to recognize these objects in new images. In-text annotation, one or many related individual textual units may be tagged for sentiment or parts of speech, for example, ‘positive,’ ‘negative,’ ‘noun,’ ‘verb,’ etc.

Data annotation can be manual, with the assistance of people, or can be automatic, with the assistance of computer programs; while the latter is faster, it is generally less accurate. However, the process could be quite time-consuming and demanding in terms of resources, and is very important to develop high-quality training datasets. Meme images were used and participants were asked to annotate it by watching these images and text collectively. If there was nothing that the annotator found as offending then the annotator had to label it as either an Islamophobic or non-Islamophobic meme.

Table 5.2 shows the Percentage Label of data. There is a split of 50/50 between

Label	Counts	Percentage
Islamophobic Memes	2000	50%
Non-Islamophobic Memes	2000	50%

Table 5.2. Percentage Label of Data

Islamophobic and non-Islamophobic memes.

Our dataset is balanced with an equal number of samples for each label, there are several key benefits:

- **Reduced Risk of Model Bias:** Balanced data does not have a bias wherein most of the test data is of the majority class. This is especially the case for classification tasks where the model may be favoring one class over the other which results in poor accuracies on the underrepresented class.
- **More Stable Training:** When the training is done for a balanced set of data it may make the convergence of the model more stable and accurate. It can foresee problems such as a gradient imbalance during optimization, which may be an evident problem given often imbalanced data.

In conclusion, it is important to have a balanced dataset since it leads to a better, fair, and robust model.

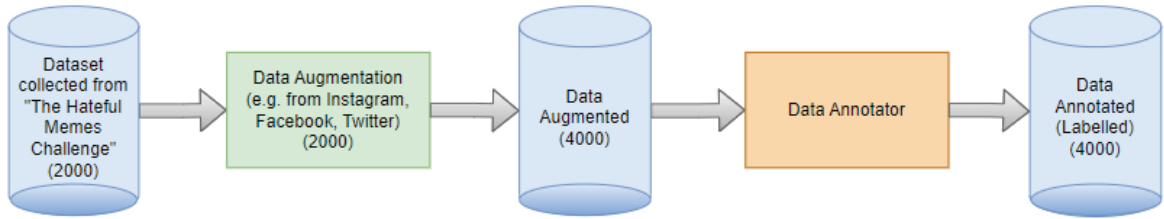


Figure 5.2: Dataset Collection

5.4 Data Extraction and Labeling:

In our approach, we process meme images by dividing them into two components: such as textual and visual data. Here we gather the text from each meme and attach it to the image path, as well as to the label belonging to a textual data set. This two-fold representation enables us to process the textual and visual data as two different modalities while still keeping a connection in case of joint analysis. We are doing this, because we are using ResNet-50 and BERT models in our multimodal and we need textual and visual data respectively, for the above two models.

The dataset is collected using the particular flow as shown in Figure 5.2.

5.4.1 Summary:

In this chapter, we describe how to obtain and expand the dataset for the identification of Islamophobia and non-Islamophobia in memes. First, we gathered a dataset from “The Hateful Memes Challenge” which presented 12,140 samples from which 2000 (Islamophobic and non-Islamophobic) memes were chosen for the research. To obtain more data, flipping and rotation techniques were used with the image thus expand-

ing the sample size to 4000. They also described data annotation as a technique of physical labeling to ensure that there is a 50/50 split between both Islamophobic and non-Islamophobic memes to ensure that there are minimum biases in the model during training. Lastly, the dataset is then divided into textual and visual forms to align the use of BERT and ResNet-50 in multimodal analysis.

Chapter 6

Evaluation and Results

6.1 Overview

In this chapter, we show the results of our multimodal approach for visual and text classification, their accuracy, and cross-entropy loss. Discuss evaluation parameters and show a graphical representation of accuracy and cross-entropy loss. We adopted uni-modal and multimodal systems to evaluate VGG-16, ResNet-50, SetFit, SVM, and BERT performances. Then we compare those unimodal accuracies and based on the results, we use the models that have high accuracy for our proposed multimodal.

6.2 Evaluation Metrics

Evaluation measures are parameters that are used to rate a model on a given job or activity. They inform you about the efficiency of the model, usually after the training session is done. In our model, accuracy, and loss functions are evaluated which are as follows:

Accuracy: The percentage of instances making up the entire picture that corresponded to predicted values. It is computed as;

```
acc = flat_accuracy(logits, label_ids)
```

```
epoch_total_val_acc += acc
```

```
valid_acc += acc
```

Mathematically, the accuracy Acc of the unified model is computed by dividing correct prediction t_c by total prediction t_p which is mathematically expressed as;

$Acc = t_c/t_p$ Using this formula, our framework achieved an accuracy of 95%.

- **Cross-Entropy Loss Function:**

The cross-entropy loss function is often used in machine learning, especially in classification-type problems. It indicates the difference between the predicted possibilities and actual class values.

```
loss = criterion(b_logits, b_labels)
```

Mathematically, to classify the labels as Islamophobic or non-Islamophobic, the argmax function is used, which predicts the output label as;

$$O_{\text{pred}} = \arg \max(O_L)$$

Where, $O_{\text{pred}} = 0$ if $p_0 < p_1$ (Islamophobic),

$O_{\text{pred}} = 1$ if $p_0 > p_1$ (Non - Islamophobic)

The cross-entropy loss is computed between the predicted labels O_{pred} and actual labels O .

$$L = - \sum_{i=1}^n O \log(O_i)$$

where, n is the number of samples. O is the actual label and O_i is the predicted probability with class i .

6.3 Results and Discussion:

We have implemented our dataset on different uni-models and a multimodel approach, from which we have created a comparison of these models as shown in Figure 6.1 and made a discussion of limitations, drawbacks, and benefits of these models are discussed below;

First, uni models were implemented for both visual and textual data. The models used for visual data classification are VGG-16 (Visual Geometry Group) and ResNet-50.

VGG-16 model for 15 epochs, achieves an accuracy of 49%, which is not acceptable for classification. A VGG-16 model might be less accurate, about 49% for visual data due to several reasons such as; its depth being relatively small compared with deeper architectures thus, it is less suitable for feature extraction for complex data types. Moreover, one major drawback of VGG-16 is its large size of the network parameters which can lead to over-fitting, if the training dataset is small. In terms of architecture, it is also quite simplistic and not innovatively designed with state-of-the-art features such as residual connections, or attention mechanisms which enhance the performance

of models in modern deep learning models.

ResNet-50 was used for visual data classification, with an epoch of 50, it shows an accuracy of 78%, which is acceptable if further improved. From the ResNet-50, there are several advantages that it presents to visual classification, especially with the 78% accuracy. Its primary strength relies on its residual connections which go a long way in helping to solve the vanishing gradient problem allowing deep networks to be trained without the same loss of performance. This leads to improved feature extraction and accuracy of the classification done by the algorithms. Furthermore, ResNet-50 architecture is well-optimized in terms of parameters to use compared to other more complex networks but yields excellent results. This is because deeper networks like VGG-16 have a higher accuracy because of their capability to capture and learn the details of patterns from visual images.

For textual data classification, three models were implemented including SetFit, Support Vector Machine (SVM), and BERT.

SetFit is implemented with the epoch of 30 and gives an accuracy of 73%. Finally, we have discussed the drawbacks of SetFit, although in an application for few-shot learning especially in textual classification. It heavily depends on the quality and sample representativeness of the few labeled examples, which may result in rather poor performance when the examples are not adequate or varied. Also, the performance of SetFit might be reduced by the limitations that are associated with a particular

pre-trained language model that SetFit uses, which, in turn, may limit the generality of processed texts. Furthermore, the parameter setting of the model may need to be fine-tuned in certain classification problems for efficient solutions.

SVM is also implemented and gives an accuracy of 85%. A disadvantage of Support Vector Machines (SVMs) might represent the fact that they are practitioners might find them problematic in certain textual classification endeavors owing to their high sensitivity to feature scaling and the delicate procedure for choosing parameters. They may have poor results when dealing with very large datasets or large-sized feature spaces because in large sizes the computational complexity of SVMs increases and it takes more time to train the model. However, SVMs may not be as able to capture non-linear feature dependencies in text as models with higher degrees of complexity such as deep learning models and this may cause performance degradation on texts that include rich context information.

BERT is implemented with the epoch of 5 and gives an accuracy of 94%, which is very promising for textual data classification, as compared to previous textual models. These aspects are again a huge advantage for text classification in the case of BERT – Bidirectional Encoder Representations from Transformers. Besides, it performs very well in analyzing language patterns and the contextual relations between words, thus improving classification efficiency. BERT uses large amounts of text data to help it achieve state-of-the-art performance across a multitude of text classification disciplines.

Uni-models	Accuracy
VGG-16	49%
ResNet-50	78%
SetFit	73%
SVM	85%
BERT	94%

Table 6.1. Model Performance Comparison

Multimodel	Accuracy
BERT + ResNet-50	95%

Table 6.2. Multimodel Accuracy

The performance comparison of different uni-model accuracies using our dataset is shown in Table 6.1.

After the evaluation of the implemented uni-models mentioned above, we used a hybrid approach by combining multiple uni-modals and designed a multimodal that gives final predictions, based on both visual and textual information from the input data. We have used ResNet-50 and BERT for visual and textual data classification respectively, given that they demonstrate the best performance as shown in Table 6.2.

The two models i.e. ResNet-50 and BERT were further evaluated for cross-entropy loss function. The cross-entropy loss function is used to measure prediction error. It measures the degree of dissimilarity between the probability distribution with the desired output of the ML model and the actual labels. The closer the references will be to the true labels the smaller the cross-entropy loss to be obtained. Cross-entropy loss for ResNet-50 and BERT models are found to be 44% and 20% respectively. Hence, the

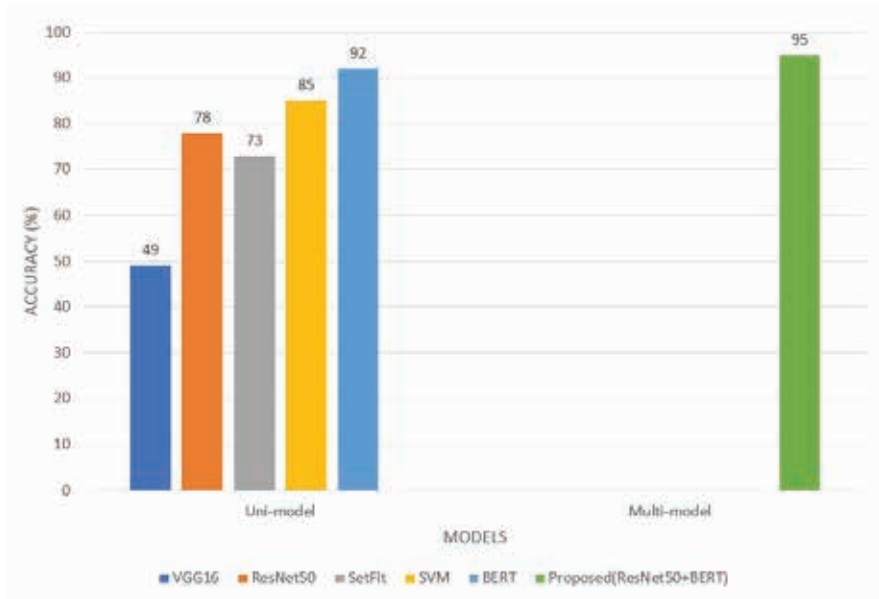


Figure 6.1: Comparison Chart

two models are considered suitable for developing a multimodal that performs detection based on both i.e. textual and visual data.

The epochs for the proposed multimodal are 5 epochs and it gives an accuracy of 95% and entropy loss of 15% as shown in Figures 6.2 and 6.3. Hence, it is very suitable for the classification of Islamophobic and non-Islamophobic memes.

6.4 Summary

The preliminary experiments show that it is possible to gain basic insights from standard uni-modal models like VGG-16, SetFit, and SVM, and further enhanced multimodal that include ResNet-50 and BERT which outperform them. The proposed approach, based on ResNet-50 and BERT for processing visual and textual data re-

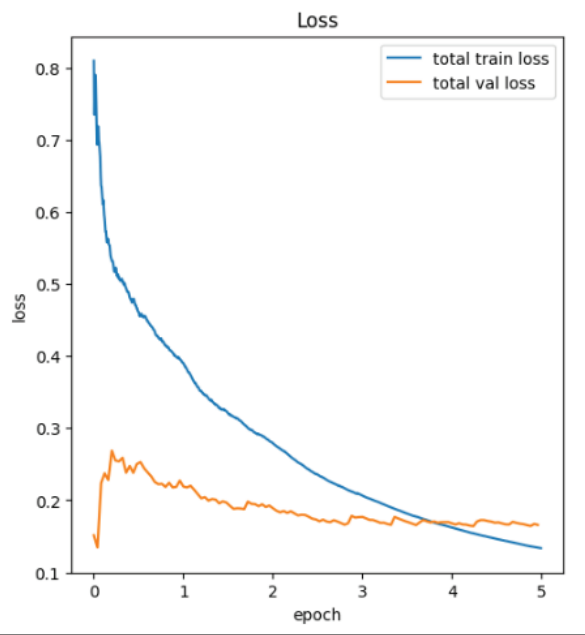


Figure 6.2: Cross-Entropy Loss

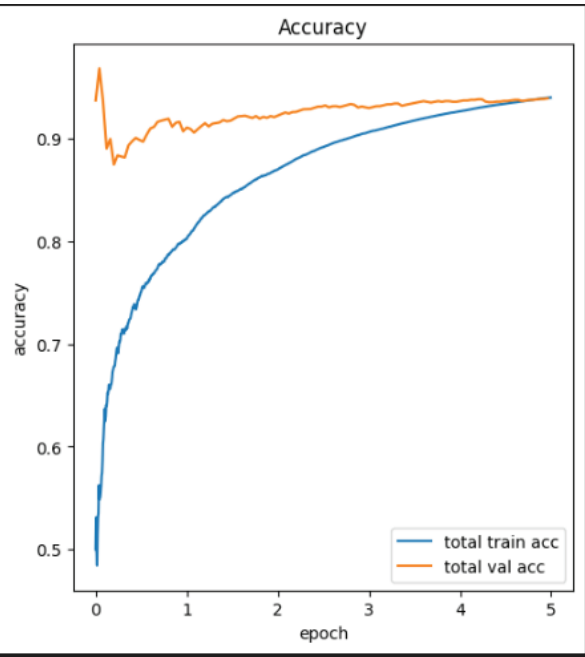


Figure 6.3: Accuracy

spectively resulted in the highest training accuracy of 95% conclusively supporting the use of multiple data inputs. The values of the cross-entropy loss metrics which is only 15% also helped in identifying the better multimodel. However, the studies reveal that the application of more profound architecture and multimodal approach can improve the level of classification and can be useful for classifying challenging tasks like Islamophobic meme analysis.

Chapter 7

Conclusion and Future Work

7.1 Overview

This thesis addresses the problem of identifying Islamophobic content, with emphasis on the role of memes in spreading hate content. We have also shown proof of the effectiveness of a range of new classification models here, including both visual and textual classifications that involved ResNet-50 as well as BERT. Thus, our contribution highlights the need to use complex models and multimodal approaches to deal with such a creative and sensitive phenomenon as Islamophobia in social media. Further, we will discuss future work that will be implemented in the future for effective content moderation processes.

In this research, we have addressed the problem of detecting and removing Islamophobic content, especially memes, from social media. Social media memes are a popular and powerful communication tool that in this case can be used to express opinions and feelings; however, its negative potential for spreading hatred and discrim-

ination is also alarming. Existing deep learning approaches are limited to unimodal approaches, which can only classify either textual or visual data, thereby not understanding the overall context of memes, including both text and visual data. In response to this, we proposed a multimodal-based technique that is capable of classifying Islamophobic memes and non-Islamophobic memes, based on both, visual and textual data. It combines multiple unimodals i.e. BERT and Resnet-50, to develop a model that can process both textual and visual data. Multiple visual models including VGG-16 and ResNet-50 have been tested, but ResNet-50 shows the best detection accuracy. Similarly, multiple textual classification models have been tested including SVM and Setfit, but modern models like BERT have better performances because of the deep learning techniques and understanding of context. The proposed unified model gives an accuracy of 95% and a cross-entropy loss of 15% while integrating both images and text data. Therefore, the results re-emphasize the necessity of deploying hierarchical architectures and multimodal methodologies for handling sophisticated classification problems including identifying Islamophobia in memes.

7.2 Limitations:

In this chapter, we discuss the following limitations and threats related to our proposed model which are as follow:

- One major source of concern is the availability and quality of the data. The

present data may also consist of noise, errors, or inconsistency in the data that may result in poor performance and accuracy of our models.

- Our models will not be able to detect and defend against adversarial attacks like; adding a small, well-calculated noise or interference to an input (e.g. images or text). This might not be able to provide an efficient and effective way of our content moderation processes.
- Our model only uses visual and textual data for Islamophobic and non-Islamophobic content detection. They cannot properly describe additional messages and some of them may be ignored.

7.3 Future Work:

In the future, we aim to expand our dataset to enhance the robustness and efficiency of our models used. Also, we are looking for ways of detecting Islamophobic content using different modalities like videos, and audio. It can describe additional messages that might be ignored in the case of using only visual and textual descriptions. Moreover, we plan to develop more robust and advanced models to detect and defend adversarial attacks. This will in turn increase the effectiveness and effectiveness of our content moderation processes.

7.4 Summary

In this chapter, we will discuss the conclusion of our results that are obtained using an advanced and efficient multimodal approach and give unique outputs. Further, for more accurate and fair performance, we will spend time on building more complete and fair datasets by including more Islamophobic and non-Islamophobic data. Lastly, we perceive multimodal content detection where we combine images, and text to comprehend any Islamophobic content and be in a position to prevent it.

Bibliography

- [1] E. Aldreabi and J. Blackburn, “Enhancing automated hate speech detection: Addressing islamophobia and freedom of speech in online discussions,” in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2023, pp. 644–651.
- [2] E. Nshom, “Introduction: understanding prejudice and communication,” in *Research Handbook on Communication and Prejudice*. Edward Elgar Publishing, 2024, pp. 10–22.
- [3] Z. Gabsi, “Critique of the islamic discourse on islamophobia,” in *Muslim Perspectives on Islamophobia: From Misconceptions to Reason*. Springer, 2024, pp. 79–116.
- [4] A. Jaleel, M. Anwar, F. Ali, R. Mukhtar, and M. Farooq, “Islamophobia content detection using natural language processing,” *Journal of Computing & Biomedical Informatics*, vol. 4, no. 02, pp. 88–97, 2023.
- [5] M. O. I. Elamin, “Counteracting islamophobia through strategic media narratives: A multi-case study approach,” *International Journal*, vol. 5, no. 10, pp. 2733–2750, 2024.
- [6] M. Y. Gada *et al.*, “An analysis of islamophobia and the anti-islam discourse: Common themes, parallel narratives, and legitimate apprehensions,” *American Journal of Islam and Society*, vol. 34, no. 4, pp. 56–69, 2017.

- [7] I. Awan, “Islamophobia on social media: A qualitative analysis of the facebook’s walls of hate,” *International Journal of Cyber Criminology*, vol. 10, no. 1, p. 1, 2016.
- [8] G. Arya, M. K. Hasan, A. Bagwari, N. Safie, S. Islam, F. R. A. Ahmed, A. De, M. A. Khan, and T. M. Ghazal, “Multimodal hate speech detection in memes using contrastive language-image pre-training,” *IEEE Access*, 2024.
- [9] Y. Qu, X. He, S. Pierson, M. Backes, Y. Zhang, and S. Zannettou, “On the evolution of (hateful) memes by means of multimodal contrastive learning,” in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 293–310.
- [10] M. S. Hee, S. Sharma, R. Cao, P. Nandi, P. Nakov, T. Chakraborty, and R. K.-W. Lee, “Recent advances in hate speech moderation: Multimodality and the role of large models,” *arXiv preprint arXiv:2401.16727*, 2024.
- [11] T.-N. Hoang and D. Kim, “Supervised contrastive resnet and transfer learning for the in-vehicle intrusion detection system,” *Expert Systems with Applications*, vol. 238, p. 122181, 2024.
- [12] A. Hamza, A. R. Javed, F. Iqbal, A. Yasin, G. Srivastava, D. Połap, T. R. Gadekallu, and Z. Jalil, “Multimodal religiously hateful social media memes classification based on textual and image data,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 8, pp. 1–19, 2024.
- [13] R. U. Mustafa and N. Japkowicz, “Monitoring the evolution of antisemitic discourse on extremist social media using bert,” *arXiv preprint arXiv:2403.05548*, 2024.
- [14] F. González-Pizarro and S. Zannettou, “Understanding and detecting hateful content using contrastive learning,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, 2023, pp. 257–268.
- [15] M. Ali and S. Zannettou, “Analyzing antisemitism and islamophobia using a lexicon-based approach.” in *ICWSM Workshops*, 2022.

- [16] R. Cao, R. K.-W. Lee, W.-H. Chong, and J. Jiang, “Prompting for multimodal hateful meme classification,” *arXiv preprint arXiv:2302.04156*, 2023.
- [17] G. Burbi, A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo, “Mapping memes to words for multimodal hateful meme classification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2832–2836.
- [18] P. T. H. Tung, N. T. Viet, N. T. Anh, and P. D. Hung, “Semimemes: A semi-supervised learning approach for multimodal memes analysis,” in *International Conference on Computational Collective Intelligence*. Springer, 2023, pp. 565–577.
- [19] Q. Mehmood, A. Kaleem, and I. Siddiqi, “Islamophobic hate speech detection from electronic media using deep learning,” in *Mediterranean conference on pattern recognition and artificial intelligence*. Springer, 2021, pp. 187–200.
- [20] A. NAZMA and C. S. RAO, “Sentiment based racism detection of tweets using deep learning modules.”
- [21] R. Velioglu and J. Rose, “Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge,” *arXiv preprint arXiv:2012.12975*, 2020.
- [22] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testugine, “The hateful memes challenge: Detecting hate speech in multimodal memes,” *Advances in neural information processing systems*, vol. 33, pp. 2611–2624, 2020.
- [23] S. Sharma, F. Alam, M. S. Akhtar, D. Dimitrov, G. D. S. Martino, H. Firooz, A. Halevy, F. Silvestri, P. Nakov, and T. Chakraborty, “Detecting and understanding harmful memes: A survey,” *arXiv preprint arXiv:2205.04274*, 2022.
- [24] S. Suryawanshi, B. R. Chakravarthi, P. Verma, M. Arcan, J. P. McCrae, and P. Buitelaar, “A dataset for troll classification of tamilmemes,” in *Proceedings of the WILDRE5–5th workshop on indian language data: resources and evaluation*, 2020, pp. 7–13.

- [25] A. K. Thakur, F. Ilievski, H.-Å. Sandlin, Z. Sourati, L. Luceri, R. Tommasini, and A. Mermoud, “Multimodal and explainable internet meme classification,” *arXiv preprint arXiv:2212.05612*, 2022.
- [26] M. S. Hee, W.-H. Chong, and R. K.-W. Lee, “Decoding the underlying meaning of multimodal hateful memes,” *arXiv preprint arXiv:2305.17678*, 2023.
- [27] L. Mathias, S. Nie, A. M. Davani, D. Kiela, V. Prabhakaran, B. Vidgen, and Z. Waseem, “Findings of the woah 5 shared task on fine grained hateful memes detection,” in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 2021, pp. 201–206.
- [28] Z. Zhang, D. Robinson, and J. Tepper, “Detecting hate speech on twitter using a convolution-gru based deep neural network,” in *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer, 2018, pp. 745–760.
- [29] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, and P. Buitelaar, “Multimodal meme dataset (multioff) for identifying offensive content in image and text,” in *Proceedings of the second workshop on trolling, aggression and cyberbullying*, 2020, pp. 32–41.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [31] M. Shafiq and Z. Gu, “Deep residual learning for image recognition: A survey,” *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022.
- [32] J. Devlin, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.