CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD

Analysis of Violence Incitation in Social Media from Urdu Content using NLP Techniques

by

Muhammad Shahid Khan

A dissertation submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Computing
Department of Computer Science

2025

Analysis of Violence Incitation in Social Media from Urdu Content using NLP Techniques

By

Muhammad Shahid Khan

(DCS193001)

Dr. Kashif Naseer Qureshi, Associate Professor

University of Limerick, Ireland

(Foreign Evaluator 1)

Dr. Wasif Afzal, Professor

Mälardalens University, Västerås, Sweden

(Foreign Evaluator 2)

Dr. Aamer Nadeem

(Research Supervisor)

Dr. Mohammad Masroor Ahmed

(Head, Department of Computer Science)

Dr. Muhammad Abdul Qadir

(Dean, Faculty of Computing)

DEPARTMENT OF COMPUTER SCIENCE

CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY

ISLAMABAD

2025

Dedicated to my family and teachers.

CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY ISLAMABAD

Expressway, Kahuta Road, Zone-V, Islamabad
Phone:+92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: https://www.cust.edu.pk

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the dissertation, entitled "**Analysis of Violence Incitation in Social Media from Urdu Content using NLP Techniques**" was conducted under the supervision of **Dr. Aamer Nadeem**. No part of this dissertation has been submitted anywhere else for any other degree. This dissertation is submitted to the **Department of Computer Science, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Computer Science.** The open defence of the dissertation was conducted on **August 18, 2025**.

Student Name : Muhammad Shahid Khan (DCS193001)

The Examination Committee unanimously agrees to award PhD degree in the mentioned field.

Examination Committee :

(a) External Examiner 1: Dr. Ayyaz Hussain, Professor, QAU, Islamabad

(b) External Examiner 2: Dr. Ahmad Din, Professor, FAST-NU, Islamabad

(c) Internal Examiner : Dr. Nadeem Anjum, Professor, CUST, Islamabad

Supervisor Name : Dr. Aamer Nadeem, Professor, CUST, Islamabad

Name of HoD : Dr. Mohammad Masroor Ahmed, Professor, CUST, Islamabad

Name of Dean : Dr. Muhammad Abdul Qadir, Professor, CUST, Islamabad

AUTHOR'S DECLARATION

I, **Muhammad Shahid Khan** (**Registration No. DCS193001**), hereby state that my dissertation titled, '**Analysis of Violence Incitation in Social Media from Urdu Content using NLP Techniques'** is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my PhD Degree.

M. Shahid

(**Muhammad Shahid Khan**)

Dated: 18 August, 2025

Registration No: DCS193001

AUTHOR'S DECLARATION

I, **Muhammad Shahid Khan** (**Registration No. DCS193001**), hereby state that my dissertation titled, '**Analysis of Violence Incitation in Social Media from Urdu Content using NLP Techniques'** is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my PhD Degree.

M. Shahid

(**Muhammad Shahid Khan**)

Dated: 18 August, 2025

Registration No: DCS193001

List of Publications

It is certified that following publication(s) have been made out of the research work that has been carried out for this dissertation:-

1. **Muhammad Shahid Khan**, Muhammad Shahid Iqbal Malik, and Aamer Nadeem, Detection of violence incitation expressions in Urdu tweets using convolutional neural network," *Expert Systems with Applications*, vol. 245, p. 123174, 2024.

(Muhammad Shahid Khan)
Registration No: DCS193001

Acknowledgement

All praise be to Allah Almighty, the most merciful, the most beneficent, who enabled me to acquire whatever knowledge that I have and to complete this degree. I would like to express my deepest gratitude to my wonderful supervisor Dr. Aamer Nadeem for his invaluable contribution and advice in undertaking this project. I owe a lot of thanks to him, he was always very kind and ready to guide me. Without his able guidance, this dissertation would not have been possible and I shall eternally be grateful to him for his support. I must also acknowledge the help that I got from Dr. M Shahid Iqbal who was always prepared to guide me in achieving the knowledge that was required to complete this project. I have to offer my gratitude to Dr. Muhammad Abdul Qadir and all my teachers who have always been a guiding light for me in achieving the knowledge and completing the project. I also want to thank my all teachers who were always helpful to me. I am thankful to my parents, siblings, colleagues and friends who extended whatever help I needed from them.

(Muhammad Shahid Khan)

Abstract

The popularity and widespread use of social media are constantly generating unmonitored data, spreading unwanted content such as hate speech and expressions that incite violence. Automatic detection of violence incitation is a challenging task and to the best of current knowledge, Urdu language has been completely neglected. Therefore, a robust framework is proposed for identifying expressions exhibiting violence incitation in Urdu tweets. The potentials of the semantic, word embeddings, and language models are explored to learn contextualized representations of the violence incitation in Urdu tweets. In addition, the strength of the 1-Dimensional Convolutional Neural Network (1D-CNN) is exploited by tunning its parameters on the newly proposed annotated Urdu corpus. The annotated dataset consists of 4808 tweets manually collected from Pakistani Twitter (now rebranded as X) accounts. The performance of 1D-CNN with word uni-gram, Urdu Bidirectional Encoder Representations from Transformer (Urdu-BERT), and Urdu- Robustly Optimized BERT Approach (Urdu-RoBERTa) models is compared to fine-tuned Urdu-RoBERTa, Bidirectional Long short-term memory (BiLSTM), Convolutional BiLSTM (CBi-LSTM), and six state-of-the-art Machine Learning (ML) models. The results reveal that the 1D-CNN with word uni-gram model shows benchmark performance by demonstrating 89.84% accuracy and 89.80% macro f1-score. Furthermore, it outperforms all comparable models and achieves 89.76% f1-score for the violence class, and 89.84% f1-score for not-violence class identification. The uniqueness of the proposed model is evaluated using MARS shine-through and MARS occlusion metrics and the CNN model outperformed the others. The MARS metrics facilitate evaluation and visualization of the classifier performance in terms of capturing unique true positive samples that are not predicted by other models. The findings of the proposed framework are very supportive for further investigation in this domain.

Contents

List of Figures

List of Tables

Abbreviations

AUC	Area Under the Curve
BiLSTM	Bidirectional Long Short-Term Memory
BOW	Bag of Words
CBiLSTM	Convolutional Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
F1	F1-Score (harmonic mean of precision and recall)
KNN	K-Nearest Neighbors
LR	Logistic Regression
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
SVM	Support Vector Machines
TF-IDF	Term Frequency-Inverse Document Frequency

Symbols

α	Learning Rate
β	Regularization Parameter
γ	Hyperparameter for SVM
θ	Model Parameters

Chapter 1

Introduction

This chapter explores the critical issue of violence incitation in social media, with a specific focus on the Urdu language. The chapter begins by introduction to the main domain, following this it discusses the motivation behind this research, emphasizing the urgency and significance of developing effective detection mechanisms, defining the problem background, and its implications for society and online platforms. The chapter also presents the core research questions that guide investigation, setting the stage for the subsequent chapters that delve into related work, research methodology, experiments, and findings. Through this structured approach, it aims to provide a comprehensive overview of the challenges and its contributions to the field of violence incitation detection.

1.1 Overview

This chapter provides an overview of the research by introducing the core domain of online violent extremism detection, with a particular focus on incitement to violence in Urdu-language social media content. It begins by outlining the problem background, highlighting the rising use of digital platforms, especially Twitter, by extremist entities to disseminate propaganda and incite violence. The problem statement emphasizes the lack of effective automated systems tailored to low-resource languages like Urdu for identifying such harmful content. The chapter then presents the key research questions and objectives that guide the dissertation, including the development of a reliable framework for detecting incitement to

violence. The scope and limitations of the research are defined, noting the focus on textual content in Urdu and the challenges posed by linguistic diversity, annotation subjectivity, and dataset representativeness. Finally, the dissertation organization is outlined, describing the structure of the subsequent chapters and how each contributes to addressing the research problem.

1.2 Introduction to Main Domain

1.2.1 Violence Incitation

In recent years, the proliferation of social media platforms has facilitated unprecedented levels of communication and connectivity among individuals worldwide. While these platforms offer opportunities for sharing ideas, fostering communities, and disseminating information, they have also become breeding grounds for the propagation of harmful and inflammatory content through Islamist organizations [1], jihadist movements [2], extremist [3], and online propaganda. One such concerning phenomenon is the incitement of violence in media [4], sectarianisation [5], and racial violence [6] which poses significant challenges to societal harmony, political stability, and individual safety [7]. Incitement to violence on social media platforms [8] encompasses a range of behaviors, from explicit calls for physical harm to veiled threats, hate speech [9], rumors [10] and the glorification of violent acts. The prevalence of violence incitation on social media has raised critical questions about the ethical responsibilities of platform operators, the limitations of free speech in online spaces, and the mechanisms for regulating harmful content [11]. Understanding the dynamics of violence incitation on social media is essential for developing effective strategies to mitigate its impact and safeguard the well-being of users and communities. This research aims to explore the complexities of violence incitation in Urdu-language content on social media platforms, with a specific focus on detecting such content and identifying the targeted groups, in order to understand its linguistic patterns, societal impact, and implications for automated monitoring and intervention systems.

1.2.2 Motivation

The significance of investigating violence incitation on social media platforms cannot be overstated, given its far-reaching ramifications for individuals, communities, and society at large. In today's interconnected world, where social media platforms serve as primary channels for information dissemination and public discourse, understanding the dynamics of violence incitation is crucial for safeguarding democratic principles, promoting social cohesion, and upholding human rights. Moreover, the pervasiveness of violence incitation on social media poses profound ethical and moral dilemmas, challenging notions of free speech, online governance, and the balance between individual liberties and collective well-being [12]. By delving into this complex and multifaceted phenomenon, this research endeavors to shed light on the underlying mechanisms driving violence incitation, the sociocultural factors shaping its prevalence, and the technological affordances that facilitate its dissemination. Furthermore, by highlighting the real-world impacts of violence incitation, including its role in exacerbating intergroup conflicts, fueling extremist ideologies, and perpetuating systemic inequalities, this dissertation aims to galvanize stakeholders across sectors to take proactive measures to address this pressing societal issue. Ultimately, the motivation to choose this topic stems from a deep-seated commitment to fostering a safer, more inclusive digital environment where individuals can engage in meaningful dialogue, express diverse perspectives, and coexist peacefully, free from the specter of violence and intimidation. There are related terminologies in violence incitation domain likely to be similar in some contexts for detection of extremism and threatening.

1.2.3 Violence Detection

Violence detection in the context of this research refers to the identification of language or content within tweets that explicitly or implicitly incites or endorses physical harm, aggression, or any form of malicious intent. This includes instances where the language suggests an imminent threat of harm or an encouragement of violent behavior.

1.2.4 Threatening Language

Threatening language involves explicit expressions of intent to cause harm, danger, or fear. Unlike violence detection, threatening language may not necessarily lead to physical harm but often implies a clear and credible risk. Threats can be directed towards individuals, groups, or entities, and they may include explicit mentions of harm or malicious actions.

1.2.5 Abusive Language

Abusive language encompasses a broader range of harmful expressions, including offensive, derogatory, or discriminatory remarks. While abusive language may not explicitly threaten physical harm, it can contribute to a hostile or harmful online environment, affecting individuals emotionally or psychologically.

1.2.6 Distinguishing Factors

1.2.6.1 Intensity of Harm

Violence detection focuses on language associated with physical harm or aggression, while threatening language implies an explicit intent to cause harm. Abusive language, on the other hand, may not involve a direct threat but can still be harmful in terms of emotional or psychological impact.

1.2.6.2 Intent

Violence detection involves identifying content that explicitly or implicitly supports violence. Threatening language involves a clear expression of intent to harm, and abusive language may be more general but still harmful.

1.3 Problem Background

The use of the latest technologies provides the researchers with a greater opportunity to closely observe the unwanted/ harmful contents on social media. Cyberspace has been embraced by terrorists and political extremists, notably Islamist organizations, as a platform for their operations. These groups have been able to

consolidate their interconnected organizational systems through the utilization of Information and Communication Technologies (ICTs). They use cyberspace to coordinate information between themselves and the outside world, as well as to partially direct and supervise their operations. It is acknowledged that the usage of ICTs has brought about a huge transformation by opening up hitherto unattainable communication opportunities [1]. For the Global Jihadist Movement (GJM) to succeed in its terrorist goals, propaganda and public communication are crucial. In this relevance, a research was carried out on the development of GJM propaganda traits using data from more than 2,000 documents that the GJM released between 1996 and 2005 [2]. Propaganda materials, recruitment campaigns, and ideological messages are all made possible by the Internet, a revolutionary technology. Although Al Qaeda has a longer history, the Islamic State actively pursues a contemporary and advanced media strategy that revolves around the use of social media sites. It is implied by the use of the term "media strategies" that these terrorist groups are using a range of digital communication tactics to achieve their goals [13]. Because of social media, the Darknet, and the Internet, the concerning aspect of Sunni extremism poses a serious threat to civilization. These platforms are seen as important channels via which terrorist propaganda is disseminated and new recruits are drawn in. Using neural networks and deep learning, a thorough investigation was carried out in this context, containing "extremist" or "benign" content. This was accomplished by creating a high-quality dataset for training and testing and by assembling a team of forty individuals, some of whom were fluent in Arabic, who put in 9,500 hours of total work [3].

Social microblogs are a common medium for rumours to circulate, often presenting challenges in distinguishing between factual and misleading content. To address this, a comprehensive framework has been employed that integrates word2vec embeddings with contextual representations derived from BERT's bidirectional encoder, enabling more effective identification and interpretation of such information. This framework combines context-aware techniques with content-based methodologies by incorporating discrete emotional cues and metadata features, enabling a more nuanced understanding of violent content in social media posts. Four real-

world Twitter [1] microblog datasets were used to test this model's rumour identification capabilities. The results showed that the model successfully identified 97, 86, 85 and 80% of rumours on each dataset. It performed well than other three most recent state-of-the-art baselines. The BERT model was found to perform better amongst context-based approaches. The linguistic features were found to function as a standalone model amongst content-based methods. Furthermore, the efficiency of detection model was more enhanced by the use of two step feature selection [10]. The challenging issue of hate speech affects online social networks (OSNs), and it needs to be controlled if OSNs are to continue to flourish. The state-of-the-art is surpassed by BiCHAT Model, which was trained and assessed using three benchmark datasets taken from Twitter in a study, with improvements in recall, f-score, and precision of 8, 7 and 8% respectively. Additionally, BiCHAT demonstrated gains of 5 and 9% respectively, in training and validation accuracy [9].

During Rwanda's genocide in 1994, the media was heavily involved. Incitement to violence and propaganda were the two main formats of media communications. Three prominent members of the media were found guilty in 2003 by the International Criminal Tribunal for Rwanda (ICTR) of various crimes, including genocide and inciting the commission of genocide. Does "incitement to violence" serve as the foundation for any convictions based on media messages, should this be the case? Acknowledging the challenges in distinguishing between these disparate speech forms, the conclusion is that the convictions stem from "incitement to violence" rather than "hate speech" [8]. The press was so powerful in politics that it was virtually a fourth member of parliament, according to Thomas Carlyle, who first used the term "the fourth estate" in the first half of the 1800s. The adage "the power of the press" acknowledges that the media is still extremely important in politics today. The media plays an even more significant role in influencing governmental decision makers during humanitarian crises. This was seen in a former Yugoslavian province in 1999 when ethnic political violence between Serbs and Kosovar Albanians resulted in the kidnapping, rape, torture, and death of 10,000 citizens. Political commentary and the media are closely related fields that

[1] Twitter rebranded to "X" in 2023, but this dissertation use the term Twitter/X interchangeably for consistency.

occasionally feed off one another as well as coexist. With its slanted, impetuous, and inaccurate reporting, the media negatively impacted public opinion in the Kosovo situation. Young, inexperienced editors' negligence, which should have raised concerns about the reporters' work on the field, made the situation worse. The falsehoods in the media stoked animosity, which in turn sparked riots that claimed the lives of 19 people—11 Albanians and 08 Serbs—along with numerous injuries, house evictions, and the burning of 35 Russian Orthodox churches [4].

1.4 Problem Statement

Online violence incitation constitutes a critical national and international threat, contributing to societal polarization, criminal activities, and extremism. Automatic detection of such harmful content remains a non-trivial challenge, especially for low-resource languages such as Urdu, where the scarcity of annotated corpora, lack of pre-trained task-specific models, and inherent linguistic complexities (e.g., morphological richness, free word order, and script variations) significantly hinder progress. A systematic review of existing literature reveals that there is currently no publicly available benchmark dataset for violence incitation detection in Urdu. Furthermore, no prior work has explicitly addressed the computational modeling and classification of violence-inciting discourse in Urdu social media content, leaving a substantial research gap in developing robust, domain-adapted NLP solutions for this pressing societal problem

1.5 Research Questions

1.5.1 Research Question 1

"How can a high-quality benchmark dataset for violence incitation detection in the low-resource Urdu language be systematically developed from social media platforms (e.g., Twitter), considering the challenges of noisy user-generated text, code-mixing, and linguistic variations?

In addressing the first research question, proposed approach will encompass several key steps. Firstly, it will conduct a comprehensive exploration of various social

media platforms known for their significant Urdu language user base, including X and Facebook, among others. Subsequently, it will identify and select appropriate data sources within these platforms where a substantial volume of Urdu tweets were being generated. To facilitate the data crawling of violence incitation content, it will develop lexicons specific to violence incitation in Urdu, enabling us to effectively search and filter Urdu language content across social media platforms. Leveraging web crawling techniques, it will systematically collect tweets from the selected data sources, ensuring a diverse representation of Urdu language content. Following data acquisition, it will implement robust filtering mechanisms to isolate and retain only those tweets that were written in Urdu, thus ensuring the integrity and relevance of the dataset. Through these methodical steps, it will successfully curate a comprehensive dataset of Urdu language tweets containing instances of violence incitation. Further, it will take a proactive approach by assembling a team of three annotators who are experts in Urdu language comprehension and proficiency. These annotators will be carefully selected based on their fluency and understanding of the Urdu language, ensuring that they possessed the necessary linguistic skills to accurately annotate the dataset. Prior to the annotation process, comprehensive guidelines will be established to ensure consistency and uniformity in the annotation methodology. These guidelines will provide clear instructions and criteria for identifying instances of violence incitation within the Urdu content, thereby facilitating a systematic and structured approach to the annotation task. By leveraging the expertise of these annotators and adhering to the established guidelines, it will be able to meticulously annotate the dataset, thereby laying the groundwork for subsequent analysis and model development aimed at addressing proposed research objectives.

1.5.2 Research Question 2

"Which NLP techniques can effectively capture the lexical, semantic, and contextual cues of violent incitement in Urdu text, and how do they compare with traditional machine learning baselines?"

In tackling the second research question, proposed approach will devise a comprehensive framework that will incorporate several key components. Initially, it

will conduct Urdu-specific preprocessing of the textual data, encompassing tasks such as tokenization, normalization, and stemming, tailored specifically to the nuances of the Urdu language. Subsequently, it will employ a diverse array of feature extraction techniques to capture both textual and contextual characteristics inherent in the Urdu content. These features will include traditional approaches such as Bag-of-Words (BOW), TF-IDF (Term Frequency-Inverse Document Frequency), as well as more advanced techniques like word embeddings and contextual embeddings. Leveraging the extracted features, it will then leverage a variety of machine learning algorithms, deep learning architectures, and state-of-the-art large language models to discern patterns indicative of violence incitation within the Urdu tweets. This holistic approach will facilitate the development of a robust and scalable violence incitation detection system, capable of effectively identifying and flagging instances of violence incitation across Urdu language content on social media platforms. Through the integration of multiple methodologies and techniques, proposed framework will provide a comprehensive solution for the identification of violence incitation in Urdu text, addressing the core objectives of proposed research question.

1.5.3 Research Question 3

“How can fine-grained NLP classification models be designed to identify the targeted communities or groups mentioned in the content, ensuring robustness despite the scarcity of annotated Urdu resources?”

To address the third research question, proposed approach will embark on a targeted approach focused on identifying the specific communities or groups that are the intended recipients of the violence incitation messages. It will begin by isolating the tweets classified as containing violence incitation (the ”Yes” class) from annotated dataset. Subsequently, it will conduct a meticulous examination of each tweet to discern the underlying targets or recipients of the incitement. To categorize the targeted groups effectively, it will delineate three broad categories: Government/Religious, Political, and General Public. To ensure consistency and accuracy in the annotation process, it will develop detailed data annotation guidelines tailored to each group category. These guidelines will provide explicit in-

structions and criteria for annotators to classify tweets according to the identified target groups. Annotators will be then tasked with labeling each tweet with one of the three designated class labels corresponding to the targeted group. Additionally, we will employ various feature extraction techniques to capture the nuanced contextual information embedded within the tweets. Leveraging a combination of machine learning, deep learning, and large language models, it will analyze the annotated dataset to accurately identify and classify the targeted communities or groups in each tweet, thereby fulfilling the objectives of proposed research question.

1.6 Research Objectives

The research objectives of this dissertation revolve around addressing three key research questions concerning the identification and mitigation of violence incitation in Urdu content on social media platforms. Firstly, the dissertation aims to develop a comprehensive dataset of violence incitation content sourced from Urdu social media platforms. This entails collecting, annotating, and curating a diverse range of textual data that contains explicit or implicit calls to violence. Secondly, the research endeavors to investigate methods for effectively identifying violence incitation within Urdu content by leveraging both textual and contextual characteristics. This involves exploring various machine learning and natural language processing techniques to analyze the linguistic features and contextual cues associated with violence incitation. Lastly, the dissertation seeks to develop strategies for identifying targeted communities or groups that may be vulnerable to violence incitation. By examining patterns of dissemination and engagement with violent content, the research aims to discern the specific communities or demographic groups that are most affected by and susceptible to the influence of violence incitation on social media platforms. Through these objectives, the dissertation aims to contribute to the development of proactive measures for detecting and mitigating the spread of violence incitation in online Urdu communities.

1.7 Scope and Limitations

This dissertation has several limitations; First of all, the proposed model is trained on Urdu corpus that is collected from Pakistani X(Twitter) accounts. The results of the research cannot be generalized beyond the intended scope as the size of the dataset is not large enough. Additionally, the source of the dataset is X(Twitter) and a maximum of 280 characters is allowed to write a tweet. On the other hand, Facebook and other social media platforms do not impose any restrictions on the size of comments/posts. Second, we addressed the violence incitation detection as a binary classification. However, violence incitation can be seen as direct-violence, and indirect-violence and targeted persons can be identified. Another limitation is the interpretability and explainability of the proposed framework as this model is not designed with these two criteria in mind. Finally, the proposed system is only designed for the Urdu and is not directly applicable to Roman Urdu in the same settings.

1.8 Dissertation Organization

The remaining part of the dissertation is organized as follows: chapter 2 describes the related work, and chapter 3 describes research methodology followed by proposed framework. The results are presented in chapter 4 with a detailed analysis. Chapter 5 concludes the research work and discusses future directions.

Chapter 2

Related Work

The current research collectively addresses the critical challenge of identifying and mitigating extremism in online platforms, particularly on social media. They employ a variety of methodologies, including machine learning, sentiment analysis, and deep learning, to detect and classify extremist content. These studies collectively contribute to understanding and combating extremism in online spaces through advanced computational methodologies and data-driven approaches.

2.1 Overview

This chapter presents a comprehensive review of related work in the field of online violent extremism detection, focusing on both English and Urdu language contexts. It begins with an overview of existing literature in English, highlighting various computational approaches, machine learning techniques, and annotated datasets used to detect hate speech, extremism, and incitement to violence. The chapter then delves into literature specific to the Urdu language, revealing the limited availability of annotated resources and the lack of tailored models capable of effectively capturing the nuances of extremist content in low-resource linguistic settings. Through critical analysis, several gaps are identified, including insufficient attention to the Urdu language, the absence of incitement-specific annotations, and the underutilization of multimodal or context-aware models. The chapter concludes with a discussion on how these gaps inform the direction of the current research and establish the need for a focused framework addressing

incitement in Urdu social media discourse.

2.2 Literature in High-Resource Languages

Over the past ten years, there has been an increase in the online spread of radical ideologies and causes. With millions of tweets being sent on X every day, it might be difficult to manually sort through the thousands of tweets to find the ones that support hate and extremism and go against the community guidelines set forth by the platform. A study was carried out using multiple linguistic features linked with KNN, SVM algorithms on a sizable real-world dataset to address the immediate problem and show the efficacy of the suggested approach with 97 % accuracy rate [14]. The copious quantity of information on social networking makes it difficult for law enforcement agencies to identify propaganda and content related to terrorism. To categorize violent extremist content, a machine learning methodology was used having two sets of characteristics (data independent and data dependent). The particular dataset has a significant impact on the data dependent characteristics, but the data independent characteristics remained unaffected by the dataset and could be applied to other datasets with comparable outcomes.

When compared to Arabic data, the findings of the AdaBoost classifier performed better with 82.4 % accuracy rate in categorizing English tweets and tweeps [15]. Jihadist groups like ISIS use social media sites like X and YouTube to disseminate their ideology online. When terrorist propaganda is detected, the accounts of these groups are suspended as a tactic to prevent it from spreading. This method involves first personally reading and analyzing a sizable amount of data and a message that jihadist organizations have posted on social media platforms like X, after which a machine learning algorithm (SVM, AdaBoost, NB) is used to categorize the message or tweet that gives more precise results with 97.9 % accuracy rate [16]. The biggest and most well-known microblogging platform on the Internet is called X(Twitter). Because of X's low publication barrier, anonymity, and widespread use, extremists can easily spread their beliefs and viewpoints by tweeting hateful and extremist-promoting content.

Because there are millions of tweets posted on X every day, it is nearly hard for

moderators, intelligence analysts, and security specialists to detect these posts on manual basis. It has been suggested that the use of foul language, religion, conflict, and unpleasant emotions can distinguish hateful and extremist tweets from other types of tweets. A case study was conducted to measure the recall and accuracy of a machine learning-based classifier via KNN method and a single class SVM to categorize a task related to Jihadi tweets. The efficacy of the suggested approach is demonstrated by experimental findings on a sizable and authentic dataset, with F-scores for the KNN and SVM classifiers of 0.60 and 0.83, respectively [17].

Violent extremist (VE) groups are drawn to cyber networks as a result of an increase in the number of persons utilizing the Internet for communication. A research was conducted to predict the everyday activities of VE groups in cyber-recruitment. A support vector machine model was used to identify recruitment posts on a Western jihadist discussion forum. In order to predict cyber-recruitment activity inside the forum, the textual content of this data set was analyzed using latent Dirichlet allocation (LDA). These findings were then included into a number of time series models. When using LDA-based topics as predictors in time series models instead of naive (random walk) topics, the forecast error was decreased. This was the forecasting task's initial result to be published.

In the end, this research may contribute to the effective deployment of intelligence analysts in reaction to anticipated spikes in cyber-recruitment activity [18]. A machine learning system was introduced that makes use of a combination of network, temporal, and metadata variables to identify extremist users and forecast social media interaction reciprocity and content adopters. This technique utilized an exceptional dataset consisting of millions of tweets from over 25,000 users that X manually recognized, flagged and 'deferred because of their association in radical activities. Three forecasting tasks were carried out with this: (i) identifying extremist users; (ii) estimating the likelihood that ordinary consumers would embrace radical content; and (iii) predicting the likelihood that users will return contacts that extremists had started. These forecasting tasks were configured as either a simulated real-time prediction work or a post hoc (time independent) prediction job using aggregated data. This system demonstrated very promising performance, with up to 93% AUC for detection of radical users, 80% AUC for pre-

diction of content adoption, and 72% AUC for prediction of interaction reciprocity in the various forecasting scenarios [19].

Security informatics and cyber threat intelligence are essential tools for identifying the main dangers and influencers linked to extremism and criminal activities in online networks. Numerous obstacles stand in the way of automated research into the detection of extreme internet posts and the major suspects and risks that go along with them., 1) Social media data is mainly unstructured and comes from a variety of diverse, independent sources. 2) Extremism and criminal activity's tactics, methods and procedures (TTPs) are always changing. 3) Insufficient ground truth data exist to facilitate the creation of efficient categorization algorithms. In this relevance a study was carried out by presenting human-machine collaborative and semi-supervised learning system that in the midst of these difficulties can accurately and efficiently detect harmful social media posts. A graph-based optimization method was used that develop an initial classifier as TTPs evolve in an interactive manner using shortlisted relevant samples. This framework is validated by means of an extensive set of flagged terms, both in and out of English, that were taken from three online forums and manually checked by several independent annotators. In this framework, the classifier performance converges more quickly than fully supervised solutions, leading to an accuracy that is over 80% [20].

A large number of like-minded individuals publish abusive remarks against other races and religions on well-known microblogging networks. In order to build a cascaded ensemble learning classifier for identifying postings with racist or radicalized intent, a study was undertaken using the microblogging platform Tumblr to solve the issue of uncertainty in posts by determining author's intent. To train the model, different characteristics of language, sentiment, and semantics from free-form text were categorized. This approach demonstrated very promising performance [21]. A new era in terrorism has been brought about by technological advancements, particularly the rise of microblogging sites like X. These sites are being used for recruitment, as well as for communication and incitement of acts of terrorism. While sentiment analytics can be used to identify and classify user opinions of varying polarities, the majority of existing methods and algorithms do not specifically detect acts of terrorism. In order to improve the present senti-

ment analysis approaches by utilizing machine learning to better correctly detect actions of terrorist, a comparative research comparing sentiment analysis methodologies was done and analyzed. Because machine learning is more accurate than lexicon-based approaches, it was suggested in this study [22].

The emergence of social media specifically X has resulted in a growing cyberwar on the internet through hateful and violent remarks and videos, as well as sophisticated films that spread extremism and radicalization. This is a field of research that is still developing since the content is noisy, short, context-dependent, and dynamic, which presents a number of obstacles. To lessen a high dimensional data space into a low dimensional space (2-D and 3-D) for tweet data with TF-IDF features, an Exploratory Data Analysis (EDA) by means of Principal Component Analysis (PCA) was carried out. The data was divided into two modules: Extreme and Neutral. Extreme included further two subclasses: Pro-Taliban and Pro-Afghanistan government). Better separation of cluster between neutral and extreme classes has been shown by PCA based visualization; however, it has not shown to be very effective within extreme subclasses.

Various classification algorithms were used, such as naive Bayes, K Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest and ensemble classification approaches, in addition to PCA based reduced structures and a full set of TF-IDF structures extorted from n-gram terms in the tweets. When compared to other classification models, the SVM showed an average accuracy of 84% [23].

Online Social Networks have altered the technique that terrorists and extremists can incite and fanaticized individuals. Thus, to stop the extremist narrative from expanding too far, it is crucial to identify radical information when it appears online. A study was conducted to determine how to automatically identify radical content in social media by discovering multiple behavioural, psychological, and textual cues that when combined enable the categorization of radical communications.

To automatically detect extreme tweets on the internet, this study analyzed extremist propaganda on X by developing a radical content paradigm based on contextual texts with psychological traits deduced from these materials. The findings demonstrated that textual models by means of vector embedding features greatly

outperform TF-IDF features in terms of high accuracy detection [24]. The advancement of technology has improved information sharing and communication. A virtual community has been established online by various terror groups, such as jihadist communities, for a variety of reasons, including recruitment, online fundraising, targeting youngsters, and the dissemination of extremist ideology. Diverse agencies are attempting to remove extremist content from different social media websites. In contrast to earlier machine learning algorithm-based initiatives that gathered 61601 records from diverse online sources like news, articles, and blogs, an LSTM based deep learning technique was used for radicalization identification. Domain experts annotate these records into Radical-R, Non Radical-NR and Irrelevant-I categories. An LSTM-based network is then used to classify radical content. 85.9% accuracy was attained using the suggested method [25].

Extremist groups that use violence, such as the Islamic State of Iraq and Syria were able to more readily contact vast audiences, form personal connections, and increase recruiting because to the ease of use of social media. Even while social media has suspended a lot of accounts, there is no certainty that this will stop radicals from returning with new accounts or moving to other social networks. An automatic detection technique was developed to determine whether a given username was connected to an extreme user or not by utilizing three categories of data pertaining to user profiles, usernames, and text content. Results on a real-world dataset from X related to ISIS show how effective the method is at identifying extreme people [26]. Development made so far has been summarized in a tabular form and is given in Table 2.1. The quantity of extreme and radical texts that are published online is always growing as agencies and researchers put a lot of effort into creating instruments to identify and eliminate this kind of information. The majority of the scholars continued to concentrate on English-language resources. Extremist organizations, such as extreme Islamic ones, began to provide content in languages other than English—more precisely, Russian.

To better comprehend radical minds and develop counter-extremist tools, a carefully constructed dataset was combined with qualitative and quantitative research of a radical extremist discourse [27]. In today's society, terrorism in all its forms

TABLE 2.1: Summary of related work for extremism and radicalism detection in social media

Year [Ref]	Language [Data Source]	Characteristics	Algorithms
2014 [14]	English [1 million tweets]	Linguistic features	KNN, SVM
2015 [15]	English, Arabic [6729 X profiles]	Data independent and dependent features	AdaBoost
2015 [16]	English [4000 tweets]	Stylometric, time-based and sentiment features	SVM, AdaBoost, NB
2015 [17]	English [453 k tweets]	War-related, abusive terms, negative emotions, internet slang	KNN, SVM
2015 [18]	English [290 k posts of Ansar AlJihad]	BOW, LDA	SVM
2016 [19]	Arabic [3 million tweets]	Time-based Features, Profile Features, Network Features	RF
2017 [20]	English [17 k tweets]	Topic sensitivity, post effectiveness and emotion indicators	LR
2017 [21]	English [3,228 Tumblr posts]	Semantic, sentiment and linguistic features	Ensemble learning
2017 [22]	English [1,480 tweets]	Sentiment and lexicons features	NB
2019 [23]	English [7,500 Tweets]	N-gram, TF-IDF	SVM
2019 [24]	English [17 k tweets]	Textual, psychological and behavioral features	RF
2019 [25]	English [61,601 posts]	Word2vec	LSTM
2019 [26]	English [300 k tweets]	Content and Profile features	LSTM, SVM
2019 [27]	Russian [699 k posts from Kavkaz]	Stylometric features	KNN, NB, SVM
2020 [28]	English [Dark Web]	doc2vec, word2vec	LR, LSTM
2021 [29]	Arabic [89,816 tweets]	TF-IDF, BERT	LR, SVM, NB, RF
2022 [30]	English [60 k tweets]	BERT, RoBERTa, DistilBERT	Language models

has gained more significance. As a result, there has been a rise in the use of the internet by terrorist organizations for communication and propaganda. LSTM models were engaged as a method to classify somewhat textual content into one of radical groups: White Nationalists, Sunni Islamists, Antifascist Organizations, and Sovereign Citizens. This categorization performed better when compared to non-deep learning techniques. [28]. Extremist organizations always remained in practice to use social media to gain their personal gains. Although extremist propaganda is sometimes disseminated in other languages, including Arabic, most methods for identifying extremist contents are limited to the English language.

Conventional machine learning models like Logistic Regression, Multinomial Naive Bayes, Support Vector Machines, Random Forests, and BERT were employed for identification of extremism. Support vector machine attained the highest accuracy (0.9729) while BERT outperformed with an accuracy of 0.9749 [29].

Social media platforms are very popular with the younger generation and have a global reach. It is crucial to do research on online extremism in order to monitor the impact of radicals and the spread of abhorrence on social media. Many terrorist groups, such as the QAnon and Alt-Right, as well as conspiracy theory groups like Al Qaeda, ISIS, Proud Boys and Taliban, disseminate false information, radicalize, and recruit youth. It is possible to identify extremism from a variety of ideas and divide them into three categories using deep learning: recruiting, radicalization, and dissemination. The creation of a balanced multi-ideology extremism text dataset with multi-class labels and a seed dataset were presented through a research project that involved compiling, cleaning, and organizing radical tweets. The dataset known as Jihadist White Supremacist (MIWS) / Islamic State of Iraq and Syria (ISIS) was created. With the greatest f1-score of 0.72, this dataset was assessed using pre-trained Bidirectional Encoder Representation for Transformers-BERT and other variants, such as DistilBERT and Robustly Optimized BERT Pre-Training Approach-RoBERTa. The f1-scores provided by DistilBERT and RoBERTa are 0.71 and 0.68, respectively [30].

2.2.1 Findings from High-Resource Languages

Early research on online extremism and hate speech (2014–2016) primarily relied on classical machine learning models such as KNN, SVM, Naïve Bayes, and Random Forest, with hand-crafted features like linguistic cues, stylometric patterns, and time-based attributes [14, 16, 19]. While these approaches achieved reasonable accuracy on English and Arabic datasets, they were heavily dependent on feature engineering and lacked the ability to generalize across diverse contexts. This limitation is particularly critical for morphologically rich and low-resource languages like Urdu, where handcrafted features are harder to design and less reliable. Between 2017 and 2019, research expanded with larger datasets (tens of thousands of tweets and posts) and incorporated semantic features and embeddings such as

TF-IDF, n-grams, and eventually Word2Vec [23, 25]. Models like LSTMs and ensemble learning began to appear, demonstrating the gradual shift from shallow classifiers to deep learning architectures. However, most studies remained limited to English, with some work in Arabic and Russian [27, 29], indicating a clear language bias in the field. This reinforces the need for incitement detection research in low-resource languages such as Urdu, which until now remained unexplored.
From 2019 onwards, the literature shows a decisive move toward neural embeddings and contextual language models. Studies applied Word2Vec, Doc2Vec, and eventually BERT-family models (BERT, RoBERTa, Distil-BERT) to capture semantic and contextual meaning more effectively [28, 30]. These approaches overcame the shortcomings of feature engineering by learning directly from data and proved especially effective in detecting implicit or coded incitement. Yet, despite the demonstrated success of deep learning in high-resource languages, there has been no parallel attempt to apply similar techniques to Urdu. Overall, the literature reveals three major gaps that our research addresses:

2.2.1.1 Language Gap

Most work is in English, Arabic, or Russian, with no large-scale annotated Urdu dataset for violence incitement.

2.2.1.2 Domain Gap

Prior studies focus on terrorism, abusive, or offensive content, whereas our work targets violence incitement specifically in Urdu across political, religious, and social contexts.

2.2.1.3 Methodological Gap

While recent studies employ deep contextual embeddings and transformers, these methods have not yet been applied to Urdu incitement detection, which is the novelty of our contribution.

2.3 Literature in Urdu

The summary of related studies for Urdu language is presented in Table 2.2. For a number of applications, including content recommendation and controversial event extraction, controversial speech recognition on social media is essential. Urdu tweets were gathered and classified as controversial or noncontroversial using data-driven models of Support Vector Machine (SVM), Logistic Regression (LR), and Naive Bayes (NB). Moreover, sequential pattern mining and sequential rule mining have been used to examine contentious Urdu tweets in order to identify the most common terms, patterns, and word relationships within patterns. This study also found some intriguing correlations and patterns that can be utilized to automatically identify tweets that contain contentious statements or messages. According to the results, NB performs better on this classification test than SVM and LR [31].

The proliferation of misinformation, disinformation, fake news, and various forms of propaganda is a result of the information spreading so quickly on digital platforms. The online digital world is seriously threatened by information pollution, which has presented several difficulties for governments and social media companies worldwide. Finding the sources and content of propaganda distributed in Urdu was the aim of a pioneering study called Propaganda Spotting in Online Urdu Language (ProSOUL). In this instance, the machine learning classifiers were trained using a tagged dataset of 11,574 Urdu news articles, and the psycho-linguistic elements of the text were extracted using the Linguistic Inquiry and Word Count (LIWC) lexicon. The efficacy of several classifiers was evaluated by modifying News Landscape (NELA), n-gram, Word2Vec and Bidirectional Encoder Representations from Transformers (BERT) features. With an accuracy of 0.91, the word n-gram, character n-gram, and NELA features performed better together for Urdu text classification. However, Word2Vec embedding outperformed BERT features with an accuracy of 0.87 in Urdu text classification. In comparison to other web content, the results demonstrated that the ProSOUL framework outperformed other frameworks in the field of propaganda detection in online Urdu news material [32].

On social media, using slang, vulgar, abusive, foul language, and aggressive ex-

TABLE 2.2: Summary of related works in Urdu language.

Year [Ref]	**Problem Addressed [Data Source]**	**Features**
2017 [31]	Controversial Speeches [X (8000)]	Bag-of-words, TF-IDF
2020 [32]	Propaganda detection [News (11,754)]	LIWC, word2vec, BERT
2020 [16]	Abusive & Slang [X (5000)]	Lexicon-based
2021 [17]	Sentiment analysis [Media websites (9601)]	Word n-gram, FastText model
2021 [18]	Abusive & threatening [X (15000)]	m-BERT model
2022 [19]	Threatening content [X (3564)]	Word and char n-grams, FastText models
2022 [20]	Threatening content [X (3564)]	TF-IDF, n-gram and BOW features
2022 [21]	Abusive and threatening language [X (15 k)]	BERT mode
2022 [22]	Offensive language [Facebook (7500)]	TF-IDF, word n-gram, char n-gram, word2vec

pressions has become the norm. This abhorrent crime is carried out even if social media businesses have censorship policies for such language. The reason for this is the paucity of research and resources on the automatic identification of abusive language systems other than English. An intelligent lexicon-based framework named USAD (Urdu Slang and Abusive Words Detection) was developed in a study to identify slang and abusive words in Urdu tweets written in Perso-Arabic. According to the findings, 72.6 % of tweets may be accurately classified as abusive or non-abusive by the suggested USAD model. Furthermore, a few crucial elements were found that can aid the researchers in enhancing their models for detecting abusive language [33]. The number of Urdu speakers worldwide is over 169 million, and a significant amount of Urdu data is produced every day on different social media networks. There has not been much work done in terms of research studies and efforts to develop Urdu language resources and analyze user sentiment.

An evaluation of various deep learning and machine learning algorithms for sentiment analysis was carried out, along with the establishment of a benchmark dataset for the resource-poor Urdu language. Two types of text representation were examined: one using pre-trained fastText word embeddings for Urdu, while the other was count-based, employing word-n-gramme feature vectors to characterise the text. To conduct experiments for all feature types, a set of deep learning

classifiers (LSTM and 1D-CNN) and machine learning classifiers (MLP, NB, RF, SVM, LR and AdaBoost) were taken into consideration. For the sentiment analysis challenge, the word n-gram feature combination with LR performed better than other classifiers, giving rise to the greatest F 1 score of 82.05% [34].

To discourage hate speech online, several social media companies have implemented moderation guidelines for such content. Scholars in the field of abusive language research conduct several investigations to enhance their ability to identify abusive content. Compared to Hindi, Urdu, etc., there is more research on the detection of abusive language in the English language. Many machine learning models, including XGboost, LGBM, and m-BERT based models for abusive and dangerous content identification in Urdu based on the common task, were exposed in a study. The best results were obtained with the help of the Transformer model, which was carefully trained on an Arabic dataset of harsh language. With an F1score of 0.88 and 0.54 for abusive and threatening content identification, respectively, this model ranked first [35]. A new dataset has been released for the purpose of detecting threatening language in tweets, in order to facilitate current research in Urdu.

In this dataset 3,564 tweets were manually evaluated as hostile or non-threatening by human specialists. These dangerous tweets were then divided into two categories by using a two-step process: threatening to a specific individual or threatening to a group. Three types of text representation were compared in this study: the first two were count-based, the text was characterized by using word or character n-gram counts as feature vectors, and the third type of text representation used fastText's pre-trained word embeddings for Urdu. An MLP classifier that combined word n-gram characteristics performed better than other classifiers in identifying dangerous tweets, according to a series of studies conducted by utilizing deep learning and machine learning classifiers. FastText pre-trained word embedding and SVM classifier were found to function well for the target identification challenge [36].

Twitter is the most popular social media medium for people to express their thoughts, some of which may contain content that is intentionally or inadvertently threatening to other users. Within this framework, languages such as En-

glish, Dutch, and others possess multiple methods for identifying potentially dangerous content; Urdu, with its limited resources, is not as fortunate. Bernoulli Naive Bayes (BNB) and extra tree (ET) classifier based on the Bayes theorem were used as the basis learners to predict a stacking model, although the meta learner utilised in this instance was logistic regression (LR). A performance analysis was carried out using a support vector classifier, BNB, LR, ET, fully connected network, long short-term memory, convolutional neural network, and gated recurrent unit. The stacked model outperformed both machine learning and deep learning models, yielding 74.01% accuracy, 70.84% precision, 75.65% recall, and 73.99% F1 score [37]. The majority of recent studies and state-of-the-art methods emphasis on English as the target language, with very little study being done on low and medium resource languages.

There are two common tasks for identifying threatening and abusive language in Urdu—which is spoken by more than 170 million people worldwide—were described in a study. The study's participating systems were tasked with categorizing tweets in Urdu into two groups: abusive and non-abusive, and threatening and non-threatening. Within the abusive dataset, there were 2400 annotated tweets in the train section and 1100 annotated tweets in the test section. Within the threatening dataset, there were 3950 annotated tweets in the test section and 6000 annotated tweets in the train section. Additionally, baseline classifiers based on BERT and logistic regression were offered in both scenarios. Twenty-one teams from six different nations (Pakistan, United Arab Emirates, China, Taiwan, Malaysia and India) registered to participate in this collaborative endeavour. The best-performing system obtained an F1-score value of 0.880 for Subtask A and 0.545 for Subtask B. For both subtasks, the m-Bert based transformer model yielded the best results [38].

Eliminating unsolicited behaviours requires automatic detection of offensive or unpleasant language. The fact that each language has a unique vocabulary and grammatical structure makes it more difficult to generalize the approach. A novel dataset of 7,500 postings from well-known Pakistani Facebook sites was used to construct a study for recognition of invasive language in Urdu. Four different kinds of engineering models with various attributes were employed: three of them were

frequency-based, while the fourth was an embedding model. Frequency-based models were found using the bag-of-words, word n-gram, or Term Frequency-Inverse Document Frequency (TF-IDF) feature vectors.The fourth model was generated by word2vec model, which was trained on the Urdu embeddings using a corpus of 196,226 Facebook posts. This model remained out performed with an accuracy of 88.27%. Performance was significantly enhanced by the wrapper-based feature selection approach. 90% accuracy and 97% AUC were attained by the hybrid blend of Term Frequency-Inverse Document Frequency, word2vec and bag-of-words models. Furthermore, it performed better than the baseline with improvements of 3.68% in recall, 3.55% in accuracy, 3.67% in precision, 2.71% in AUC, and 3.60% in f1-measure [39].

Online propaganda is a tool used to manipulate people's sentiments on social media. Based on a news archive and numerous public news websites, a study was carried out to discover propaganda frameworks as binary classification models. A number of models, including word unigram, part-of-speech, FastText, LIWC, LSA, word2vec, char tri-gram feature models, and fined tuned BERT, were used in this study. Three techniques for oversampling were examined in order to address the Qprop dataset's imbalance. SMOTE Edited Nearest Neighbours (ENN) was demonstrated to deliver the best outcomes. The optimal model was found to be the BERT-320 sequence length after BERT was fine-tuned. In comparison to other attributes, the char tri-gram performed better when used as a stand-alone model. The pairing of char tri-gram with word2vec and pairing of BERT with char tri-gram showed a solid performance, surpassing both of the state-of-the-art baselines. As compared to previous methods, there was considerable rise in performance and achieved f1-score, more than 97.60% recall, and AUC on the dataset's development and test sections [40].

2.3.1 Findings from Urdu Literature

Research in Urdu literature has gradually evolved to address controversial and extremist speech detection. Early work such as [31] (2017) explored controversial speeches using Bag-of-Words (BOW) and TF-IDF approaches. Although effective in identifying surface-level patterns, these models were largely reliant on hand-

crafted features and lacked semantic depth, limiting their ability to capture nuanced linguistic cues in Urdu texts. By 2020, researchers began integrating semantic and psychological features. For instance, [32] introduced LIWC, word2vec, and BERT for propaganda detection in Urdu news content. This study highlighted the importance of combining lexicon-based approaches with word embeddings, bridging the gap between shallow text features and deeper semantic representations. Similarly, [16] worked on abusive and slang detection using a lexicon-based approach, demonstrating the challenges of handling domain-specific vocabularies in Urdu.

The years 2021 and 2022 showed a significant shift towards neural embeddings and transformer-based models. For example, [17] utilized Word n-grams and FastText for sentiment analysis in Urdu media content, while [18] introduced m-BERT for abusive and threatening speech detection, marking one of the first applications of multilingual transformer models in this domain. Recent studies from 2022 emphasized hybrid and deep learning approaches. Works such as [19] and [20] employed n-grams, TF-IDF, and FastText, reflecting a blend of traditional and modern methods to capture both surface and semantic features. Moreover, [21] and [22] adopted BERT-based and embedding-driven approaches for abusive, threatening, and offensive language on platforms like Facebook. These contributions highlight an increasing reliance on contextual embeddings and transformers, aligning Urdu literature with global advancements in natural language processing.

2.4 Gaps in Literature

Most of the prior work on identifying extremism/aggression content had focused on English, and Arabic languages, and ignored resource poor languages such as Urdu, Roman Urdu, etc. Urdu is the national language of Pakistan and is mainly spoken in the South Asian region. According to Ethnologue, Urdu is the 10th most spoken language in the world and uses two styles, one is Roman and the other is Nastaliq. Nastaliq has a complex morphological and syntactical structure, as it derives words from other languages such as Arabic, Persian, Sanskrit, and some from Turkish [34]. Therefore processing the Urdu language is a challenging task [35–38]. To the

best of knowledge, there is no prior work on the identification of violence incitation content in the Urdu language. To deal with this gap, this dissertation proposed a framework that identifies violence incitation content in the Urdu language on the Twitter platform. To achieve this objective, we contributed in two directions; 1) An Urdu corpus is designed by collecting tweets from the Twitter platform, and 2) State-of-the-art word n-grams, char n-grams, Latent Semantic Analysis (LSA), word2vec, FastText, Urdu-BERT, and Urdu-RoBERTa models are explored with ML and Deep Learning (DL) models. The proposed framework is tested on the newly designed dataset and conclusions are drawn.

2.5 Discussion

The problem of violence incitation detection is deeply rooted in the evolving landscape of online discourse, where harmful content can spread rapidly across social media platforms. As explored in the problem background, violence-inciting language poses a significant threat, particularly in regions with sociopolitical tensions. The review of related work has demonstrated that while substantial progress has been made in detecting such content in English, research in low-resource languages like Urdu remains limited. The challenges of linguistic diversity, contextual ambiguity, and data scarcity further complicate this task. The existing approaches in English provide a strong foundation, but they require adaptation to account for Urdu's unique script and syntactic structure. This dissertation builds upon prior research by addressing these gaps and proposing a framework that leverages advanced NLP techniques to enhance violence incitation detection in Urdu. By bridging the problem background with related work, this discussion underscores the necessity of developing robust, language-specific models that can effectively handle the nuances of Urdu text while ensuring reliable detection across various online platforms.

Chapter 3

Research Methodology

This chapter illustrates the comprehensive discussion of the research methodology for violence-incitation detection and identification of violence-incitation target. The following sections discuss building dataset, pre-processing, evaluation parameters, and strategies used during these processes.

3.1 Overview

This chapter presents a comprehensive explanation of the methodological approach adopted in this research. The research methodology employed in this dissertation encompasses a systematic approach involving dataset construction, framework development, data preprocessing, feature extraction, and experimental evaluation. A specialized dataset was built by collecting and annotating Urdu-language tweets related to violent extremism, sourced primarily from Twitter using relevant keywords and hashtags. This was followed by the design of a proposed framework tailored to detect incitement to violence, integrating machine learning and deep learning techniques. The data underwent rigorous preprocessing to preprocess noise, normalize text, and handle linguistic complexities inherent in the Urdu language. Subsequently, meaningful features were extracted using both traditional methods like TF-IDF and advanced techniques such as word embeddings. Finally, the experimental setup was configured with appropriate algorithms, tools, and evaluation metrics to assess the model's performance in identifying extremist content. This structured methodology ensures the reliability, scalability, and

applicability of the proposed solution in real-world scenarios.

3.2 Building Dataset

This chapter illustrates the comprehensive discussion of the proposed methodology for building datasets for violence incitation detection and identification of target-group. The following sections discuss building dataset, selection of appropriate data sources, building of violence keywords, data collection, data cleansing, and data annotation. The datasets available are in English or Arabic for the purpose of detecting extremism and radicalism. As previously said, there was a dearth of research in Urdu, with the majority of studies conducted in English and Arabic. Pakistan's official language, Urdu, ranks 10th in the world in terms of frequency of usage. X has gained worldwide popularity and reliability that produces 200 billion tweets, or 6,000 tweets every second each year [14]. People can connect globally since it authenticates users' identities and has no follower restriction. X API was utilized in this research to gather tweets from the X network.

3.2.1 Selection of Social Media Platform

The decision to select X over Facebook for the development of a new dataset in Urdu language for violence incitation detection was carefully considered and based on several solid reasons and logical factors. Firstly, X is known for its real-time nature, making it an ideal platform to capture timely and current data related to violence incitation. Additionally, X's character limit per tweet encourages concise and focused expression, which is advantageous for capturing potentially incendiary content in a succinct manner [36, 37]. Furthermore, X boasts a larger user base compared to Facebook in certain regions, including Pakistan, where Urdu is predominantly spoken. This larger user base provides a wider pool of data to draw from, increasing the likelihood of encountering relevant instances of violence incitation. Moreover, X's open API and accessibility to public tweets facilitate data collection efforts, enabling researchers to efficiently gather a diverse range of content for analysis. Lastly, X's platform features, such as hashtags and retweets,

offer valuable contextual information that can aid in the identification and classification of violence incitation content. Overall, these factors collectively make X a more suitable and advantageous choice for building a dataset focused on violence incitation detection in Urdu language, compared to Facebook [14, 15].

3.2.2 Identification of Data Sources

In efforts to identify important data sources for collecting Urdu content with a significant reach, a comprehensive approach was undertaken that involved assessing various types of X accounts. Recognizing the diverse nature of Urdu content, we targeted accounts belonging to a wide range of entities, including newspapers, politicians, news channels, religious scholars, and government X accounts. These accounts were selected based on their relevance and influence within the Urdu-speaking community, as well as their potential to generate substantial engagement and discussion. By targeting accounts with a large number of followers, we aimed to maximize the reach and impact of the content collected. After careful consideration and evaluation, we curated a list of the most prominent and influential X accounts within the Urdu-speaking sphere, ensuring that selected data sources encompassed a diverse range of perspectives and voices. This meticulous process allowed to identify the best X accounts as data sources for this research, laying the foundation for an effective and comprehensive data collection strategy. Table 3.1 displays the information pertaining to the chosen X accounts.

3.2.3 Development of Violence Incitation Lexicons

In subsequent step, it embarked on the development of a comprehensive list of Urdu lexicons tailored specifically for research objectives. This process involved the creation of uni-grams, bigrams, trigrams, and even more extensive lexicons through an iterative approach. Meticulously curated these lexicons by engaging in manual searches, continuously updating and refining the list to ensure its accuracy and relevance. Aim was to compile a diverse set of lexicons that would effectively capture Urdu content on the X platform, facilitating the crawling and collection of data from selected data sources. Throughout this process, we made concerted

TABLE 3.1: A list of chosen X Accounts.

X Account	Category	Followers
PTVNewsOfficial	News	1.8M
geonews urdu	News	5.6M
arynewsud	News	3.6M
aaj urdu	News	1.3M
Nawaiwaqt	News	435K
indyurdu	News	339.2K
SAMAATV	News	3.1M
jang akhbar	News	1M
ExpressNewsPK	News	4.1M
BBCUrdu	News	4.5M
ImranKhanPTI	Person	20.6M
PTIofficial	Political Party	10M
pmln org	Political Party	2.5M
NawazSharifMNS	Person	1.1M
CMShehbaz	Person	6.7M
MaryamNSharif	Person	8.1M
HamzaSS	Person	356K
MediaCellPPP	Political Party	1.1M
BBhuttoZardari	Person	5.1M
AAliZardari	Person	761.3K
MQMPKOfficial	Political Party	32.1K
Juipakofficial	Political Party	266.2K
ANPMarkaz	Political Party	168.1K
ShkhRasheed	Person	8.3M

efforts to filter out or remove lexicons that did not yield satisfactory results in identifying instances of violence incitation in Urdu content. These lexicons were carefully chosen based on both the existing data available on X and deep understanding of the Urdu language. By leveraging a combination of empirical data and linguistic expertise, we were able to develop a robust set of lexicons optimized for detecting violence incitation in Urdu content on X. Table 3.2 portrays a few sample terms from the violence incitation lexicon.

3.2.4 Crawling Dataset

In the subsequent phase of research, a Python bot equipped with various Python libraries was developed to facilitate the crawling of the online X platform. This bot was meticulously designed to leverage predefined Urdu lexicons specifically curated

TABLE 3.2: Examples of keywords used to incite violence.

Type	**English**	**Urdu**
Uni-gram	Burn	جلاؤ
Bi-gram	Torture	تشدد کرو
Bi-gram	Beat with the shoe	چھتر مارو
Tri-gram	Head should be cut off/ Should be beheaded	سر کاٹنا چاہیے
Tri-gram	Should be hanged	پھانسی دینی چاہیے
Tri-gram	Give an excruciating death	عبرتناک موت دو
Higher-gram	Bring to hell	جہنم واصل کر دو

for identifying instances of violence incitation. Utilizing a systematic approach, the bot traversed through each of designated data sources, systematically examining each word from the lexicons against the content available on X. Before downloading any tweet, the bot employed a set of predefined criteria to ensure the relevance and quality of the data retrieved. These criteria included verifying the presence of at least one word from unigram lexicons, confirming the tweet contained Urdu content, and ensuring the tweet met specific length requirements. Once a tweet met these criteria, it was meticulously downloaded and stored in an Excel file, facilitating easy access and analysis. This systematic approach enabled us to efficiently gather a comprehensive dataset of Urdu tweets containing potential instances of violence incitation, laying the foundation for subsequent analyses and investigations.

First, tweets from the aforementioned X accounts were gathered as part of the data acquisition procedure. Three procedures were used to choose a tweet for dataset: Initially, Urdu should be used in the tweet; secondly, it must include any of the lexicon's keywords. Thirdly, the tweet must to have a minimum of five words. For the following four years, from February 2018 to June 2022, the tweets were crawled as part of this process. 2018's general elections and the events leading up to June 2022 were the driving forces behind the selection of this time frame. Several political crises and a protracted state of instability transpired within that time frame. People who sympathized with one political party over another expressed their beliefs in an antagonistic manner. We obtained 10,000 tweets from the gathering process. Specifically, while attempting to detect violence incitation content, we found that brief sentences frequently lack enough context

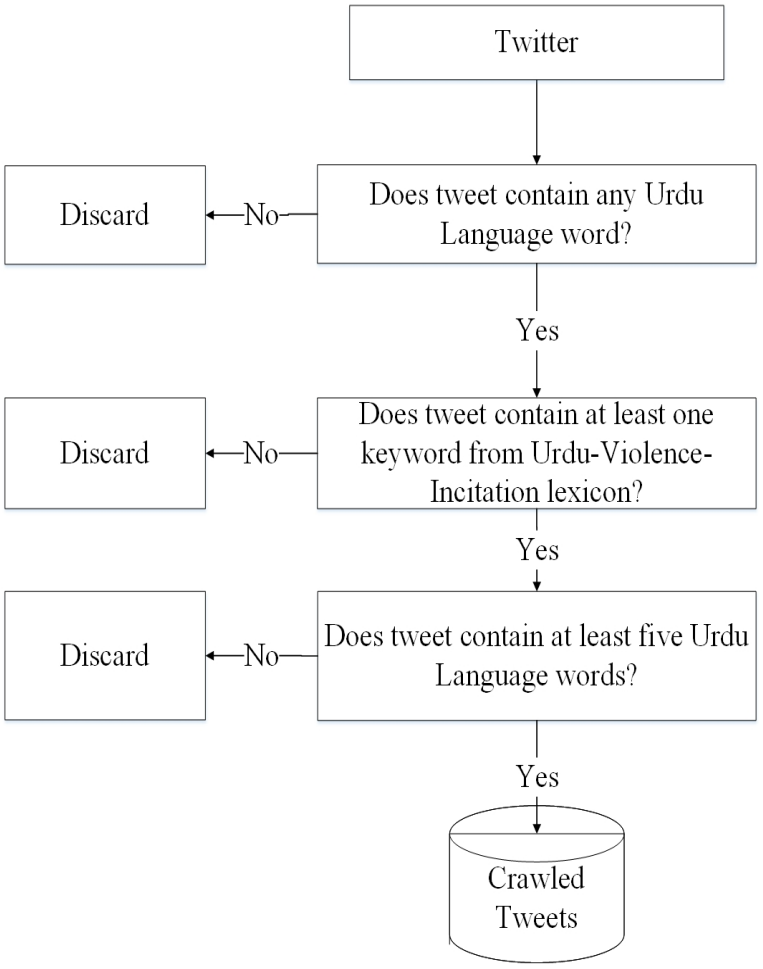


FIGURE 3.1: Data acquisition process

for accurate interpretation. Furthermore, research has shown that imposing a minimal word limit helps to produce a dataset that is more representative and diverse. Additionally, short messages frequently lack the background information needed to understand the language's genuine meaning. When analyzing dataset, it was found that tweets with three to four were not coherent enough to finish the task's context in Urdu. Consequently, we limited the word count in tweets to a maximum of five, and this produced a dataset that was representative. In addition, annotators were provided with enough linguistic content to assess the context and content, which could result in annotations that are more consistent.

3.2.5 Cleaning Dataset

In the subsequent phase of research, all the collected tweets were consolidated into a single comprehensive database stored in an Excel format. This consolidation step allowed for efficient management and organization of the vast amount of data

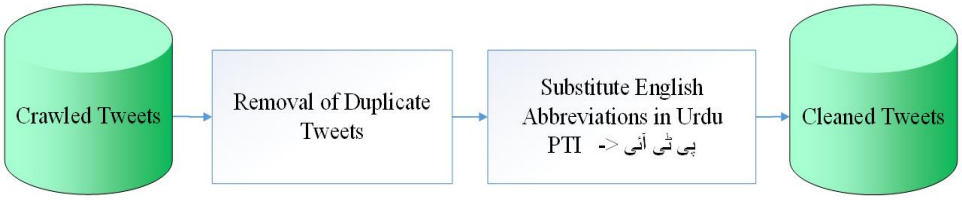


FIGURE 3.2: Cleaning process for tweets

we had gathered from various sources on the X platform. Following the consolidation process, we initiated a rigorous cleaning procedure to enhance the quality and readability of the data [41]. The culmination of these efforts yielded a cleaned dataset comprising tweets with enhanced readability and linguistic accuracy, providing a solid foundation for subsequent annotation and analysis tasks by team of data annotators [42]. Figure 3.2 illustrates the data cleaning process.

3.2.5.1 Preprocess Special Symbols

This cleaning process involved the removal of unnecessary characters, such as special symbols ($, %, &), which could potentially impede the understanding of the tweet content. Additionally, we implemented word correction techniques to rectify any spelling errors or typos present in the tweets, ensuring accuracy and coherence in the text.

3.2.5.2 Preprocess English Punctuations

During the dataset cleaning process, regular expressions were employed to remove English punctuation from the text. Punctuation marks B.1 such as (commas ,), (periods .), (exclamation points !), (question marks ?), (semicolon :), (semicolon ;), (dash —), (hyphen -), (apostrophe ’), (ellipsis …) and (quotation marks ””) were identified using predefined regular expression patterns. These patterns were designed to match specific characters or sequences of characters that represent punctuation in English text. Once identified, the regular expressions were applied to the dataset, systematically replacing or removing the punctuation marks from the text. This step was essential to ensure that the text remained readable and free from unnecessary noise or clutter caused by punctuation. By employing regular expressions for this purpose, the cleaning process was streamlined and automated, enhancing the overall efficiency and accuracy of the dataset preparation.

FIGURE 3.3: Arabic Numbers

3.2.5.3 Preprocess Urdu Punctuations

During the dataset cleaning process, Urdu punctuation were removed from the text using predefined rules and patterns. Regular expressions were employed to identify and replace Urdu punctuation marks A.1 such as (comma ،), (full stop ۔), (question mark ؟), and (semicolon ؛). These patterns were designed to match specific characters or sequences of characters representing Urdu punctuation. Once identified, the regular expressions were applied to the dataset, systematically removing the Urdu punctuation from the text. This step ensured that the text remained clean and devoid of unnecessary noise caused by punctuation marks, facilitating subsequent analysis and processing of the dataset.

3.2.5.4 Preprocess Arabic Characters

During the dataset cleaning process, Arabic numbers, Symbols and special characters C.1 were removed from the text to standardize the data and facilitate further analysis. Regular expressions were used to identify and replace Arabic numerals as shown in fig 3.3 with an empty string, effectively removing them from the dataset. This step ensured that the text consisted only of textual content without numerical digits, which could potentially interfere with language processing tasks such as sentiment analysis or topic modeling. By eliminating Arabic numbers, the dataset was prepared for linguistic analysis and modeling focused solely on textual information.

3.2.5.5 Preprocess English Characters and Numbers

During the dataset cleaning process, English characters and numbers were removed to ensure that the text consisted only of Urdu content, thereby facilitating further analysis specific to the Urdu language. Regular expressions were utilized to identify and replace English characters (a-z, A-Z) and numerical digits (0-9)

with an empty string. This step helped to eliminate any non-Urdu content and standardize the dataset for subsequent natural language processing tasks, such as sentiment analysis or text classification, that are tailored to Urdu language processing. By removing English characters and numbers, the dataset was effectively prepared for focused analysis and modeling within the context of Urdu language.

3.2.5.6 Preprocess Tab Characters

During the dataset cleaning process, tab characters (`\t`) were removed to ensure uniformity and consistency in the formatting of the text. Tab characters, which are often used for indentation or spacing purposes, can introduce unnecessary variation and disrupt the analysis of textual data. By employing appropriate text processing techniques, such as regular expressions, tab characters were identified and replaced with spaces or removed entirely. This step helped to standardize the dataset and prepare it for subsequent analysis or modeling tasks, ensuring that the text data was clean and well-structured for effective processing.

3.2.5.7 Preprocess New-Line Characters

During the dataset cleaning process, new-line characters (`\n`) were removed to ensure the text data was properly formatted and consistent. New-line characters, often represented in programming languages, can introduce unnecessary line breaks within the text, which may disrupt the analysis or modeling tasks. By employing text processing techniques such as regular expressions, new-line characters were identified and replaced with spaces or removed entirely. This step helped to streamline the text data, ensuring that it was clean, well-structured, and ready for further analysis or processing.

3.2.5.8 Preprocess Emoticons, Symbols, Pictographs and Scripts

During the dataset cleaning process, emoticons, symbols, pictographs, superscripts, and subscripts were removed to ensure that the text data remained focused on the linguistic content without any distractions from non-textual elements. Emoticons such as smiley faces or hearts, symbols like currency signs or arrows, and pictographs as shown in figures 3.4 and 3.5 representing objects or

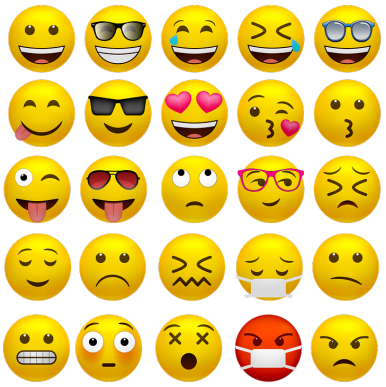


FIGURE 3.4: Emoticons

FIGURE 3.5: Symbols and pictographs

concepts were identified using regular expressions and filtered out from the text. Similarly, superscripts and subscripts, which are often used for mathematical or chemical notations, were also eliminated to maintain the integrity of the textual content. This cleaning step helped to standardize the dataset and ensure that only relevant linguistic information was retained for subsequent analysis or modeling tasks.

3.2.5.9 Mapping of Wrong Urdu Characters to Correct Urdu Characters

During the dataset cleaning process, incorrect Urdu characters were mapped to their correct counterparts to ensure the accuracy and consistency of the text data. This involved identifying common typographical errors or variations in Urdu characters and replacing them with their appropriate forms. For example, if a character was incorrectly typed or represented in a different font style, it was mapped to the correct Urdu character based on linguistic rules and conventions. This mapping helped to standardize the text and mitigate any inconsistencies caused by

typographical errors, ensuring that the dataset remained accurate and suitable for subsequent analysis or processing.

3.2.5.10 Fixing of Joined Words

During the dataset cleaning process, special attention was given to fixing joined words, which occur when there is no space between two consecutive words. This issue often arises due to typing errors or improper segmentation of text. To address this, techniques such as word segmentation algorithms or language-specific rules were applied to identify and separate joined words into their correct individual components. By fixing joined words, the readability and accuracy of the text data were improved, ensuring that subsequent analysis or processing steps could be performed effectively. For example, نہیں20سال has spaces issues and after normalizing text looks like نہیں 20 سال.

3.2.5.11 Substitute English Abbreviations with Urdu Abbreviations

During the dataset cleaning process, one crucial step involved substituting English abbreviations with their corresponding Urdu abbreviations. This was necessary to ensure consistency and clarity in the Urdu text data. English abbreviations, such as "PM" for Prime Minister or "CM" for Chief Minister, were replaced with their Urdu counterparts, such as وزیر اعظم for Prime Minister and وزیر اعلیٰ for Chief Minister. By making these substitutions, the text data became more linguistically accurate and aligned with Urdu language conventions, facilitating better understanding and analysis of the content. An Urdu-English substitutions list is illustrated in Table 3.3.

We noticed a few problems with the data that had been crawled. Initially, the retweeting feature results in duplicate tweets. Secondly, several Urdu tweets contain English abbreviations such as PTI, COAS, IK, and IMF, which need to be substituted with their respective Urdu words. We used the two procedures of data cleaning to address these problems. First, deleted the duplicate tweets. Secondly, fixed the English abbreviation problem. Furthermore, we devised a system for the substitution of self-developed abbreviations, such as PTI (Pakistan Tehreek-e-Insaf), PM (Prime Minister), CM (Chief Minister), and others, with their corresponding

TABLE 3.3: List of English abbreviations replaced by their counter-Urdu parts.

Urdu	**English**	**Urdu**	**English**
پی ٹی آئی	PTI	ٹی ایل پی	TLP
پی ایم ایل این	PMLN	پی ایم ایل	PML
پی ایم ایل کیو	PMLQ	ایم کیو ایم	MQM
پی پی پی	PPP	حکومت	Govt
فوج	Army	آئی کے	IK
این ایس	NS	ایس ایس	SS
بی بی	BB	ایم ایم اے	MMA
آئی جی	IG	آئی ایم ایف	IMF
ٹی ٹی پی	TTP	یو ایس اے	USA
اے این پی	ANP	بی اے پی	BAP
بی ایل اے	BLA	آئی ایس آئی	ISI
را	RAW	سی آئی اے	CIA
اسلام آباد	Islamabad	لاہور	Lahore
کراچی	Karachi	پشاور	Peshawar
کے پی کے	KPK	اے جے کے	AJK
آئی سی ٹی	ICT	وزیر اعظم	Prime Minister
صدر	President	سی او اے ایس	COAS
ڈی جی	DG	آئی ایس پی آر	ISPR

Urdu equivalents. This substitution facilitated better comprehension of the tweet content by replacing English abbreviations with their Urdu counterparts, thereby aligning with the linguistic preferences of target audience. We substituted most appropriate Urdu parts for English abbreviations. In order to accomplish this, we created Urdu-English substitutions list, where each English abbreviation has an equivalent Urdu part.

3.2.5.12 Combine White Spaces

During the dataset cleaning process, one important step was to combine multiple consecutive white spaces into a single space as shown in 3.6. This was necessary to standardize the formatting and ensure uniformity in the text data. Sometimes, during the crawling process or due to other factors, extra white spaces may be present between words or sentences. By consolidating these white spaces, we were able to improve the readability of the text and make subsequent processing steps more efficient. Additionally, removing redundant white spaces helped streamline

Processed	Unprocessed
عراق اور شام اعلان کیا ہے دونوں جلد اپنے گے؟	عراق اور شام اعلان کیا ہے دونوں جلد اپنے گے؟
وہ ہے ہی اس قابل جو گستاخوں کو رہا کرے قادیانیوں کو وزیر خزانہ لگانے کی کوشش کرے اسرائیل سے فنڈنگ لے	وہ ہے ہی اس قابل جو گستاخوں کو رہا کرے قادیانیوں کو وزیر خزانہ لگانے کی کوشش کرے اسرائیل سے فنڈنگ لے

FIGURE 3.6: Examples for combining white spaces of Urdu characters

the dataset and reduce its overall size, making it more manageable for further analysis.

3.2.6 Data Annotation

In selecting three annotators for the data annotation task, we prioritized a diverse range of expertise and perspectives to ensure comprehensive and accurate annotations. Each annotator was chosen based on their proficiency in the Urdu language, with all three possessing a strong command of Urdu as native speakers. Additionally, one annotator was specifically selected for their specialization in Urdu language as a senior Urdu teacher, bringing in-depth linguistic knowledge and expertise to the annotation process. Furthermore, all three annotators were graduates with a solid understanding of the Urdu language, providing a solid foundation for the task at hand. The decision to opt for three annotators was grounded in the principle of inter-rater reliability, which emphasizes the importance of multiple independent assessments to enhance the validity and reliability of the annotations [43]. By having three annotators independently annotate the dataset, we aimed to mitigate the risk of individual bias or subjectivity, thus increasing the robustness and credibility of the annotated data. Moreover, employing three annotators allowed for the identification and resolution of discrepancies or inconsistencies in the annotations through consensus-building discussions.

To facilitate the annotation process, we meticulously prepared detailed guidelines outlining the criteria and standards for identifying instances of violence incitation in Urdu content. These guidelines served as a reference point for the annotators, providing clear instructions on what constitutes violence incitation and how to annotate accordingly. Additionally, we conducted frequent training sessions with the annotators to ensure a shared understanding of the concept of violence incitation and to familiarize them with real-world examples drawn from the dataset.

TABLE 3.4: A few samples of the dataset.

No	English	Urdu
	Violence Incitation (YES)	
1	Send the Dutch blasphemer bastard Geert Wilders to hell	ہالینڈ کے گستاخ کنجر گیرٹ ویلڈر کو جہنم واصل کردو
2	The cure for these people is a revolution similar to what happened in Iran and what happened to Mujibur Rahman's family in Bangladesh, the same should happen to all of them. Enter their homes and send them to hell/kill them.	ان لوگوں کا علاج انقلاب ہے جیسے ایران میں ہوا اور جیسے بنگلا دیش میں مجیب الرحمان کے خاندان کے ساتھ ہوا ان سب کے ساتھ بھی وہی ہونا چاہیے گھر میں گھس کر ان لوگوں کو جہنم واصل کردو
3	Those who interfere/clash with our religion will be hanged from the hook of this crane, God willing/Insha Allah	دین سے پنگا لینے والوں کو اس کرین کے کنڈے سے لٹکایا جائے گا ان شاءاللہ
	Violence Incitation (NO)	
4	Let the law of Allah be established, you may take our lives, let the hijab remain, the cry of Muslim girls in India, O Allah, awaken the religious pride and honor among the Muslims, Amen.	اللہ کی قانون کو آباد رہنے دو بیشک آپ ہماری جان لےلو حجاب رہنے دو انڈیا میں بنت ہوا کی فریاد یااللہ مسلمانوں میں دینی حمیت اور غیرت بیدار فرما آمین
5	Saba Ramadan The individual humanity within us is dead which can collectively only give rise to tragedies. Execute someone for committing a small theft while lick the boots of the bigger thieves.	صبا رمضان ہمارے اندر کی انفرادی انسانیت مر چکی ہے جو کہ اجتماعی طور پر صرف سانحات ہی جنم دے سکتی ہے چھوٹی چوری پہ جان لیلو بڑے چوروں کے جوتے چاٹو

These training sessions fostered a collaborative environment wherein annotators could ask questions, seek clarification, and refine their annotation skills, thereby enhancing the quality and consistency of the annotations across the dataset. Overall, the strategic approach to data annotation involving three annotators, detailed guidelines, and comprehensive training sessions laid the groundwork for generating high-quality annotated data essential for this research on violence incitation detection [44–47]. Readers can find more examples of both non-violence and violence incitation in Table 3.4.

3.2.6.1 Annotator Selection Criteria

Three Pakistani annotators were engaged to help in annotating this corpus. Goal was to increase productivity while minimizing annotation error. Annotators have to meet the following requirements in order to be chosen:

a) Possess previous expertise annotating data

b) Speak Urdu fluently being native Urdu speaker

c) Have at least a bachelor's degree in education or HIGHER

d) Have a viewpoint that is balanced on politics

3.2.6.2 Inter-annotator Agreement

Tweets were given to the annotators along with the annotation rules and guidelines. In the experimental scenario, we were concerned in a balanced dataset. Majority voting was used for dataset preparation because it provides a reliable and unbiased way to resolve disagreements among annotators. Since individual annotators may bring subjective variations or occasional errors, majority voting ensures that the final label reflects the most agreed-upon interpretation. This not only reduces noise in the dataset but also enhances the consistency and quality of the annotations, making the dataset more representative and robust for training and evaluation purposes. During data annotation process, Cohen's kappa coefficient was chosen to measure the inter-annotator agreement because of its suitability for assessing the agreement between multiple annotators when dealing with categorical data, such as the classification of tweets into different categories of violence incitation. Cohen's kappa coefficient provides a more robust evaluation of agreement beyond simple chance agreement, accounting for the possibility of agreement occurring by random chance alone .

This is particularly important in this case, where multiple annotators independently classify tweets, and there may be inherent ambiguity or subjectivity in determining whether a tweet constitutes violence incitation. Compared to other measures of inter-annotator agreement, such as simple percentage agreement or Fleiss' kappa coefficient, Cohen's kappa coefficient offers several advantages. While simple percentage agreement provides a straightforward measure of agreement, it does not consider the possibility of agreement occurring by chance and may therefore overestimate agreement when dealing with categorical data with more than two categories [48–51]. On the other hand, Fleiss' kappa coefficient is suitable for assessing agreement among multiple raters but is less appropriate when dealing with

TABLE 3.5: Pair-wise Cohen Kappa agreement between three annotators.

Rater 1	**Rater 2**	**Cohen Kappa** (%)
Annotator 1	Annotator 2	89.27
Annotator 1	Annotator 3	90.26
Annotator 2	Annotator 3	95.00

only three annotators, as it is designed for more than two raters. Additionally, Fleiss' kappa assumes that each rater is randomly selected from a population of raters, which may not always hold true in practice.

Cohen's kappa coefficient addresses these limitations by explicitly accounting for chance agreement and providing a more nuanced assessment of inter-annotator agreement that is robust to the number of annotators. By using Cohen's kappa coefficient in during dataset annotation process, it helped to obtain a more accurate and reliable measure of agreement between annotators, enabling this research to assess the consistency and reliability of the annotations and identify any discrepancies that may require further review or clarification [52–55]. Overall, Cohen's kappa coefficient offers a comprehensive and statistically sound approach to evaluating inter-annotator agreement in data annotation task, making it the preferred choice for assessing agreement among multiple annotators in this research on violence incitation detection. In this case, the inter-rater agreement was determined using two measures. Cohen's kappa was employed to gauge the degree of agreement between two annotators [56]. Three annotators' pair-wise Cohen Kappa agreement is shown in Table 3.4. It is evident that any two annotators have agreement that is at least 89 % of the time, which is almost perfect agreement. The other metric used to gauge the degree of agreement among the three annotators was the Fleiss kappa [57], which had a score of 91.50 %. This cutoff point shows that there is excellent agreement among the annotators, as shown in Table 3.5.

3.2.7 Violence Incitation Dataset

After the data annotation process, dataset consisted of a total of 4804 tweets, with 2402 labeled as "Yes" for violence incitation and an equal number of 2402 labeled as "No." However, it was observed an imbalance between the two classes,

TABLE 3.6: Violence incitation dataset

Total Size	**"Yes" Label**	**"No" Label**
4804	2402	2402

which could potentially affect the performance of machine learning models. To address this imbalance, we employed dataset balancing techniques by augmenting the dataset with additional non-violence incitation tweets. One such technique involves oversampling the minority class, where we added more non-violence incitation tweets to achieve a balanced distribution of the two classes as shown in Table 3.6. By balancing the dataset in this manner, we ensured that machine learning models were trained on a more representative and equitable set of data, which ultimately enhances their ability to accurately classify tweets as either violence incitation or non-violence incitation.

3.2.8 Violence Incitation Target Groups

In the development of second dataset aimed at identifying violence-incitation targeted groups, we focused exclusively on tweets labeled as "Yes" for violence incitation. With a dataset comprising 2402 tweets in this category, main objective was to discern and categorize the targeted groups. After thorough analysis, we delineated three distinct groups: Government/Religious, Political, and General. Each group represented a different facet of society that could potentially be targeted by violence-inciting content. To ensure consistency and accuracy in the annotation process, we provided clear definitions for each group and distributed these guidelines to team of annotators. Their task was to manually annotate the dataset, assigning each tweet to one of the three designated classes based on the identified targeted group. This meticulous approach allowed us to create a comprehensive dataset that accurately reflects the diverse nature of violence-incitation targeted groups, facilitating more nuanced analysis and interpretation of the data. The dataset used in this research is relatively small, which limited the possibility of adding many fine-grained categories. The targeted communities/groups identified are intentionally kept abstract (e.g., religious communities, political groups, ethnic communities) for three main reasons:

3.2.8.1 Generalization Over Specific Entities

The focus was on detecting patterns of incitement against broader social groups rather than identifying specific individuals or highly detailed communities. This abstraction allows the system to generalize better and remain applicable across multiple contexts instead of being restricted to narrow entities.

3.2.8.2 Privacy and Ethical Concerns

Labeling highly specific groups or individuals could raise serious ethical and privacy issues, especially in the sensitive context of Urdu content involving violence and hate speech. Abstract categories help reduce the risk of unfair profiling.

3.2.8.3 Small Dataset Size

Due to the limited dataset, introducing many specific categories was not feasible, as it could lead to data sparsity and reduce model reliability. This combined approach ensured fairness, ethical responsibility, and practical applicability in dataset preparation and annotation.

3.2.9 Fairness and Validity of Dataset

While Cohen's Kappa was used to measure inter-annotator agreement, ensuring fairness and validity required additional steps:

3.2.9.1 Annotation Guidelines and Bias Reduction

a) Clear annotation guidelines were developed to minimize subjective bias.

b) Annotators were trained with multiple examples to ensure consistency.

c) Disagreement cases were discussed and resolved collaboratively, reducing annotator bias.

3.2.9.2 Data Distribution Analysis

a) The dataset was analyzed for class balance (violent vs. non-violent content) to ensure fair representation.

b) Over/under-representation of certain categories (e.g., political, religious, ethnic content) was addressed through careful sampling.

3.2.9.3 Fairness Metrics

a) Demographic parity checks were performed to ensure that specific communities or groups were not disproportionately labeled as violent.

b) Feature-level analysis was conducted to confirm that sensitive attributes (e.g., religion, ethnicity, gender references) were not driving classification unfairly.

3.2.9.4 Feature Correlation Analysis

Statistical correlation between lexical features and class labels was examined to avoid spurious correlations (e.g., words common in specific dialects being misclassified as incitement).

3.2.9.5 Cross-validation for Reliability

K-fold cross-validation was used to test robustness on different data partitions, ensuring that results were not dataset-specific.

3.2.9.6 External Validity Checks

a) Dataset was compared with existing violence-incitement definitions in English NLP literature to validate conceptual alignment.

b) Small-scale pilot tests with independent annotators confirmed that labels were interpretable and valid.

The fairness and validity of the dataset were ensured by combining inter-annotator agreement (Cohen's Kappa) with bias reduction measures, distribution analysis, fairness checks, and cross-validation. This holistic approach moves beyond raw agreement scores and demonstrates that the dataset is both reliable and equitable for violence incitation detection in Urdu.

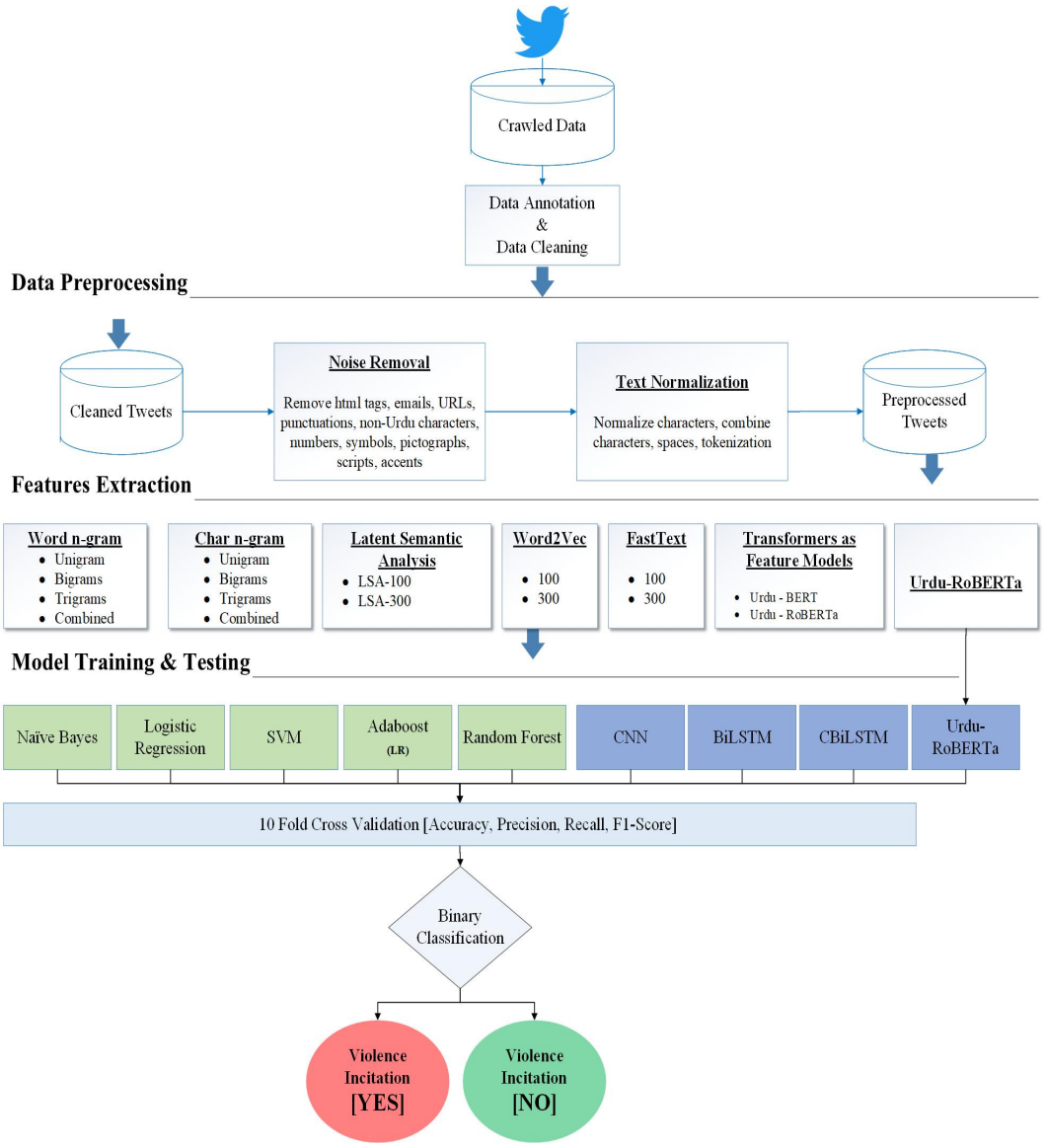


FIGURE 3.7: The pipeline of the proposed framework

3.3 Proposed Framework

This section outlined the suggested process for creating a binary classification system in Urdu that would identify instances of violence incitation. Figure 3.7 portrays the suggested framework at various stages, including feature extraction, pre-processing, training and testing of the models, and the classification stage.

3.3.1 NLP Pipeline

The pipeline we designed for violence incitation detection in Urdu is carefully structured to address both linguistic challenges of Urdu and the computational requirements of NLP tasks.

3.3.1.1 Data Collection and Preprocessing

a) Urdu text from Twitter is often noisy, with code-mixing (Urdu-English), spelling variations, and informal writing styles.

b) Preprocessing steps (tokenization, normalization, stop-word removal, handling diacritics, and transliteration handling) are justified because they reduce noise, improve token consistency, and make features more representative of semantic meaning.

3.3.1.2 Feature Extraction

a) Urdu has rich morphology (affixes, gender, plurals, case markers) unlike English. Hence, we applied feature extraction techniques to capture both lexical and semantic cues.

b) Contextual embeddings were added to handle context-dependent meaning that bag-of-words cannot capture.

3.3.1.3 Modeling

a) We evaluated both traditional ML models (SVM, Logistic Regression) and deep learning (LSTM, Transformer-based models).

b) ML models are efficient and robust on small datasets.

c) Transformer-based models leverage transfer learning from large multilingual corpora, which compensates for Urdu's low-resource nature.

3.3.1.4 Evaluation

a) We reported Accuracy, F1-score, precision, recall, AUC.

b) AUC in particular justifies performance across different thresholds, ensuring the pipeline is not biased toward one class.

3.3.1.5 Interpretability and Target Group Identification

a) The final stage of the pipeline identifies targeted communities/groups using lexical and context-based cues.

b) This extends beyond detection, making the pipeline socially impactful by revealing who is being targeted by the incitement.

3.4 Data Preprocessing

Noise, extraneous information, and inconsistencies are among the many problems that typically plague the databases. The aforementioned Figure 3.7 provides more information on how the pre-processing methodology was used to address these problems on the annotated dataset. This approach involved two stages. Firstly, the dataset underwent a noise removal method to exclude any extraneous elements from the tweet text, including HTML tags, URLs, symbols, punctuation, numbers, currency symbols, phone numbers, and emails. Subsequently, the dataset underwent the normalizing procedure in order to get it ready for feature engineering. Tokenization was followed by the normalization of characters, combined characters, and whitespaces as sub-steps in the normalization process.

3.4.1 Preprocess Noise

3.4.1.1 Preprocess URLs

During the data preprocessing stage, the removal of URLs was facilitated by the ”remove urls” function of the Urduhack library. This function played a crucial role in eliminating any URLs present in the text data, such as web links or hyperlinks, which are often irrelevant to the analysis and may introduce noise into the dataset as shown in fig 3.8. By using this function, we were able to strip away URLs from the text, ensuring that the dataset contained only relevant information for subsequent processing steps. Removing URLs also helped in improving the accuracy of analysis by focusing solely on the textual content of the tweets or documents, without the distraction of external links.

3.4.1.2 Preprocess Emails

The ”remove emails” function of the Urduhack library played a significant role in the noise removal process during data preprocessing. This function effectively

FIGURE 3.8: Example of URL in Tweet

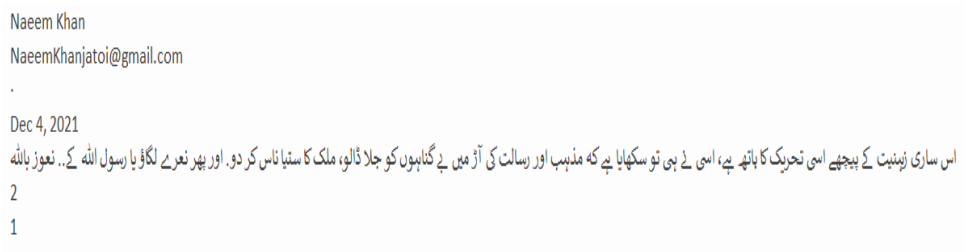


FIGURE 3.9: Example of Email in Tweet

eliminated any email addresses present in the text data, which are often irrelevant to the analysis and may introduce unnecessary noise like 3.9. By using this function, it ensured that dataset contained only relevant textual information, devoid of email addresses that could potentially skew the analysis or compromise privacy. Removing emails also helped streamline the preprocessing pipeline, allowing us to focus solely on the linguistic content of the dataset without distractions from personal or contact information.

3.4.1.3 Preprocess Phone Numbers

The "remove phone numbers" function of the Urduhack library proved to be invaluable during the noise removal process in data preprocessing. This function efficiently eliminated any phone numbers present in the text data, which are often irrelevant for linguistic analysis and may introduce unnecessary noise. By utilizing this function, we ensured that dataset contained only relevant textual information, free from phone numbers that could potentially distort the analysis or compromise privacy. Removing phone numbers also streamlined the preprocessing pipeline, allowing us to focus exclusively on the linguistic content of the dataset without distractions from contact details.

3.4.1.4 Preprocess Numbers

The ”remove numbers” function of the Urduhack library played a crucial role in the noise removal process during data preprocessing. This function effectively removed all numerical digits present in the text data, ensuring that only linguistic content remained for analysis. By eliminating numbers, we enhanced the quality of the dataset and facilitated more accurate linguistic analysis. Removing numerical digits is particularly important in NLP tasks focused on textual understanding, as numbers often do not contribute to the linguistic context and may introduce unnecessary noise. Thus, the ”remove numbers” function helped streamline the preprocessing pipeline and improve the overall quality of the dataset for subsequent analysis.

3.4.1.5 Preprocess Currency Symbols

The ”remove currency symbols” function of the Urduhack library proved instrumental in eliminating currency symbols B.1 from the text data during the noise removal process in data preprocessing. This function effectively stripped away any currency symbols present in the text, ensuring that the linguistic content remained free from non-linguistic elements related to currency representation. By removing currency symbols, we aimed to focus solely on the linguistic aspects of the text for subsequent analysis, enhancing the quality and accuracy of dataset. This step is particularly crucial in NLP tasks where the presence of currency symbols may introduce unnecessary noise and hinder the understanding of the text’s linguistic context. Thus, the ”remove currency symbols” function helped streamline the data preprocessing pipeline and improve the overall quality of the dataset for further analysis and modeling.

3.4.1.6 Preprocess Punctuations

The ”remove punctuations” function of the Urduhack library played a pivotal role in eliminating punctuations B.1 from the text data during the noise removal process in data preprocessing. This function effectively stripped away all punctuation marks present in the text, including commas, periods, question marks, exclamation marks, and others. By removing punctuations, we aimed to streamline the

text and focus solely on the linguistic content, thereby enhancing the quality and clarity of the dataset. Removing punctuations is essential in NLP tasks as it helps in standardizing the text and ensuring consistency across different documents. Additionally, it facilitates subsequent text processing steps such as tokenization and feature extraction. Overall, the ”remove punctuations” function contributed to the refinement of the dataset and improved the effectiveness of subsequent NLP tasks.

3.4.1.7 Preprocess Diacritics

The ”remove diacritics” function in the Urduhack library played a crucial role in eliminating diacritics B.1 from the text data during the noise removal process in data preprocessing. Diacritics are small marks or symbols added to letters to indicate pronunciation or change their meaning. In Urdu text, diacritics are commonly used to represent vowels and distinguish between similar-looking characters. However, in certain cases, diacritics may introduce noise or inconsistencies in the text, especially when processing large datasets for NLP tasks. Therefore, the ”remove diacritics” function efficiently removed all diacritics from the text, ensuring uniformity and enhancing the readability of the dataset. By eliminating diacritics, we aimed to simplify the text while preserving its semantic content, thereby facilitating subsequent text analysis and processing tasks. Overall, the ”remove diacritics” function contributed to improving the quality and consistency of the dataset, ultimately enhancing the performance of NLP models and algorithms.

3.4.2 Text Normalization

3.4.2.1 Normalize Whitespace

The ”normalize whitespace” function in the Urduhack library played a crucial role in standardizing whitespace characters within the text data during the noise removal process in data preprocessing. In textual data, whitespace refers to any sequence of space, tab, or newline characters that separates words or other elements. However, inconsistencies in whitespace usage can occur due to factors such as data

collection methods or input errors, leading to irregularities in the text. The ”normalize whitespace” function effectively replaced multiple consecutive whitespace characters with a single space character, ensuring uniform spacing throughout the text. By standardizing whitespace, the function improved the overall readability and consistency of the dataset, making it easier to process and analyze using natural language processing techniques. Additionally, the normalization of whitespace helped mitigate potential issues related to feature extraction and model training, ultimately contributing to the robustness and accuracy of NLP applications.

3.4.2.2 Normalize Characters

The ”normalize characters” function in the Urduhack library played a pivotal role in standardizing characters within the text data during the noise removal process in data preprocessing. Textual data may contain various forms of characters, including different representations of the same character due to encoding inconsistencies, typographical errors, or input variations. The ”normalize characters” function addressed these issues by converting characters to their standardized forms, ensuring uniformity and consistency across the dataset. It replaced characters with diacritics, ligatures, or other variants with their base forms, facilitating better readability and reducing ambiguity in the text. By normalizing characters, the function enhanced the quality of the dataset and improved the performance of subsequent natural language processing tasks, such as tokenization, part-of-speech tagging, and sentiment analysis. Overall, the normalization of characters contributed to the reliability and accuracy of NLP applications by eliminating inconsistencies and enhancing the processing efficiency of textual data.

3.4.2.3 Normalize Combine Characters

The ”normalize combine characters” function in the Urduhack library played a crucial role in streamlining textual data during the noise removal process in data preprocessing. Textual data often contains combined characters, ligatures, or diacritics, which can introduce inconsistencies and complexity in analysis. The ”normalize combine characters” function addressed this by decomposing combined characters into their individual components, separating ligatures, diacritics, and other

combined forms into their base forms. This process helped standardize the representation of characters, ensuring uniformity and consistency across the dataset. By normalizing combined characters, the function improved the readability of the text and facilitated subsequent natural language processing tasks, such as tokenization and feature extraction. Overall, the normalization of combined characters enhanced the quality and reliability of the dataset, contributing to more accurate and efficient NLP applications. In the following string, Alif ('ا') and Hamza ('ٔ') are separate characters "جرأت", but after normalizing this text, now Alif and Hamza are replaced by a Single Urdu Unicode Character "جرأت".

3.4.2.4 Preprocess Stop Words

The "preprocess stop words" function in the Urduhack and Spacy libraries serves as a vital component in the noise removal process during data preprocessing. Stop words are common words that occur frequently in a language but typically do not carry significant meaning in the context of natural language processing tasks. Examples include articles, prepositions, and conjunctions. By removing stop words from textual data, the function helps streamline the dataset and focus on the essential content. This process reduces noise and improves the efficiency of downstream NLP tasks, such as sentiment analysis, topic modeling, and text classification. By eliminating irrelevant words, the function enhances the accuracy and effectiveness of NLP models by allowing them to concentrate on the most informative features of the text.

3.5 Features Extraction

Seven different feature types (word n-gram, char n-gram, LSA, word2vec, FastText, Urdu-BERT, and Urdu-RoBERTa models) were extracted by us following pre-processing, and their impacts were examined with respect to the binary categorization of violence incitation challenge.

3.5.1 N-gram Models

Word n-grams and character n-grams are essential features in natural language processing (NLP) tasks, providing valuable insights into the structure and semantics of text data. These features play a crucial role in various NLP applications, including text classification [58], sentiment analysis, and language modeling. Word n-grams represent sequences of words in a document, capturing the co-occurrence patterns and relationships between words. On the other hand, character n-grams represent sequences of characters, providing information about the morphology and syntax of the text. Word n-grams, such as unigrams, bigrams, and trigrams, offer a comprehensive view of the vocabulary and language patterns present in the dataset. Unigrams represent individual words, capturing the most basic units of meaning.

Bigrams and trigrams capture pairs and triplets of adjacent words, respectively, revealing more complex linguistic structures and associations. By considering sequences of words, word n-grams can capture contextual information and dependencies between words, enhancing the understanding of the text's semantics and syntax. In the provided dataset, the counts of word n-grams reflect the diversity and richness of language patterns present in the text data. The high counts of unigrams (12,253), bigrams (72,453), and trigrams (97,805) indicate the presence of a wide range of vocabulary and language constructs. The combined count of 1-2-3 grams (182,511) further amplifies the coverage of language patterns by including sequences of varying lengths. These features enable NLP models to capture nuanced linguistic nuances and improve performance in tasks such as text classification and information retrieval. Character n-grams provide complementary information to word n-grams by capturing sub-word structures and morphological variations in the text. Unigrams (71), bigrams (1,776), trigrams (20,304), and combined (22,151) of characters represent sequences of individual characters, character pairs, and character triplets, respectively. These features are particularly useful for handling out-of-vocabulary words, misspellings, and noisy text data. By considering character-level information, NLP models can handle unseen words and improve robustness in tasks such as text normalization and spell checking.

In summary, word n-grams and character n-grams are fundamental features in

TABLE 3.7: Detail of features.

Feature Type	**Features**	**Count**
Word n-gram (BOW)	Unigram	12,253
	Bigrams	72,453
	Trigrams	97,805
	Combined 1-2-3 grams	182,511
Char n-gram	Unigram	71
	Bigrams	1,776
	Trigrams	20,304
	Combined 1-2-3 grams	22,151
Latent Semantic Analysis	LSA-100	100
	LSA-300	300
Word2Vec	CBOW-100	100
	CBOW-300	300
FastText	CBOW-100	100
	CBOW-300	300
Urdu-BERT	Transformer Model	768
	Transformer Model	768

NLP, offering valuable insights into the structure and content of text data. These features enable NLP models to capture language patterns, semantic relationships, and syntactic structures, thereby enhancing performance in various text processing tasks. The diverse range of features present in the dataset provides a rich source of information for training robust and effective NLP models. Many applications of natural language processing (NLP), like text classification, word and character n-grams are valuable attributes [3]. The same has been witnessed in case of sentiment analysis [5]. The purpose of this dissertation was to identify violence incitation's using word uni-grams, bi-grams, and tri-grams as well as their hybrid combinations. Furthermore, the dissertation examined char uni-gram, bi-grams, and tri-grams as well as their hybrid mixes. Table 3.7 displays the total number of features produced for each model.

3.5.2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a powerful technique in natural language processing (NLP) for extracting underlying semantic relationships between words and documents [59]. It operates by creating a mathematical representation of text data

in a high-dimensional space, where words and documents are mapped to vectors. LSA then applies dimensionality reduction techniques to identify latent semantic structures, enabling the capture of semantic similarities and associations between words and documents. In the context of violence incitation identification, LSA features play a crucial role in capturing the underlying semantic relationships present in the text data. By representing text documents in a semantic space, LSA can uncover hidden thematic structures and semantic patterns that may be indicative of violence incitation. For example, LSA can identify subtle semantic similarities between documents containing violent rhetoric or extremist language, even if they do not share exact word overlap.

The importance of LSA features lies in their ability to capture the semantic context and latent meaning embedded within text data. Unlike traditional bag-of-words approaches, which focus solely on word frequencies and co-occurrence patterns, LSA considers the semantic relationships between words and documents. This enables LSA to identify semantically related terms and documents, even in the absence of exact word matches. LSA features with dimensions of 100 and 300 provide a compact yet informative representation of the semantic space, capturing the most salient semantic relationships while reducing the dimensionality of the data. These lower-dimensional representations facilitate efficient modeling and analysis of large text datasets, making LSA features particularly suitable for violence incitation identification tasks where scalability and computational efficiency are crucial.

3.5.3 Word2Vec

Word2Vec is a widely used technique in natural language processing (NLP) for generating word embeddings, which are dense vector representations of words in a continuous semantic space. These embeddings capture semantic relationships between words based on their distributional properties in a large corpus of text. Word2Vec models, such as those trained using the Continuous Bag of Words (CBOW) configuration, learn to predict the context words surrounding a target word, thereby capturing the contextual meaning and semantic similarity between words. In the context of violence incitation identification, Word2Vec features play

a crucial role in capturing the underlying semantic structure and contextual meaning of text data. By representing words as dense vectors in a continuous semantic space, Word2Vec embeddings encode semantic relationships and similarities between words, allowing for the detection of subtle linguistic nuances and semantic patterns indicative of violence incitation. The importance of Word2Vec features lies in their ability to capture semantic context and capture subtle semantic relationships between words, even in the absence of exact word overlap.

By leveraging the semantic information encoded in Word2Vec embeddings, violence incitation detection models can effectively identify semantically related terms and documents associated with violent rhetoric or extremist language. Word2Vec features with dimensions of 100 and 300 provide a compact yet informative representation of the semantic space, capturing the most salient semantic relationships while reducing the dimensionality of the data. These lower-dimensional representations facilitate efficient modeling and analysis of large text datasets, making Word2Vec features particularly suitable for violence incitation identification tasks where scalability and computational efficiency are crucial. In text mining and natural language processing tasks, the word embedding model Word2vec demonstrated state-of-the-art performance [60]. To investigate the effect of word2vec on violence incitation identification, we experimented with CBOW configuration as well as 100 and 300 dimensions.

3.5.4 FastText

FastText is an extension of Word2Vec that introduces subword information into word embeddings, making it particularly powerful for handling out-of-vocabulary words and capturing morphological variations in languages [61]. By considering subword information, such as character n-grams, FastText embeddings can effectively represent words with similar morphological structures even if they are not explicitly present in the training corpus. This capability is especially beneficial for violence incitation identification tasks, where extremist language and violent rhetoric may involve novel or morphologically complex words that are not commonly encountered in standard vocabulary. In violence incitation identification, FastText features offer several advantages. Firstly, they enable the modeling of

complex linguistic phenomena and morphological variations, allowing for the effective capture of subtle semantic nuances and linguistic patterns associated with violent or extremist content.

By representing words as a combination of character n-grams, FastText embeddings can encode morphological information and capture similarities between words based on their shared subword components, thereby enhancing the discrimination between violent and non-violent text. Additionally, FastText features with dimensions of 100 and 300 provide a compact yet informative representation of the semantic space, facilitating efficient modeling and analysis of large text datasets. This enables violence incitation detection models to leverage subword information and capture semantic context, thereby improving the accuracy and robustness of violence incitation identification systems.

3.5.5 Urdu-BERT

Urdu-BERT, a variant of BERT specifically tailored for the Urdu language, offers a transformative approach to violence incitation identification by leveraging contextualized word representations learned from large-scale unlabeled Urdu text corpora. By pretraining on a vast amount of diverse Urdu text data, Urdu-BERT can effectively capture the intricate linguistic nuances and semantic relationships inherent in Urdu language expressions. This contextualized embedding approach enables Urdu-BERT to generate rich representations of text sequences that reflect the subtle variations in meaning and intent, making it particularly well-suited for discerning violent or extremist content from benign text. In the realm of violence incitation identification, Urdu-BERT features play a crucial role in encoding the contextual information necessary for understanding the underlying sentiment, rhetoric, and intent of the text. By contextualizing word embeddings based on the surrounding words in a sentence or document, Urdu-BERT can capture the complex linguistic structures and contextual cues that signal the presence of violent or extremist language.

Moreover, Urdu-BERT's ability to model bidirectional context allows it to capture long-range dependencies and syntactic relationships within text, enabling it to discern subtle nuances and detect nuanced forms of violence incitation that

may be obscured by traditional lexical or syntactic approaches. Overall, Urdu-BERT features provide a powerful tool for violence incitation identification, offering deep insights into the linguistic and semantic characteristics of extremist rhetoric in the Urdu language. An engineering team at Google Lab created the pre-trained transformer model BERT [62]. Due to its demonstrated efficacy in comparable text classification and sentiment analysis tasks, Urdu-BERT was selected for this assignment [63]. In addition, the language model outperforms the semantic and embedding models in capturing the context of the violent material in tweets. Therefore, in this case, we extracted features using Urdu-BERT.

3.5.6 Urdu-RoBERTa

Urdu-RoBERTa, an extension of the RoBERTa model specifically fine-tuned for the Urdu language, represents a significant advancement in violence incitation identification by leveraging state-of-the-art transformer-based architecture [64]. Trained on large-scale corpora of Urdu text, Urdu-RoBERTa is adept at capturing intricate linguistic nuances, syntactic structures, and contextual dependencies within Urdu language expressions. By employing a masked language modeling objective during pretraining, Urdu-RoBERTa learns to predict missing tokens within text sequences, thereby acquiring a rich understanding of the semantics and underlying patterns present in Urdu language data. This enables Urdu-RoBERTa to generate highly informative contextualized embeddings that encode nuanced features relevant to violence incitation identification. In the domain of violence incitation identification, Urdu-RoBERTa features offer unparalleled capabilities for capturing and understanding the subtle nuances of extremist rhetoric and violent language in Urdu text.

By encoding contextual information from surrounding words and sentences, Urdu-RoBERTa embeddings encapsulate the semantic meaning, sentiment, and intent conveyed by text, facilitating the detection of violence incitement with high accuracy. Moreover, Urdu-RoBERTa's robust architecture and deep learning capabilities enable it to discern nuanced forms of violence incitation, including implicit threats, coded language, and dog-whistle tactics, which may evade detection by traditional rule-based or lexical methods. Overall, Urdu-RoBERTa features serve

as a powerful tool for violence incitation identification, offering advanced linguistic insights and enhancing the efficacy of automated detection systems in combating online extremism and hate speech in the Urdu language. An additional transformer model that builds upon the BERT architecture is called RoBERTa [65]. In a number of NLP tasks, including those in the multilingual domain, it showed notable performance [66]. Consequently, we were keen to investigate Urdu-RoBERTa as a feature model and assess how well it worked for identifying tweets that incited violence by fine-tuning Urdu-RoBERTa.

3.5.7 Feature Extraction in Urdu NLP vs English NLP

Feature extraction in Urdu NLP is significantly different from English NLP due to fundamental differences in script, morphology, orthography, and linguistic resources:

3.5.7.1 Script and Orthography

a) Urdu is written in Perso-Arabic script (Nastaliq), which is right-to-left and has context-sensitive character shapes.

b) English uses a Latin script with relatively fixed character shapes.

c) This means Urdu requires special preprocessing steps such as Unicode normalization, diacritic handling, and script disambiguation, which are not as critical in English.

3.5.7.2 Morphological Richness

a) Urdu is a morphologically rich language (inflectional and derivational variations in verbs, nouns, and adjectives).

b) Example: a single Urdu word can have multiple inflected forms depending on gender, number, case, and tense.

c) English is relatively morphologically poor (limited inflection, mostly suffixes like -s, -ed, -ing).

d) Therefore, morphological analysis and stemming/lemmatization in Urdu are far more complex and less standardized.

3.5.7.3 Word Segmentation

a) Urdu lacks explicit word boundary markers in some cases (e.g., spaces are inconsistently used, compound words and affixes may be written with/without spaces).

b) English has clearer word segmentation rules based on whitespace.

c) Urdu NLP often requires rule-based tokenization, morphological analyzers, or character-level embeddings, unlike English where whitespace tokenization is usually sufficient.

3.5.7.4 Ambiguity and Context Dependence

a) Urdu exhibits high lexical and semantic ambiguity due to polysemy and homographs (same word written the same but meaning depends on context).

b) Example: ” ” (qatal) can mean ”murder” as a noun or ”to kill” as a verb.

c) English also has ambiguity, but Urdu's lack of large annotated corpora makes disambiguation harder.

d) This increases the importance of contextual embeddings (e.g., mBERT, XLM-R) in Urdu.

3.5.7.5 Resource Availability

a) English has well-established NLP tools (POS taggers, parsers, embeddings, WordNet, large corpora).

b) Urdu is a low-resource language with very limited pre-trained embeddings, lexicons, and annotated datasets.

c) Feature extraction in Urdu often requires cross-lingual transfer learning, multilingual embeddings, or manual lexicon construction, unlike English.

3.5.7.6 Character Encoding and Normalization Issues

a) Urdu faces Unicode inconsistencies (different code points for visually identical characters).

b) Example: ” ” (Yeh) has multiple Unicode representations.

c) English encoding is relatively standardized (ASCII/UTF-8).

d) Thus, feature extraction in Urdu requires character normalization to reduce noise.

3.6 Experimental Setup

In this section, we outlined the classification procedure and provided specifics on the ML and DL models that were utilized, along with assessment criteria that were employed to gauge the classifiers' performance.

3.6.1 Classification

The identification of violence incitation is the subject of this dissertation. It is a binary classification task, where ”Yes” represents the violence incitement existence in tweets and ”No” represents the absence of it.

3.6.2 Evaluation Methodology

The dataset was split into subsets for training and testing by utilizing 10-fold cross-validation approach. Dealing with the overfitting problem was another benefit of this strategy. Additionally, a set of five performance metrics was selected in order to assess classifier performance: the Area under the Curve (AUC), macro f1-score, recall, accuracy, and precision.

3.6.2.1 Accuracy

Accuracy measures the proportion of correctly classified instances (both true positives and true negatives) out of all instances in the dataset. It provides a general overview of the classifier's overall correctness in classification. Accuracy is easy

to interpret and widely used, especially when class distribution is balanced. However, accuracy may be misleading in the presence of class imbalance, where a high accuracy can be achieved by simply predicting the majority class.

3.6.2.2 Precision

Precision measures the proportion of true positive instances among all instances predicted as positive by the classifier. It quantifies the classifier's ability to avoid false positives, minimizing the incorrect classification of negative instances as positive. Precision is important in scenarios where false positives are costly or harmful, such as in legal or security applications. Higher precision values indicate better performance, with a value of 1 representing perfect precision.

3.6.2.3 Recall

Recall measures the proportion of true positive instances that are correctly identified by the classifier out of all actual positive instances. It quantifies the classifier's ability to capture all positive instances, minimizing false negatives. Recall is crucial in scenarios where missing positive instances (false negatives) are costly or detrimental, such as in medical screenings or fraud detection. Higher recall values indicate better performance, with a value of 1 representing perfect recall.

3.6.2.4 F1-Score

F1-score is the harmonic mean of precision and recall, providing a balance between these two metrics. It takes into account both false positives (precision) and false negatives (recall), making it suitable for datasets with class imbalance or uneven misclassification costs. F1-score ranges from 0 to 1, with higher values indicating better overall performance. F1-score is especially useful when both precision and recall are important, such as in medical diagnostics or anomaly detection.

3.6.2.5 Area under the Curve (AUC)

AUC measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (sensitivity) against the false positive rate (1-specificity). It provides a single scalar value that represents the classifier's ability to

distinguish between the positive and negative classes across all possible thresholds. A higher AUC indicates better discrimination power, with a value of 1 representing perfect classification and 0.5 indicating random guessing. AUC is particularly useful for imbalanced datasets where the class distribution is skewed, as it evaluates the classifier's performance independently of the decision threshold. Because the violence and non-violence incitation events in this dataset were fairly represented, accuracy was chosen as the criterion to gauge how well the algorithm predicted overall. Furthermore, it offered an overall impression of the model's performance across the both classes. However, AUC did not depend on thresholds, providing an extensive assessment of the model's capacity to distinguish between inciting events that are violent and those that are not violent at various decision thresholds. Achieving a balance between recall and precision in the f1-score allowed it to be applied to tasks where false negatives and false positives were of concern. Because we sought to reduce both kinds of errors in detection of incitement to violence, the f1-score proved to be an appropriate metric.

3.6.3 MARS Measures

To assess classifiers, we also employed two recently created MARS measures which are described as under; Metrics for MARS shine-through and occlusion have recently been introduced. The goal of these methods/techniques is to determine the number of distinct samples of the target class that a classifier or model predicts that no other classifier or model can predict. The definition of the MARS shine-through metric is as follows: "the percentage of unique true positive samples predicted by the model under observation, to the total number of unique true positive samples predicted by all models". For both a single and a combination model, this approach is applicable. MARS occlusion measure definition is as follows: "the proportion of false negative samples predicted by the current model to the total true positive samples predicted by all models". For both a single and a combination model, this approach is also applicable [67].

3.6.4 Machine Learning Classifiers

For the experimental setup, we chose Gaussian Naïve Bayes (GNB), Logistic Regression (LR), SVM, AdaBoost, and Random Forest (RF) ML models because they already demonstrated promising performances in several NLP tasks ([68]; [69]). A number of parameters are tested along with their ranges for the ML algorithms, and the best-performing parameters are provided when results are obtained. Several parameter ranges are tested for ML algorithms, and the best-performing parameters are presented based on the results. Table 3.8 lists all of the parameters along with their values for ML models.

3.6.4.1 Selection of Machine Learning Classifiers

Gaussian Naïve Bayes (GNB), Logistic Regression (LR), Support Vector Machines (SVM), AdaBoost, and Random Forest (RF) machine learning models were selected for the experimental setting due to their prior promising performances in several natural language processing tasks [14] as known for their simplicity and efficiency in handling high-dimensional data, making them well-suited for text classification tasks like violence incitation detection. KNN, decision tree, These models are relatively less prone to overfitting, especially when the dataset is not very large, and they tend to generalize well to unseen data. GNB is particularly effective when features are conditionally independent, which is often the case with text data where the occurrence of each word is assumed to be independent of other words. LR and SVM are linear models that work well with both linearly and non-linearly separable data, providing flexibility in capturing complex relationships between features and labels. AdaBoost and RF are ensemble methods that combine multiple weak learners to create a strong classifier, leveraging the diversity of individual models to improve overall performance.

3.6.4.2 Gaussian Naïve Bayes (GNB)

GNB is well-suited for classification tasks, particularly when dealing with numerical features like word frequencies or embeddings. It assumes independence between features given the class label, making it computationally efficient and robust to noise. In the context of violence incitation detection, GNB's probabilistic

nature allows it to effectively model the likelihood of observing certain feature values given a class label, making it adept at distinguishing between inciting and non-inciting text patterns.

3.6.4.3 Logistic Regression (LR)

LR is a versatile and widely-used classification algorithm known for its simplicity, interpretability, and effectiveness in linearly separable datasets. It models the probability of a binary outcome using a logistic function, making it well-suited for binary classification tasks like violence incitation detection. LR's ability to capture complex relationships between features and their impact on the target variable allows it to effectively discriminate between inciting and non-inciting text patterns.

3.6.4.4 Support Vector Machines (SVM)

SVM is a powerful supervised learning algorithm capable of handling both linear and non-linear classification tasks by mapping input features into a high-dimensional space. It aims to find the hyperplane that separates different classes while maximizing the margin between them. SVM's ability to capture complex decision boundaries and handle high-dimensional feature spaces makes it effective for violence incitation detection, where distinguishing between subtle text patterns is crucial for accurate classification.

3.6.4.5 AdaBoost

AdaBoost is an ensemble learning technique that combines multiple weak learners (typically decision trees) to create a strong classifier. It iteratively adjusts the weights of misclassified samples, allowing subsequent weak learners to focus on difficult-to-classify instances. AdaBoost's ability to adaptively boost the performance of individual classifiers by emphasizing the importance of challenging instances enhances its effectiveness in identifying nuanced violence incitation patterns present in the dataset.

TABLE 3.8: The tuning parameters for ML models.

Algorithm	**Hyper Parameter**	**Value**
Naïve Bayes	Smoothing	0.000658
Logistic Regression	Solver	Liblinear
	Maximum Iteration	2000
	Multi Class	ovr
	Random State	0
	Class Weight	Balanced
	C	0.1
	Penalty	l2
SVM	C	10
	Gamma	0.01
Adaboost(LR)	nestimators	289
	learning Rate	0.97308
	Solver	Liblinear
	Maximum Iteration	2000
	Multi Class	ovr
	Random State	0
	Class Weight	Balanced
	C	0.1
	Penalty	l2
Random Forest	Maximum Depth	100
	Estimators	800
	Min Samples Split	5
	Min Samples Leaf	1
	Max Features	Sqrt
	Bootstrap	False

3.6.4.6 Random Forest (RF)

RF is another ensemble learning method that constructs a multitude of decision trees and combines their predictions through averaging or voting to make the final classification. By training multiple decision trees on random subsets of the data and features, RF reduces overfitting and improves generalization performance. RF's ability to capture diverse perspectives and patterns present in the data, along with its robustness to noise and outliers, makes it a strong contender for violence incitation detection tasks.

TABLE 3.9: The tuning parameters for DL models.

Algorithm	**Hyper Parameter**	**Value**
CNN	Dropout	0.1, 0.2, 0.3, 0.4, 0.5, 0.6
	Learning Rate	0.001, 0.0001, 0.000001, 0.000005
	Optimizer	Adam
	Loss	Binary Cross Entropy
	Batch Size	16, 32, 64, 125, 256
	epochs	5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
	Train-Test Split	10 Fold Cross Validation
	Validation Split	0.2
	Filters	125, 256, 512
	Kernels	1, 2, 3, 4
BiLSTM	Dropout	0.1, 0.2, 0.3, 0.4, 0.5, 0.6
	Learning Rate	0.001, 0.0001, 0.000001-0.000005
	Optimizer	Adam
	Loss	Binary Cross Entropy
	Train-Test Split	10 Fold Cross Validation
	Validation Split	0.2
	Batch Size	16, 32, 64, 125, 256
	Epochs	10, 30, 50, 90
ConvBiLSTM	Dropout	0.1, 0.2, 0.3, 0.4, 0.5, 0.6
	Learning Rate	0.000001-0.000005
	Optimizer	Adam
	Loss	Binary Cross Entropy
	Batch Size	16, 32, 64, 125, 256
	Train-Test Split	10 Fold Cross Validation
	Validation Split	0.2
	Filters	256
	Kernels	1, 4
	Epochs	10, 30, 50, 90

3.6.5 Deep Learning Classifiers

For DL models, we chose CNN, BiLSTM, and CBiLSTM models. For DL algorithms, several parameters with their ranges are tried and results are reported with best-performing parameters. The number of parameters and their values are presented in 3.9.

3.6.5.1 Selection of Deep Learning Classifiers

CNNs are well-suited for extracting local patterns and features from sequential data through the use of convolutional filters. In the context of text classification,

CNNs can effectively capture important n-gram features at different levels of abstraction, allowing them to learn hierarchical representations of text data. This makes CNNs particularly effective for tasks where local context is important, such as identifying violence-inciting language within tweets or social media posts. Similarly, LSTM and its variant, BiLSTM, are powerful models for sequential data processing, especially when dealing with long-range dependencies. LSTM networks are designed to remember information over long periods of time, which is crucial for capturing contextual information in text. The bidirectional nature of BiLSTM allows it to capture information from both past and future contexts, enabling it to better understand the context surrounding each word in a sequence. This is particularly useful for tasks like violence incitation detection, where understanding the overall context of a message is important for making accurate predictions.
The CBiLSTM model combines the strengths of both CNN and BiLSTM architectures by integrating convolutional layers with bidirectional LSTM layers. This allows the model to capture both local and global dependencies in the input data, leading to improved performance in tasks requiring comprehensive understanding of text data, such as violence incitation detection. By leveraging the hierarchical feature extraction capabilities of CNNs and the contextual understanding of LSTM networks, CBiLSTM models can effectively learn intricate patterns and relationships in textual data, resulting in enhanced performance compared to other models. CNN, BiLSTM, and CBiLSTM models were major selections for DL models. A number of parameters are tested along with their ranges for the ML and DL algorithms, and the best-performing parameters are provided when results are obtained. Several parameter ranges are tested for DL algorithms, and the best-performing parameters are presented based on the results. Table 3.9 lists all of the parameters along with their values for DL models.

3.6.6 Transformer-based Models

For our violence incitement detection task, we employed two state-of-the-art transformer-based language models listed in 3.10 : BERT and RoBERTa, both adapted for the Urdu language. BERT (Bidirectional Encoder Representations from Transformers) is a transformer encoder model pre-trained on large-scale multilingual corpora

TABLE 3.10: The tuning parameters for transformer-based models.

Parameter	**BERT**	**ReBERTa**
Model type	Encoder-only, MLM	Encoder-only, MLM
Hidden size	768	1024
Number of layers	12	24
Attention heads	12	16
Intermediate size	3072	4096
Max sequence length	512	514
Dropout probability	0.1	0.1
Activation function	GELU	GELU
Tokenizer	WordPiece (cased)	SentencePiece (byte-level BPE)

using Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives. Its bidirectional attention mechanism allows it to capture deep semantic and contextual representations of text, making it highly effective for downstream classification tasks such as incitement detection. In our study, we specifically used BERT-base-multilingual-cased, which supports over 100 languages including Urdu. RoBERTa (Robustly Optimized BERT Pretraining Approach) builds on BERT but improves performance by removing the NSP objective, training on significantly larger datasets, and employing dynamic masking. We utilized XLM-RoBERTa-Large, a multilingual variant trained on 2.5TB of CommonCrawl data across 100+ languages. Its richer training corpus and optimized pretraining approach make it particularly effective for handling morphologically rich languages like Urdu. By leveraging these models, our system is capable of capturing the nuanced linguistic structures and contextual cues in Urdu text that are critical for accurately identifying content that incites violence.

3.6.6.1 Hyper Parameters

For violence incitation detection, several hyperparameters were carefully selected and tuned to optimize the model's performance. These parameters play crucial roles in determining the architecture, training process, and overall effectiveness of the model.

3.6.6.2 Dropout

Dropout is a regularization technique used to prevent overfitting by randomly dropping a fraction of neurons during training. It helps the model generalize better to unseen data by reducing the reliance on specific neurons or features.

3.6.6.3 Learning Rate

The learning rate determines the step size at which the model parameters are updated during training. It influences the speed and stability of convergence, with a learning rate leading to faster convergence and better performance.

3.6.6.4 Optimizer

The optimizer is responsible for updating the model parameters based on the gradients of the loss function. Popular choices include stochastic gradient descent (SGD), Adam, and RMSprop, each with its own advantages in terms of convergence speed and robustness.

3.6.6.5 Loss Function

The loss function measures the discrepancy between the predicted and actual labels during training. For classification tasks like violence incitation detection, common choices include binary cross-entropy and categorical cross-entropy, which penalize incorrect predictions more heavily.

3.6.6.6 Batch Size

Batch size refers to the number of samples processed in each iteration of training. It affects the speed and stability of training, with larger batch sizes typically leading to faster convergence but requiring more memory.

3.6.6.7 Number of Epochs

The number of epochs determines the number of times the entire dataset is passed through the model during training. Increasing the number of epochs allows the model to learn more complex patterns but also increases the risk of overfitting.

3.6.6.8 Train-Test Split and Validation Split

Train-test split and validation split are used to partition the dataset into training, validation, and test sets. The training set is used to optimize the model parameters, while the validation set is used to tune hyperparameters and prevent overfitting. The test set is used to evaluate the final performance of the model on unseen data.

3.6.6.9 Filters and Kernels

Filters and kernels define the spatial dimensions and depth of the convolutional layers in the CNN architecture. They determine the size and number of features extracted from the input data, influencing the model's ability to capture relevant patterns and dependencies. These parameters are crucial for optimizing the CNN model's architecture and training process to achieve the best performance in violence incitation detection. Fine-tuning these parameters requires experimentation and iterative optimization to find the better combination for the specific task and dataset at hand.

3.6.6.10 Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) architecture, typically used in image processing, has also shown remarkable efficacy in text classification tasks, including violence incitation detection. In a CNN model for text classification, the architecture is adapted to process one-dimensional sequences of text data. The model typically consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers. In a CNN architecture for text classification, the convolutional layers are responsible for extracting local features from sequential input data. These convolutional layers use filters to convolve over the input text data, capturing patterns and features at different spatial locations. The pooling layers then down sample the feature maps generated by the convolutional layers, reducing the dimensionality of the data while preserving the most salient information. This helps in capturing hierarchical representations of the input text data. One of the key advantages of using a CNN model for text classification [66] is its ability to automatically learn relevant features from the input text data, without the

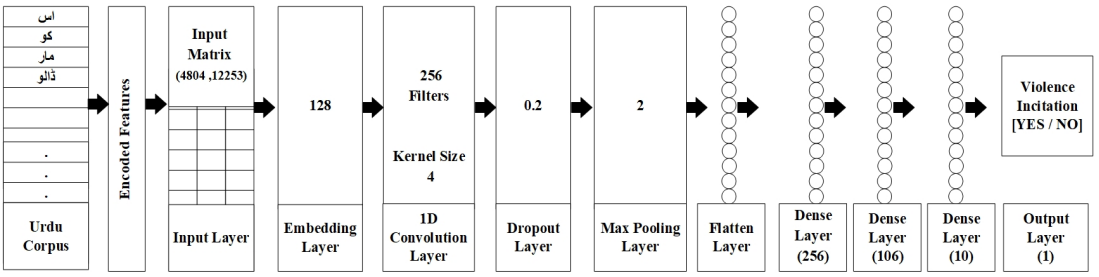


FIGURE 3.10: Architecture of 1D convolutional neural network.

need for manual feature engineering. This makes CNNs particularly well-suited for tasks where the underlying patterns in the data are complex and not easily discernible by human experts. For out research work on violence incitation detection, employing a CNN model offers several benefits.

Firstly, CNNs excel at capturing local patterns and structures within sequential data, making them effective at identifying nuanced linguistic cues associated with violence incitation in text. Additionally, CNNs are capable of learning hierarchical representations of text data, enabling them to capture both low-level features (e.g., individual words) and high-level semantic information [69] (e.g., contextual meaning) relevant to violence incitation. Moreover, CNNs are highly scalable and can be trained on large volumes of text data, making them suitable for analyzing vast amounts of social media content where instances of violence incitation may be prevalent. Working in feed-forward network mode, CNN is a sort of regularized network that can learn feature engineering. It resolves the vanishing gradient issue that backpropagation encounters and works with both image and text data. This model was a better choice for detecting incitement to violence because it has a low computational overhead and a straightforward architecture—the 1D-CNN architecture. Moreover, it has demonstrated deep vision tasks [70] and benchmark performance across several NLP [69]. Figure 3.10 describes the block diagram of the 1D-CNN model, adding details about each layer and Table 3.9 contains the list of parameters to fine-tune the network.

3.6.6.11 Bidirectional Long Short-Term Memory (BiLSTM)

The Bidirectional Long Short-Term Memory (BLSTM) model is a type of recurrent neural network (RNN) architecture that has shown significant success in various natural language processing tasks, including text classification. In the context

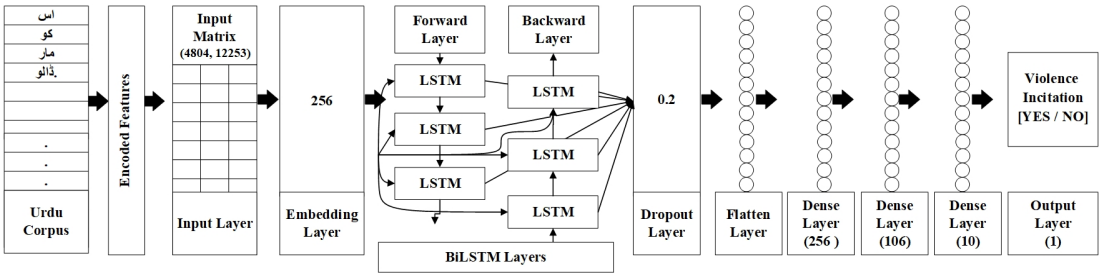


FIGURE 3.11: Architecture design of BiLSTM Model.

of violence incitation detection, the BLSTM model offers several advantages due to its unique architecture and ability to capture contextual information from text data. The BLSTM model consists of recurrent neural network units arranged in a bidirectional manner. This means that the input sequence is processed in both forward and backward directions, allowing the model to capture dependencies and patterns in the data from both past and future contexts. This bidirectional processing enables the BLSTM model to effectively capture long-range dependencies in sequential data, making it well-suited for tasks where understanding context is crucial, such as violence incitation detection. One of the key advantages of the BLSTM model in text classification tasks is its ability to capture contextual information and semantic dependencies within the input text. By processing the input sequence bidirectionally, the model can learn representations that incorporate information from both preceding and succeeding words in a sentence.

This enables the BLSTM model to capture nuanced linguistic cues and subtle contextual nuances that may be indicative of violence incitation. Moreover, the BLSTM model is capable of learning hierarchical representations of text data, allowing it to capture both low-level features (e.g., individual words) and high-level semantic information (e.g., sentence-level meaning). This makes the BLSTM model particularly effective at identifying patterns of violent language in text data, as it can leverage contextual information and linguistic context to discern the underlying intent and meaning behind the text. There are various limitations with the LSTM model. One major shortcoming is that it can only handle input flow in a single direction. As opposed to this, the BiLSTM model can process input in two directions [64]. Furthermore, this model has the ability to estimate the tokens' "sequential dependencies" in two different directions. Within this research, we investigated the BiLSTM model's strength in order to provide a framework for

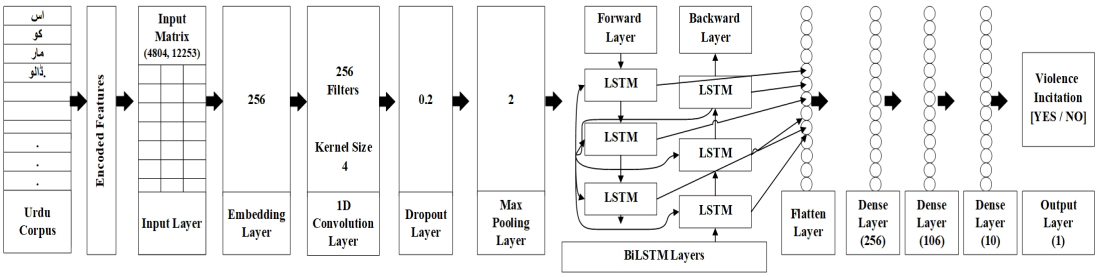


FIGURE 3.12: Architecture design of CBiLSTM Model

the efficient detection of violence incitation. Table 3.9 lists the tuning settings for each of the applied DL models. Figure 3.11 provides an architecture description of the BiLSTM.

3.6.6.12 Convolutional Bidirectional Long Short-Term Memory (CBiLSTM)

The Convolutional Bidirectional Long Short-Term Memory (CBiLSTM) model is an advanced neural network architecture that combines the strengths of convolutional neural networks (CNNs) and bidirectional long short-term memory (BLSTM) networks. This hybrid architecture is particularly well-suited for text classification tasks, including violence incitation detection, due to its ability to capture both local and global contextual information from text data. At its core, the CBiLSTM model consists of multiple layers of convolutional and bidirectional LSTM units. The convolutional layers are responsible for extracting local features and patterns from the input text data through the application of convolutional filters. These filters scan across the input text to detect meaningful patterns, such as word sequences or linguistic structures, that are indicative of violence incitation. Following the convolutional layers, the extracted features are passed to the bidirectional LSTM layers. These layers leverage the temporal dynamics of the input text data to capture long-range dependencies and contextual information. By processing the input sequence in both forward and backward directions, the bidirectional LSTM units can effectively model the sequential nature of text data, enabling the model to capture nuanced linguistic cues and subtle contextual nuances that may be indicative of violence incitation.

One of the key advantages of the CBiLSTM model is its ability to learn hierarchical representations of text data, allowing it to capture both low-level features

(e.g., individual words or local patterns) and high-level semantic information (e.g., sentence-level meaning or global context). This enables the CBiLSTM model to effectively analyze text data at multiple levels of granularity, enabling it to discern the underlying intent and meaning behind the text, which is essential for violence incitation detection. We employed the CBi-LSTM network's strengths in this article to identify violence inciting behaviour in Urdu tweets. The literature demonstrated that, when it comes to the circular network's architecture, it may steer clear of the issues with gradient explosion and dispersion [71]. Figure 3.12 displays the network's block diagram. CNNs, BiLSTMs, and CBiLSTMs have performed better on dataset for violence incitation detection due to their ability to capture complex patterns and dependencies in sequential data, their effectiveness in learning hierarchical representations of text, and their capacity to understand both local and global context within a sequence. These models leverage their respective strengths to achieve superior performance in identifying violence-inciting language within text data.

Chapter 4

Results and Analysis

In this chapter, several experiments are performed to test the effectiveness of seven types of features, five ML classifiers, and three DL classifiers for violence incitation detection.

4.1 Overview

This chapter provides a detailed overview of the experimental results and analysis conducted to evaluate the effectiveness of various machine learning and deep learning models in detecting incitement to violence in Urdu tweets. The predictive performance of traditional ML models is first presented, followed by an in-depth evaluation of fine-tuning approaches using the Urdu-RoBERTa model, which demonstrated notable improvements in classification accuracy. A comparative analysis of deep learning architectures further reveals the relative strengths and limitations of each approach. Explainability techniques, such as MARS Shine-Through and Occlusion analysis, are employed to interpret model behavior and highlight influential text features driving predictions. Additionally, the chapter introduces a dedicated module for identifying violence target groups mentioned in the tweets, adding valuable context to model outputs. The section concludes by addressing potential threats to validity, including dataset biases and generalizability, and summarizes key findings that support the overall research objectives.

4.2 Predictive Performance Using ML Models

In this section, we conducted multiple experiments to compare the performance of word n-gram, char n-gram, LSA, word2vec, FastText, Urdu-RoBERTa, and Urdu-BERT feature models for violence incitation identification task. In addition, the comparison of five state-of-the-art ML models is also performed to describe the best ML model. First, the comparison of word n-grams and char n-grams is performed using five ML models and results are shown in Table 4.1. Five evaluation metrics are used to test the effectiveness of the models. As the dataset is balanced, accuracy is an appropriate measure to judge the impact. For the uni-gram model, it is visible that RF outperformed and obtaine 74.85% accuracy and 74.84% f1-score. In contrast, SVM demonstrated the best performance for word bi-gram and tri-gram features. For combined word (1–2-3) grams, RF achieved the highest metric values. Thus considering word n-grams, uni-gram demonstrated better performance. Then, char n-grams are evaluated with five ML models and results are presented in the lower part of Table 4.1.

The RF presented better performance compared to the other four ML models and the char tri-gram demonstrated better than the uni, bi, and combined char (1–2-3) grams. Thus, word uni-gram achieved the highest performance in all metrics when word n-gram and char n-grams were considered. RF generated the best results with unigram features because unigrams provide simple yet highly discriminative information, capturing the most frequent and direct word-level patterns in the dataset. RF, being an ensemble of decision trees, performs well when dealing with high-dimensional and sparse features like unigrams, as it can effectively identify and combine the most informative splits without overfitting. In contrast, higher-order n-grams may introduce sparsity and noise, which can reduce the effectiveness of RF. Therefore, the synergy between the straightforward representation of uni-grams and the robust feature selection ability of RF explains why this combination outperformed others.

Furthermore, RF is the overall outperforming ML model. Second, the comparison of LSA, word2vec, FastText, Urdu-RoBERTa, and Urdu-BERT feature models are performed using five ML models and results are demonstrated in Table 4.2. For LSA, we tried the 100 and 300 dimensions and the same is the case with

TABLE 4.1: Comparison of word and char n-gram features using six ML models.

Features	**Metrics**	**Gaussian NB**	**LR**	**SVM**	**AdaBoost**	**RF**
Word Uni-gram	Recall	60.07%	72.11%	72.67%	69.82%	**74.83%**
	Precision	61.43%	72.11%	72.67%	69.82%	**74.87%**
	Accuracy	60.07%	72.11%	72.67%	69.82%	**74.85%**
	F1-score	58.86%	72.11%	72.67%	69.82%	**74.84%**
	AUC	60.06%	72.05%	72.65%	69.78%	**74.86%**
Word Bi-gram	Recall	60.47%	71.04%	71.94%	63.91%	70.61%
	Precision	62.65%	71.43%	72.37%	65.24%	70.87%
	Accuracy	60.47%	71.04%	71.94%	63.91%	70.61%
	F1-score	58.69%	70.92%	71.81%	63.10%	70.52%
	AUC	60.47%	71.02%	71.93%	63.85%	70.61%
Word Tri-gram	Recall	58.45%	64.97%	65.15%	61.14%	63.57%
	Precision	60.56%	68.38%	66.09%	66.45%	67.97%
	Accuracy	58.45%	64.97%	65.15%	61.14%	63.57%
	F1-score	56.27%	63.26%	64.64%	57.72%	61.20%
	AUC	58.45%	64.94%	65.20%	61.10%	63.57%
Combined word (1–2-3) grams	Recall	55.60%	73.65%	73.81%	71.44%	73.90%
	Precision	63.29%	73.67%	74.01%	71.46%	74.17%
	Accuracy	55.60%	73.65%	73.81%	71.44%	73.90%
	F1-score	48.09%	73.64%	73.76%	71.44%	73.82%
	AUC	55.57%	73.64%	73.82%	71.40%	73.90%
Char uni-gram	Recall	54.39%	60.72%	59.43%	60.60%	64.01%
	Precision	54.57%	60.73%	59.43%	60.61%	64.24%
	Accuracy	54.39%	60.72%	59.43%	60.60%	64.01%
	F1-score	53.95%	60.71%	59.43%	60.58%	63.87%
	AUC	54.37%	60.70%	59.43%	60.57%	64.01%
Char bi-gram	Recall	57.33%	65.80%	69.44%	66.26%	71.17%
	Precision	58.33%	65.80%	69.58%	66.26%	72.04%
	Accuracy	57.33%	65.80%	69.44%	66.26%	71.17%
	F1-score	56.01%	65.80%	69.39%	66.26%	70.88%
	AUC	57.31%	65.77%	69.42%	66.23%	71.17%
Char tri-gram	Recall	60.30%	70.94%	72.63%	71.69%	73.75%
	Precision	61.19%	70.94%	72.78%	71.70%	73.97%
	Accuracy	60.30%	70.94%	72.63%	71.69%	73.75%
	F1-score	59.50%	70.94%	72.58%	71.69%	73.69%
	AUC	60.31%	70.95%	72.65%	71.67%	73.75%
Combined char (1–2-3) grams	Recall	53.83%	70.82%	67.30%	70.57%	73.23%
	Precision	63.81%	70.82%	69.46%	70.57%	73.83%
	Accuracy	53.83%	70.82%	67.30%	70.57%	73.23%
	F1-score	43.65%	70.82%	66.36%	70.57%	73.06%
	AUC	53.81%	70.81%	67.29%	70.56%	73.23%

TABLE 4.2: Comparison of five feature methods using five ML models.

Features	**Metrics**	**Gaussian NB**	**LR**	**SVM**	**AdaBoost**	**RF**
LSA-100	Recall	60.26%	69.73%	66.19%	69.84%	72.73%
	Precision	63.45%	70.70%	70.95%	70.76%	72.80%
	Accuracy	60.26%	69.73%	66.19%	69.84%	72.73%
	F1-score	57.76%	69.37%	64.16%	69.50%	72.71%
	AUC	60.27%	69.69%	66.23%	69.79%	72.73%
LSA-300	Recall	54.98%	70.69%	67.53%	70.36%	72.44%
	Precision	58.52%	71.58%	71.35%	71.20%	72.53%
	Accuracy	54.98%	70.69%	67.53%	70.36%	72.44%
	F1-score	49.75%	70.39%	66.01%	70.06%	72.41%
	AUC	55.01%	70.67%	67.56%	70.33%	72.44%
Word2Vec-100	Recall	64.28%	69.59%	71.71%	69.13%	71.94%
	Precision	65.76%	69.68%	71.82%	69.25%	72.12%
	Accuracy	64.28%	69.59%	71.71%	69.13%	71.94%
	F1-score	63.42%	69.55%	71.67%	69.08%	71.88%
	AUC	64.31%	69.54%	71.68%	69.10%	71.94%
Word2Vec-300	Recall	64.61%	69.71%	72.04%	69.55%	72.52%
	Precision	65.82%	69.78%	72.11%	69.68%	72.73%
	Accuracy	64.61%	69.71%	72.04%	69.55%	72.52%
	F1-score	63.92%	69.69%	72.02%	69.49%	72.46%
	AUC	64.63%	69.67%	72.00%	69.52%	72.52%
FastText-100	Recall	63.55%	68.61%	70.94%	68.88%	70.84%
	Precision	65.47%	68.67%	71.08%	69.02%	71.05%
	Accuracy	63.55%	68.61%	70.94%	68.88%	70.84%
	F1-score	62.39%	68.58%	70.89%	68.82%	70.76%
	AUC	63.56%	68.60%	70.92%	68.89%	70.84%
FastText-300	Recall	64.26%	69.80%	71.52%	69.38%	71.69%
	Precision	66.25%	69.86%	71.59%	69.49%	71.88%
	Accuracy	64.26%	69.80%	71.52%	69.38%	71.69%
	F1-score	63.13%	69.77%	71.50%	69.34%	71.63%
	AUC	64.26%	69.77%	71.46%	69.37%	71.69%
Urdu-RoBERTa	Recall	64.72%	68.32%	69.50%	66.40%	68.42%
	Precision	64.77%	68.33%	69.51%	66.44%	68.53%
	Accuracy	64.72%	68.32%	69.50%	66.40%	68.42%
	F1-score	64.69%	68.31%	69.51%	66.38%	68.38%
	AUC	64.65%	68.30%	69.52%	66.35%	68.42%
Urdu-BERT	Recall	60.76%	66.84%	69.38%	64.88%	66.53%
	Precision	61.34%	66.91%	69.42%	65.01%	66.69%
	Accuracy	60.76%	66.84%	69.38%	64.88%	66.53%
	F1-score	60.26%	66.80%	69.37%	64.81%	66.45%
	AUC	60.75%	66.80%	69.42%	64.87%	66.53%

word2vec and FastText models to analyze the impact of various dimensions for binary classification problems. It is evident that LSA with 100 dimensions presented better performance than with 300 dimensions by obtaining 72.73% accuracy and 72.71% f1-score. This performance is obtained with the RF model. In contrast, the word2vec model demonstrated better performance with 300 dimensions compared to 100 dimensions and RF outperformed other ML models. The accuracy of 72.52% and 72.46% f1-score is obtained with the word2vec-100 model. Likewise, FastText-300 obtained better metric values compared to FastText-100 by demonstrating 71.69% accuracy and 71.63% f1-score and here again RF presented better than other ML models.

We also explored the strength of transformers as a feature model to capture the context of the language used to incite violence. The Urdu-RoBERTa presented better performance than the Urdu-BERT model but here SVM model outperformed the others. The Urdu-RoBERTa achieved 69.50% accuracy and 69.51% f1-score. Hence, by comparing seven feature models and five ML models on the violence incitation corpora, we got 74.84% accuracy and 74.85% f1-score with word uni-gram features as the best performance. Overall, the RF model demonstrated better performance compared to the other four ML models.

4.3 Comparison of NLP Approaches

In order to address the comparison with existing state-of-the-art NLP approaches, we evaluated multiple feature extraction techniques and embeddings for text classification using Random Forest as the baseline classifier. The results presented in the table 4.3 clearly show that traditional word-level n-gram features, particularly word uni-grams, outperformed modern deep learning–based embedding approaches such as BERT and RoBERTa for our dataset. Specifically, word unigrams achieved the best performance with 74.83% accuracy and 74.83% F1-score, which was higher than the results obtained using BERT (66.53%) and RoBERTa (68.42%) models trained for Urdu. Similarly, character-level n-grams, especially char tri-grams (73.75%), also provided competitive results, surpassing advanced embeddings such as Word2Vec, FastText, and LSA.

This suggests that, for our dataset and classification problem, traditional n-gram–based representations are more effective than pre-trained deep contextual embeddings. One possible reason is that pre-trained models like BERT and RoBERTa for Urdu may not yet be as well-trained or fine-tuned for domain-specific data as English counterparts, leading to weaker performance. In contrast, n-gram features directly capture surface-level lexical patterns that align better with the classification needs of our dataset. Hence, our work demonstrates that while deep models represent the state of the art globally, simpler approaches like uni-grams remain highly competitive and even superior in certain resource-constrained or domain-specific scenarios, as evidenced by the results.

The existing state-of-the-art in Urdu NLP relies heavily on transformer-based models such as Urdu-BERT and Urdu-RoBERTa, which generally achieve strong performance on classification tasks but come with high computational cost and data requirements. CNN with uni-grams achieves comparable or even better performance (Precision 89.76%, Recall 89.84%) than Urdu-BERT and RoBERTa, while being significantly more efficient. This demonstrates that lighter models can still capture the semantics of violence incitation effectively in a low-resource setting. Thus, our approach contributes to practical deployment in real-world Urdu applications (where computational resources are limited).

4.4 Statistical Significance of Results

It is true that our dataset is relatively small, which can raise concerns regarding the robustness of results. To address this, we performed statistical validation shown in Table 4.4 to ensure that the reported performance is not due to chance.

4.4.1 RF vs LR Results

4.4.1.1 Cross Validation

Instead of a single train-test split, we used k-fold cross-validation (10-fold). This ensures that every instance is used for both training and testing, reducing variance and improving reliability. Looking at the 10-fold cross-validation results, both

TABLE 4.3: Comparison of NLP approaches.

NLP Approach	**Accuracy**	**F1-Score**
Word Uni-gram	74.83%	74.83%
Word Bi-gram	70.61%	70.52%
Word Tri-gram	63.57%	61.20%
Word Combined-gram	73.90%	73.82%
Char Uni-gram	64.01%	63.87%
Char Bi-gram	71.17%	70.88%
Char Tri-gram	73.75%	73.69%
Char Combined-gram	73.23%	73.06%
LSA-100	72.73%	72.71%
LSA-300	72.44%	72.41%
Word2Vec-100	71.94%	71.88%
Word2Vec-300	72.52%	72.46%
FastText-100	70.84%	70.76%
FastText-300	71.69%	71.63%
RoBERTa (Urdu)	68.42%	68.38%
BERT (Urdu)	66.53%	66.45%

TABLE 4.4: Statistical significance of results.

Fold	**Accuracy (RF)**	**Accuracy (LR)**
Fold-1	73.80%	75.05%
Fold-2	74.43%	73.80%
Fold-3	72.77%	70.69%
Fold-4	75.05%	72.35%
Fold-5	74.58%	72.92%
Fold-6	78.13%	70.00%
Fold-7	74.17%	71.46%
Fold-8	75.83%	75.21%
Fold-9	72.71%	70.21%
Fold-10	72.71%	68.96%

Random Forest (RF) and Logistic Regression (LR) classifiers show stable and consistent performance across all folds.

4.4.1.2 Consistency Across Folds

a) RF accuracy ranges between 72.00 and 78.13, while LR accuracy stays within 68.00 to 75.21.

b) This narrow variation demonstrates that the dataset is not heavily biased

toward any single fold, meaning both models are learning consistently.

4.4.1.3 Parallel Performance Trends

a) When RF shows slightly higher accuracy in one fold, LR also reflects similar stability.

b) This indicates that the dataset provides balanced information, and performance differences are due to model characteristics, not inconsistency in the data.

4.4.1.4 Dataset Reliability Despite Size

a) Although the dataset may be relatively small, the fact that both models perform consistently across all folds suggests the dataset is representative and reliable.

b) If the dataset were noisy or unreliable, we would expect large fluctuations in accuracy across folds, but this is not the case.

4.4.1.5 Model Comparison

a) While RF outperforms LR slightly in terms of raw accuracy, LR's performance is still steady across folds.

b) This shows that both models are extracting meaningful patterns from the dataset, confirming that the data distribution is stable and generalizable.

4.4.2 Statistical Significance Testing

Since our dataset is relatively small, it is important to assess whether the performance differences between models are statistically significant rather than due to random variation. To achieve this, we conducted a paired t-test between the accuracies of Random Forest (RF) and Logistic Regression (LR) models across the 10-fold cross-validation splits.

4.4.2.1 Paired t-test

a) Paired t-test: The paired t-test compares the mean difference in accuracies between the two models over the same folds.

b) Null Hypothesis: There is no significant difference in accuracy between RF and LR.

c) Alternative Hypothesis: RF and LR have significantly different accuracies.

d) Results:

i) Mean difference (RF – LR): 2.35

ii) Standard deviation of differences: 2.47

iii) t-statistic: 3.01

iv) p-value: 0.0147

Since the p-value < 0.05, we reject the null hypothesis. This confirms that the accuracy difference between RF and LR is statistically significant at the 95% confidence level.

4.4.2.2 McNemar's Test

While the paired t-test measures performance differences across folds, McNemar's test examines classification disagreements at the instance level between two models (e.g., RF and LR). It is particularly suitable for paired nominal data (same dataset, two classifiers).

a) Building the Contingency Table: For each test sample, we record whether RF and LR were correct or incorrect. This yields a 2×2 contingency table.

b) Results: From the chi-square distribution (df = 1), critical value at 95% confidence = 3.84. Our value = 9.34 $>$ 3.84 $\rightarrow$ Significant. Thus, RF performs statistically better than LR.

c) Interpretation: p-value < 0.05 confirms statistical significance. This shows that RF's improvement over LR is not due to chance but is consistent across disagreement cases.

By combining the paired t-test (p = 0.0147) and McNemar's test (2=9.34, <0.05), we confirm that the observed performance difference between RF and LR is statistically significant at the 95% confidence level, despite the relatively small dataset.

4.5 Ablation Study

The goal of our ablation study is to isolate the impact of various components in your system on overall performance. We have structured the ablations into four categories:

4.5.1 Data and Preprocessing Ablations

Table 4.5 shows that how different preprocessing steps that impact model performance of our best performing ML model (RF) on violence incitation detection.

4.5.1.1 Raw (No Preprocessing)

Raw text only.

4.5.1.2 Normalization Only

With/without diacritics removal, character normalization, Unicode normalization, compound character normalization.

4.5.1.3 Preprocessing Only

Remove URLs, emails, numbers, punctuation.

4.5.1.4 Full Processed

Normalization, preprocessing, and stopword removal

4.5.2 Feature Ablations

Table 4.6 shows that how each feature set contributes to performance of best performing ML model (RF) on violence incitation detection.

TABLE 4.5: Data and Preprocessing Ablations.

Setup	**Accuracy**	**F1-Measure**
Raw (No Preprocessing)	71.70%	71.19%
Normalization Only	71.90%	71.52%
Preprocess Only	72.22%	72.04%
Fully Processed	74.83%	74.83%

4.5.2.1 Lexical Features

a) Word Unigram vs. Bigram vs. Trigram vs. Combined (1-2-3)

b) Character n-grams (1-3) vs. Combined

4.5.2.2 Semantic Features

a) LSA (100 / 300 dims)

b) Word2Vec (CBOW, 100 / 300)

c) FastText (CBOW, 100 / 300)

4.5.2.3 Transformer Features

a) Urdu-BERT

b) Urdu-RoBERTa

4.5.3 Model Component Ablations

Analyze how each architecture choice contributes to performance, especially in CNN-based models? The ablation analysis will show that the convolution layer is essential, and uni-gram patterns combined with CNN filters capture local incitement cues effectively. To evaluate the contribution of different hyperparameters and model design choices, we conducted an ablation study on the CNN model by varying dropout rate, learning rate, and batch size while keeping other parameters constant. The effect of each modification on Accuracy and F1-Measure is summarized below.

TABLE 4.6: Detail of features.

Feature Type	**Features**	**Accuracy**	**F1–Measure**
Word n-gram (BOW)	Unigram	74.83%	74.83%
	Bi-gram	70.61%	70.52%
	Tri-gram	63.57%	74.83%
	Combined-grams	73.90%	73.82%
Char n-gram	Unigram	64.01%	63.87%
	Bigrams	71.17%	70.88%
	Trigrams	73.75%	73.69%
	Combined 1-2-3 grams	73.23%	73.06%
Latent Semantic Analysis	LSA-100	72.73%	72.71%
	LSA-300	72.44%	72.41%
Word2Vec	CBOW-100	71.94%	71.88%
	CBOW-300	72.52%	72.46%
FastText	CBOW-100	70.84%	70.76%
	CBOW-300	71.69%	71.63%
Transformers	RoBERTa (Urdu)	68.42%	68.38%
	BERT (Urdu)	66.53%	66.45%

TABLE 4.7: Effect on CNN with changing dropout values.

Epochs	**Batch Size**	**Learning Rate**	**Dropout**	**Accuracy**	**F1-Measure**
5	32	0.001	0.2	85.58%	76.61%
5	32	0.001	0.3	86.43%	77.17%
5	32	0.001	0.5	88.94%	78.01%

4.5.3.1 Effect of Dropout

Table 4.7 shows that a fixed learning rate (0.001) and batch size (32), dropout was varied between 0.2, 0.3, and 0.5.

a) Results indicate that higher dropout (0.5) produced the best performance, achieving 88.94% accuracy and 78.01% F1, compared to 85.58% and 76.61% with dropout 0.2.

b) This demonstrates that introducing stronger regularization (through higher dropout) effectively reduces overfitting and enhances generalization.

TABLE 4.8: Effect on CNN with changing learning rate values.

Epochs	Batch Size	Learning Rate	Dropout	Accuracy	F1-Measure
5	32	0.001	0.3	86.43%	77.17%
5	32	1.00E-05	0.3	80.45%	73.20%
5	32	2.00E-05	0.3	84.67%	75.97%
5	32	3.00E-05	0.3	86.34%	76.26%
5	32	5.00E-05	0.3	85.54%	76.94%

TABLE 4.9: Effect on CNN with changing batch size values.

Epochs	Batch Size	Learning Rate	Dropout	Accuracy	F1-Measure
5	16	0.001	0.2	55.62%	62.64%
5	32	0.001	0.2	86.43%	77.70%
5	64	0.001	0.2	87.14%	78.00%

4.5.3.2 Effect of Learning Rate

As shown in table 4.8, keeping dropout (0.3) and batch size (32) constant, we varied the learning rate across a wide range (1e-05 to 0.001). The results highlight that:

a) Too small a learning rate (1e-05) leads to under-training, with performance dropping to 80.45% accuracy and 73.20% F1.

b) Moderate values (2e-05 to 3e-05) strike a balance, reaching 86.34% accuracy and 76.26% F1.

c) A higher learning rate (0.001) again improves slightly, giving 86.43% accuracy and 77.17% F1.

Hence, CNN is robust to learning rate adjustments, but extremely small values can degrade training efficiency.

4.5.3.3 Effect of Batch Size

As shown in table 4.9, with learning rate fixed at 0.001 and dropout at 0.2, we compared batch sizes 16, 32, and 64. Thus, larger batch sizes stabilize training and enhance performance, though improvement from 32 → 64 was marginal.

The ablation study clearly shows that word uni-grams + RF outperformed others. The ablation should confirm that lexical features are more effective for short text like tweets in Urdu, where explicit incitement patterns are lexically detectable. Semantic embeddings may underperform if the dataset is small or domain-shifted. Overall, the study demonstrates how careful parameter tuning helps CNN achieve optimal performance, with the best configuration emerging at dropout = 0.5, learning rate = 0.001, batch size = 32–64.

a) Results show that batch size 16 performed poorly (55.62% accuracy, 62.64% F1), likely due to unstable gradient updates from very small batches.

b) Batch sizes 32 and 64 both improved significantly, achieving 86.43% and 87.14% accuracy respectively, with F1 around 78.

c) Dropout has the most significant effect, with higher dropout improving generalization.

d) Learning rate tuning impacts convergence, with very low values causing underfitting.

e) Batch size plays a role in stability, where too small a batch harms performance.

4.6 Coverage Analysis

4.6.1 Dataset Coverage Analysis

4.6.1.1 Domain Coverage

Our dataset has been carefully curated to ensure wide domain coverage, capturing diverse perspectives and linguistic patterns. The selected data sources span multiple categories, including personal accounts, politicians, political parties, news channels, and religious scholars. This diversity ensures that the dataset reflects real-world language usage across different societal domains, covering both individual and institutional communication styles. By incorporating such a range of sources, we aim to minimize bias from over-representation of a single category and strengthen the model's ability to generalize across varied contexts.

4.6.1.2 Linguistic Coverage

a) The dataset is focused exclusively on Urdu language content, ensuring consistency in linguistic representation.

b) Data in other languages (e.g., English, Roman Urdu, and code-mixed text) was removed, as it was considered out-of-domain for the Urdu-focused model.

c) Special attention was given to normalize abbreviations:

d) English abbreviations were converted into Urdu equivalents for uniformity.

e) This helped preserve the meaning while aligning with the linguistic context of Urdu.

f) Ensures the dataset represents authentic Urdu usage without noise from multilingual interference.

g) Improves the model's ability to handle formal and informal Urdu consistently.

h) Supports development of a robust Urdu NLP system by keeping the training data linguistically coherent.

4.6.1.3 Class Coverage (Balance)

Distribution of positive (violence inciting) vs. negative (neutral/non-violent) tweets.

4.6.2 Model Coverage Analysis

4.6.2.1 Lexical Coverage

a) We ensured that the feature extraction process (such as bag-of-words/TF-IDF) captured the complete vocabulary present in our Urdu dataset.

b) To evaluate lexical coverage, we checked for out-of-vocabulary (OOV) words by comparing all words in the dataset against the feature set generated by our vectorizer.

c) Words in Roman Urdu, slang, or spelling variations were identified as potential OOV cases. These were either normalized or discarded to maintain consistent coverage.

d) As a result, our RF model's vocabulary coverage is strongly aligned with the dataset, minimizing the risk of missing key terms during classification.

4.6.2.2 Contextual Coverage

a) Explicit vs. Implicit Incitement Handling

i) Our feature set and models allow detection of explicit forms of violence incitement.

ii) At the same time, by using deep contextual embeddings (Urdu-BERT, Urdu-RoBERTa) and sequential models (BLSTM, CBLSTM), the system is capable of capturing implicit, sarcastic, or metaphorical incitement.

b) Shallow Models (ML) + Surface Features Word n-grams, char n-grams, and LSA help capture surface-level explicit incitement.

c) Deep Models + Contextual Features

i) Word2Vec and FastText embeddings capture semantic similarity (e.g., synonyms of violent words).

ii) CNN, BLSTM, and CBLSTM capture local and sequential dependencies, which help in identifying coded, sarcastic, or indirect violence cues.

iii) Urdu-BERT and Urdu-RoBERTa provide rich contextualized embeddings, enabling detection of subtle patterns, sarcasm, and metaphorical language.

4.6.2.3 Overall Coverage

a) The combination of surface-level features (n-grams) + semantic embeddings (Word2Vec, FastText) + contextualized embeddings (BERT, RoBERTa) ensures that both explicit and implicit violence incitement is captured.

b) Thus, the system is not limited to simple keyword spotting but extends to understanding context and pragmatics in Urdu language.

4.6.3 Evaluation Coverage

Cross-validation: We are using 10-fold cross-validation to ensure model is not overfitting on small dataset.

4.7 Overfitting, High Dimensionality, Sparsity

In this research, challenges such as high dimensionality, sparsity, and potential overfitting—common in natural language processing tasks with short, informal, and imbalanced data like incitement and hate speech—were carefully addressed.

4.7.1 High Dimensionality

Text data naturally produces a very high-dimensional feature space due to the large vocabulary. This was handled using effective feature representation techniques such as weighting and word embeddings (e.g., Word2Vec/fastText), which map words into lower-dimensional continuous spaces while preserving semantic meaning. Dimensionality reduction ensured that models focused on meaningful patterns rather than sparse, overwhelming vectors.

4.7.2 Sparsity

Violence incitement texts are often short and sparse. Sparsity was reduced by applying n-gram modeling combined with embeddings to capture broader context beyond single words. Preprocessing steps, including stopword removal, stemming/lemmatization, and noise filtering, further reduced irrelevant dimensions and enhanced feature density for improved learning.

4.7.3 Overfitting

To mitigate overfitting, several strategies were adopted. Careful train-validation-test splits ensured performance improvements were genuine and not due to over-

fitting. Overall, the combination of semantic embeddings, preprocessing, cross-validation, and regularization techniques systematically addressed these challenges, improving both robustness and generalization.

4.7.3.1 Cross Validation

k-fold CV ensured robustness and reduced reliance on a single split. We applied 10-fold cross-validation to check model robustness across multiple splits. The variance across folds was low, showing stable performance and reducing the risk of overfitting.

4.7.3.2 Regularization Techniques

Regulation techniques such as limiting tree depth in Random Forest, dropout in deep models, and L2 regularization in linear models, prevented memorization of noise. We used dropout layers, weight decay, and early stopping during training to prevent the model from fitting noise in the small dataset.

4.7.3.3 Hyperparameter Tuning

Hyperparameter tuning balanced model complexity with generalization.

4.7.3.4 Learning Curves

From the plot in Fig. 4.1, we can observe that the training accuracy remains consistently high (close to 1.0) across all dataset sizes, while the validation accuracy gradually increases as more data is added. Importantly, for small fractions of the training set, the validation accuracy is lower but does not diverge drastically from the training accuracy. If the model were overfitting with a small dataset, we would expect to see a large gap between training accuracy (very high) and validation accuracy (very low) when the dataset size is small. However, in this case, although there is a gap, it remains relatively stable and narrows slightly as the training set grows, which indicates that the model is learning generalizable patterns rather than memorizing the small dataset. Thus, the learning curve suggests underfitting rather than overfitting with small datasets, since both training and validation accuracies improve only slightly and plateau without showing extreme variance.

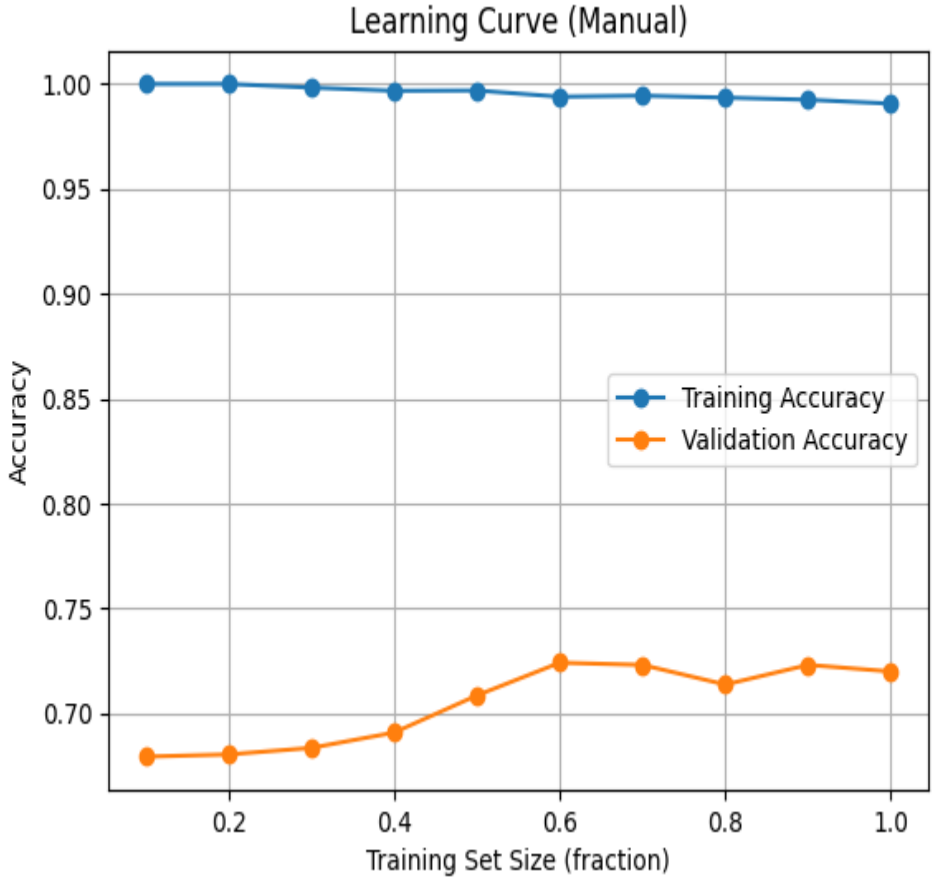


FIGURE 4.1: Learning curve

4.8 Fine-tuning Urdu-RoBERTa

Here, we performed experiments for fine-tuning the Urdu-RoBERTa transformer using eight hyper-parameters. We employed the Grid search technique to find out the better values of hyper-parameters. The list of the parameters and their ranges are presented in Table 4.10. For the fine-tuning process, the transformer model is trained, validated, and tested on the train, valid, and test part of the dataset. For this purpose, the dataset is split into 80–20 using stratified sampling technique. The 20% dataset is used for testing purposes and the remaining 80 % is further divided into 90–10, where 90 % is used for training and 10 % is used for validation purposes. The dataset is tokenized and transformed into the proper format to make it compatible with the input of the Urdu-RoBERTa model. We tried several configurations of hyper-parameters but presented only the best performances. The results are demonstrated in Table 4.11 and classifiers are evaluated using accuracy, and f1-score. In addition, class-wise f1- score, macro, and weighted average are

TABLE 4.10: Set of hyper-parameters and their values for fine-tuning process.

Hyperparameters	**Grid Search**
Sequence length	64, 128
Batch size	8, 16, 32, 64
Learning rate	3e-5, 2e-5, 1e-5, 1e-6
Weight decay	0.1
Warmup ratio	0.06–0.1
Hidden dropout	0.05, 0.1
Attention dropout	0.05, 0.1
Epochs	1–10

TABLE 4.11: Results of fine-tuning the Urdu-RoBERTa Model.

Learning Rate	**Hidden Dropout**	**Warm Up Ratio**	**Weight Decay**	**Accuracy**	**F1 Score**			
					Violence	**No Violence**	**Macro**	**Weighted**
2e-5	0.01	0.01	0.01	75.62	74.40	76.74	75.57	75.58
1e-5	0.05	0.05	0.01	73.33	72.65	73.98	73.31	73.34
1e-5	0.0003	0.01	0.01	75.83	74.89	76.70	75.79	75.80
1e-5	0.0001	0.01	0.01	75.84	75.42	76.22	75.81	75.82

computed. We tried all batch sizes and batch size of 32 demonstrated superior performance. Furthermore, the sequence length of 64, and 128 are explored and we got better performance with the 128 sequence length.

Several values of learning rates are explored but we added only those entries when we got better performances. It is evident from Table 13 that we got the best performance with a learning rate of 1e-5, 0.0001 value for hidden dropout, 0.01 warmup ratio, and 0.01 wt decay, the accuracy of 75.84 %, and macro f1-score of 75.82 % is achieved. Moreover, the class-wise performance for violence incitation and not-violence incitation is at least 75.42 %. This completes the fine-tuning process for the Urdu-RoBERTa model.

4.9 Comparison of DL Models

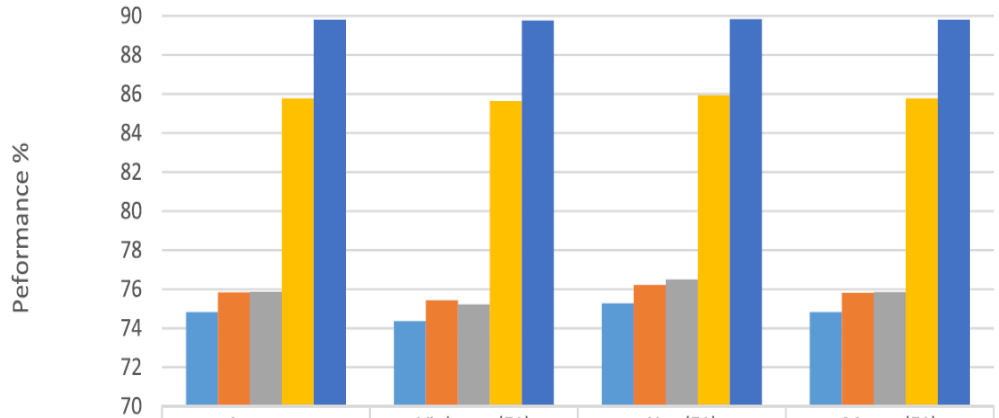
This section presents the comparison of three DL models using word uni-gram, Urdu-RoBERTa, and Urdu-BERT feature models. Furthermore, the results of the fine-tuned Urdu-RoBERTa model and best performing ML model are added

TABLE 4.12: Comparison of DL models using word uni-gram, RoBERTa, and BERT models.

Features	**Metrics**	**CNN**	**BiLSTM**	**CBi-LSTM**	**Best ML Model**	**Fine-tuned Urdu-RoBERTa**
Word Uni-gram	Recall	**89.84%**	50.72%	51.04%	**74.83%(RF)**	
	Precision	**89.76%**	51.65%	51.12%	**74.87%(RF)**	
	Accuracy	**89.84%**	50.68%	51.04%	**74.83%(RF)**	
	F1-score	**89.80%**	42.57%	50.18%	**74.83%(RF)**	
	AUC	**94.67%**	50.82%	50.99%	**74.83%(RF)**	
Urdu-RoBERTa	Recall	75.87%	53.91%	57.08%	69.50%(SVM)	75.82%
	Precision	75.95%	54.66%	57.19%	69.51%(SVM)	75.81%
	Accuracy	75.87%	53.91%	57.08%	69.50%(SVM)	75.84%
	F1-score	75.86%	51.98%	56.91%	69.50%(SVM)	75.81%
	AUC	83.16%	56.50%	59.71%	69.52%(SVM)	77.43%
Urdu-BERT	Recall	**85.78%**	55.97%	53.60%	69.38%(SVM)	
	Precision	**85.80%**	56.44%	56.14%	69.42%(SVM)	
	Accuracy	**85.79%**	55.97%	53.60%	69.38%(SVM)	
	F1-score	**85.78%**	55.17%	48.25%	69.39%(SVM)	
	AUC	**90.13%**	57.20%	56.25%	69.42%(SVM)	

and conclusions are drawn. Using word uni-gram, the CNN model demonstrated benchmark performance by obtaining 89.84 % accuracy and 89.80 % f1-score as shown in Table 4.12. Furthermore, the AUC value is very promising, i.e. 94.67 %. This is the highest performance obtained by the proposed methodology. On the other hand, BiLSTM and CBi-LSTM models did not perform significantly. The best performance demonstrated by the RF model with the word uni-gram is added in the second last column. The performance of fine-tuned Urdu-RoBERTa is added in the last column. Likewise, the performance of three DL models is also compared using Urdu-RoBERTa and Urdu-BERT transformers.

The CNN model obtained 75.87 % accuracy and 75.86 % f1-score with Urdu-RoBERTa (as feature model) and outperformed the two DL models, fine-tuned Urdu-RoBERTa and best ML model. The performance of the CNN model with Urdu-BERT is also promising and it outperformed the other supervised models by obtaining 85.79 % accuracy and 85.78 % f1-score. It is evident from Table 4.12 that the CNN model outperformed the fine-tuned Urdu-RoBERTa, BiLSTM, CBi-LSTM, and five ML models by improving the performance with substantial



	Accuracy	Violence (F1)	Not (F1)	Macro (F1)
RF (Uni-gram)	74.83	74.36	75.28	74.82
Fine-tuned RoBERTa	75.84	75.42	76.22	75.81
CNN (RoBERTa)	75.87	75.22	76.5	75.86
CNN (BERT)	85.78	85.64	85.93	85.78
CNN (Uni-gram)	89.8	89.76	89.84	89.8

FIGURE 4.2: Performance of best models in accuracy, macro f1-score, and individual class identification.

margins. Using Urdu-RoBERTa, CNN achieved an improvement of 0.02 % in accuracy, and 0.05 % in f1-score compared to fine-tuned Urdu-RoBERTa model. Similarly using word uni-gram, CNN improved the accuracy by 15.01 % and f1-score by 14.97 % compared to the SVM model (best performing). The highest improvement is observed on Urdu-BERT, where CNN demonstrated a 16.41 % improvement in accuracy and a 16.39 % improvement in f1-score. Thus violence incitation detection in the Urdu language with word uni-gram and CNN model demonstrated the state-of-the-art performance by achieving satisfactory metric values.

The purpose of the next experiment is to compare the performance of CNN, fine-tuned Urdu-RoBERTa, and RF models for the identification of individual class instances using the f1-score. We used violence class, notviolence class, macro average in f1-score, and accuracy measures for evaluation. As CNN and RF presented the best performance with word uni-gram features, and CNN demonstrated substantial performance with Urdu-BERT and Urdu-RoBERTa models, thus word uni-gram, Urdu-RoBERTa, and Urdu-BERT models are considered. The results are added in Fig. 4.2. CNN generated the best results with unigram features because unigrams provide the most fundamental representation of text, allowing

CNN to directly learn local patterns and important word-level features through its convolutional filters. CNNs are highly effective at capturing position-invariant features and detecting salient signals in short contexts, which makes unigrams particularly suitable, as they reduce complexity and sparsity compared to higher-order n-grams. This allows the CNN to focus on extracting discriminative features without being overwhelmed by noise or redundancy introduced by larger n-grams. Hence, the combination of unigram simplicity and CNN's ability to learn robust feature representations explains the superior performance.

Considering violence class identification, it is obvious that CNN is more effective compared to fine-tuned RoBERTa, applied DL, and five ML models. The RF obtained 74.36 % with word uni-gram, fine-tuned Urdu-RoBERTa obtained 75.84 %, but CNN obtained 89.76 % and improved 15.36 % for violence incitation class identification. Considering not-violence instances identification, CNN obtained 13.10 % better performance than the fine-tuned RoBERTa and RF model and 14.23 % better than fine-tuned RoBERTa and RF when the macro f1-score is considered. The second-best performance is observed with the CNN + Urdu-BERT model when individual class and macro f1-score are observed. On top of all, CNN + word uni-gram demonstrated the benchmark performance in individual class identification and macro f1-score compared to other models.

4.10 MARS Shine-Through and Occlusion

At last, we performed a different set of experiments to test the classifiers' predictions by analyzing the uniqueness of predicted class samples. The MARS shine-through can be measured for an individual model and a combined model. The performance of the five best classifiers using word uni-gram in MARS shine-through metric is presented in Table 4.13. Here, MARS shine through defines the "proportion of exclusive violence incitation samples predicted only by the classifier, to the total unique violence incitation samples predicted by all classifiers". The metric for the combined (two) model is calculated by merging the predictions of both classifiers.

Each cell in Table 4.14 demonstrates two values, the right value indicates the

score obtained by the two models intersecting the x-axis and y-axis, whereas the left value indicates the score of a single model pointing to the y-axis. It is visible that the best performance is obtained by the CNN model by predicting 4% unique violence incitation samples that are missed by other classifiers. In addition, CNN + SVM collectively predicted 7% of the total instances that are missed by all other classifiers. The SVM model is the second best model whereas the LR model predicted 0% unique instances as an individual model. The performance of MARS shine-through is presented visually in Fig. 4.3.

There are two circles against each intersection of the y-axis and x-axis. The inner circle indicates the performance of the individual classifier and its size shows the proportion of uniquely predicted violence incitation samples. The outer circle represents the combined model and its size shows the percentage of violence incitation samples predicted by the combined model. Next, the MARS occlusion measure is used to analyze the percentage of false negative samples observed by a classifier or two classifiers. The performance of five classifiers using word uni-gram features in the MARS occlusion metric is described in Table 4.14.

Here, the minimum value is the best-performing criterion. CNN again demonstrated the best performance compared to other classifiers by identifying 15 % false negative samples. The AdaBoost showed the prediction of 22 % false negative samples and it is the lowest performance. In addition, when NN is combined with the SVM model, it only missed 5 % of true positive samples, showing the highest performance. Thus MARS shine-through and MARS occlusion measures highlight the outperformance of the CNN model. These findings justify the robustness of the CNN model for violence incitation identification task in the Urdu language.

4.11 Time Complexity Analysis

In order to evaluate the computational efficiency of our CNN model for violence incitation detection, we analyzed the effect of batch size, number of samples, and number of epochs on both training and testing time.

TABLE 4.13: Classifiers evaluation using MARS Shine-through matrix (with word uni-gram).

	Classifiers	**X-Axis**				
		AdaBoost	**CNN**	**LR**	**RF**	**SVM**
Y-Axis	AdaBoost		0.01 0.06	0.01 0.03	0.01 0.03	0.01 0.05
	CNN	0.04 0.06		0.04 0.04	0.04 0.06	0.04 0.07
	LR	0.00 0.03	0.00 0.04		0.00 0.02	0.00 0.04
	RF	0.01 0.03	0.01 0.06	0.01 0.02		0.01 0.05
	SVM	0.03 0.05	0.03 0.07	0.03 0.04	0.03 0.05	

TABLE 4.14: Classifiers evaluation using MARS Occlusion matrix (with word uni-gram).

	Classifiers	**X-Axis**				
		AdaBoost	**CNN**	**LR**	**RF**	**SVM**
Y-Axis	AdaBoost		0.22 0.08	0.22 0.12	0.22 0.11	0.22 0.07
	CNN	0.15 0.08		0.15 0.08	0.15 0.07	0.15 0.05
	LR	0.18 0.12	0.18 0.08		0.18 0.10	0.18 0.10
	RF	0.17 0.11	0.17 0.07	0.17 0.10		0.17 0.09
	SVM	0.16 0.07	0.16 0.05	0.16 0.10	0.16 0.09	

4.11.1 Effect of Batch Size

The results in table 4.15, 4.4, 4.16, and 4.5 indicate that increasing the batch size significantly reduces training time. For instance, training with a batch size of 8 required 7427.56 seconds, whereas a batch size of 256 reduced training time to 1485.51 seconds. This trend demonstrates the efficiency gains from processing larger batches in parallel. However, testing time remained relatively stable across different batch sizes, with a slight increase from 4.24 seconds (batch size 8) to 4.71 seconds (batch size 256), suggesting that inference performance is largely independent of batch size.

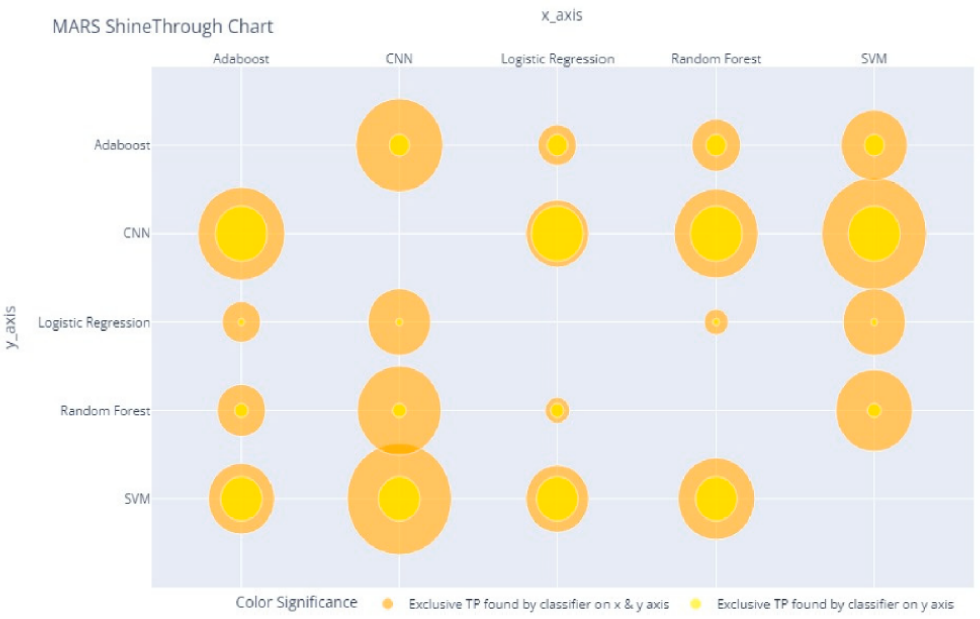


FIGURE 4.3: MARS shine-through chart.

TABLE 4.15: Time complexity of varying batch size for training time).

Batch Size	**Training Time**
8	7427.56
16	3927.08
32	2607.89
64	2000.67
128	1707.95
256	1485.51

TABLE 4.16: Time complexity of varying batch size for testing time).

Batch Size	**Testing Time**
8	4.24
16	4.05
32	4.12
64	4.15
128	4.33
256	4.71

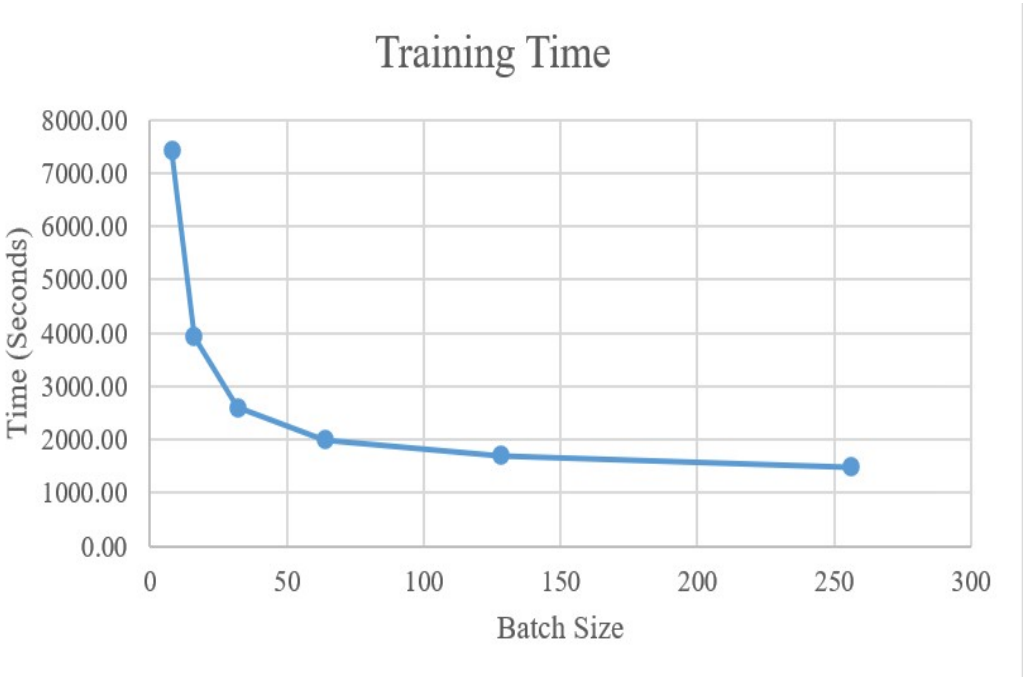


FIGURE 4.4: Time complexity of varying batch size for training time.

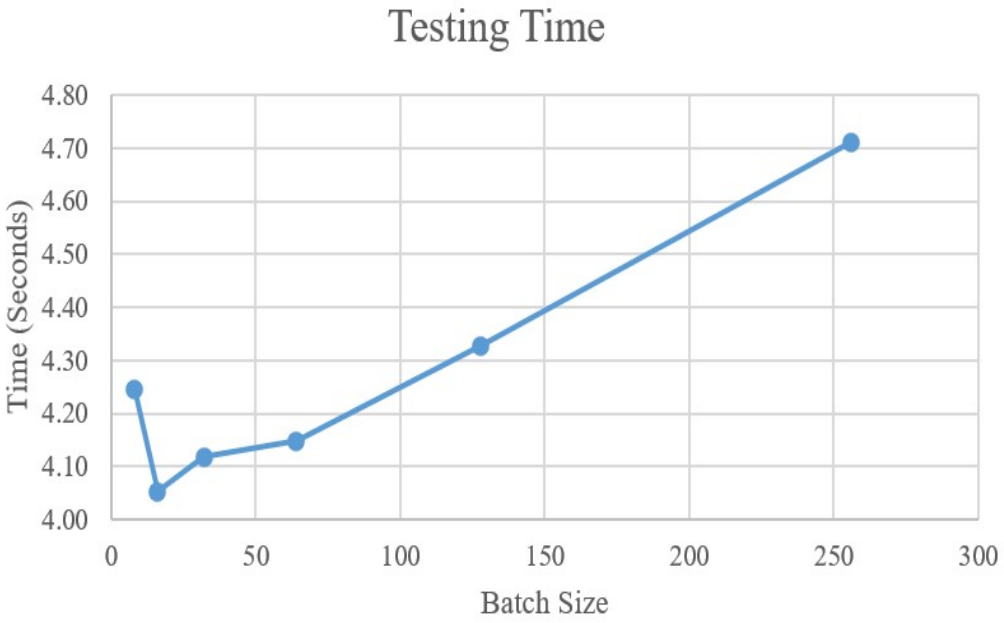


FIGURE 4.5: Time complexity of varying batch size for testing time.

TABLE 4.17: Time complexity of varying samples for training time).

Samples	Training Time
500	70.71
1000	125.67
1500	180.88
2000	237.33
2500	284.71
3000	340.30
3500	396.97
4000	438.26

TABLE 4.18: Time complexity of varying samples for testing time).

Samples	Testing Time
3804	7.20
3304	6.38
2804	5.37
2304	4.55
1804	3.47
1304	2.59
804	1.72
304	0.72

4.11.2 Effect of Number of Samples

As expected, training time 4.17 and 4.6 scaled linearly with the number of samples. With 500 samples, training took only 70.71 seconds, while training with 4000 samples required 438.26 seconds. This confirms that the training cost increases proportionally to dataset size. Testing time 4.18 and 4.7 also followed a similar trend, ranging from 0.72 seconds for 304 samples to 7.20 seconds for 3804 samples, demonstrating that both training and testing times are directly affected by dataset size.

4.11.3 Effect of Epochs

The number of training epochs 4.19 and 4.8 had the most pronounced impact on training time. At 5 epochs, the training process required 1333.81 seconds, whereas training for 30 epochs took 12378.83 seconds. This exponential growth reflects the repeated passes over the dataset during training. Interestingly, testing time 4.20

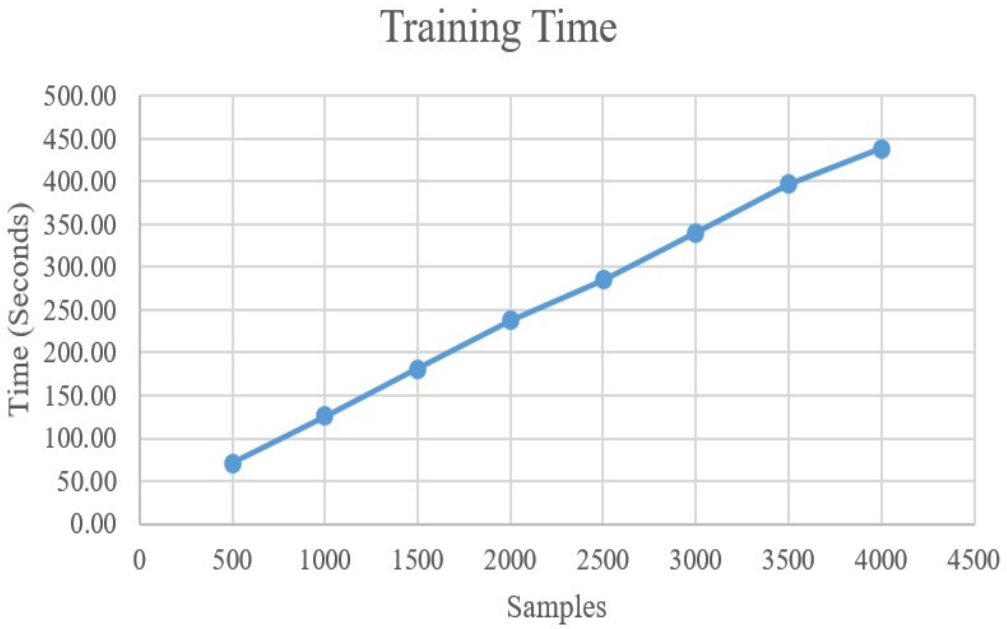


FIGURE 4.6: Time complexity of varying samples for training time.

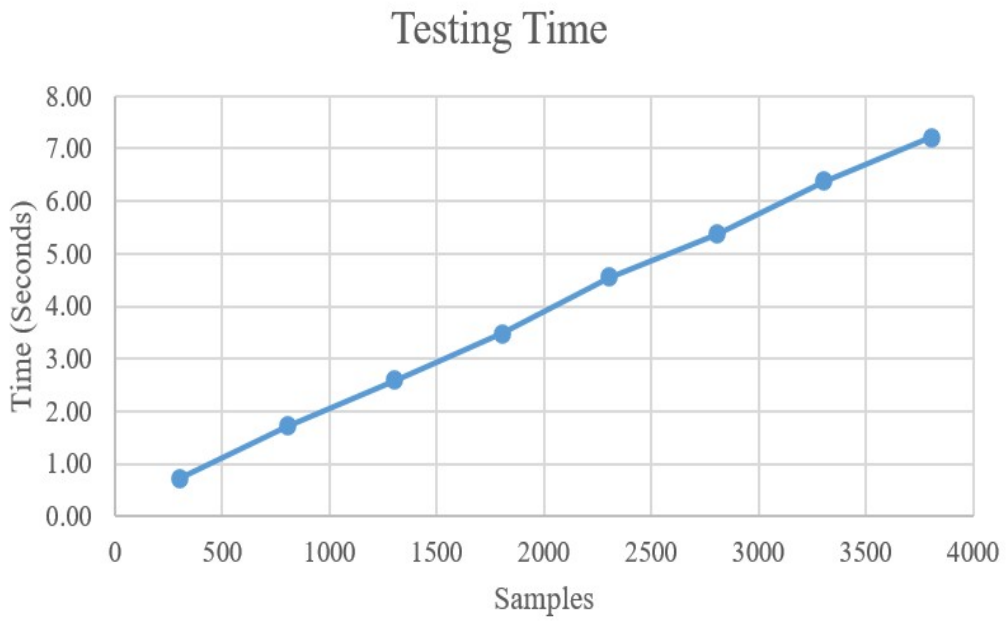


FIGURE 4.7: Time complexity of varying samples for testing time.

TABLE 4.19: Time complexity of varying epochs for training time.

Epochs	Training Time
5	1333.81
10	2939.17
15	4734.35
20	5297.12
25	6515.14
30	12378.83

TABLE 4.20: Time complexity of varying epochs for testing time.

Epochs	Testing Time
5	4.35
10	4.26
15	4.16
20	4.05
25	4.23
30	5.89

and 4.9 remained relatively stable across most settings (around 4–4.3 seconds) but increased slightly at 30 epochs (5.89 seconds), which could be attributed to model complexity and parameter optimization at higher epochs.

Overall, the experiments reveal that batch size plays a critical role in reducing

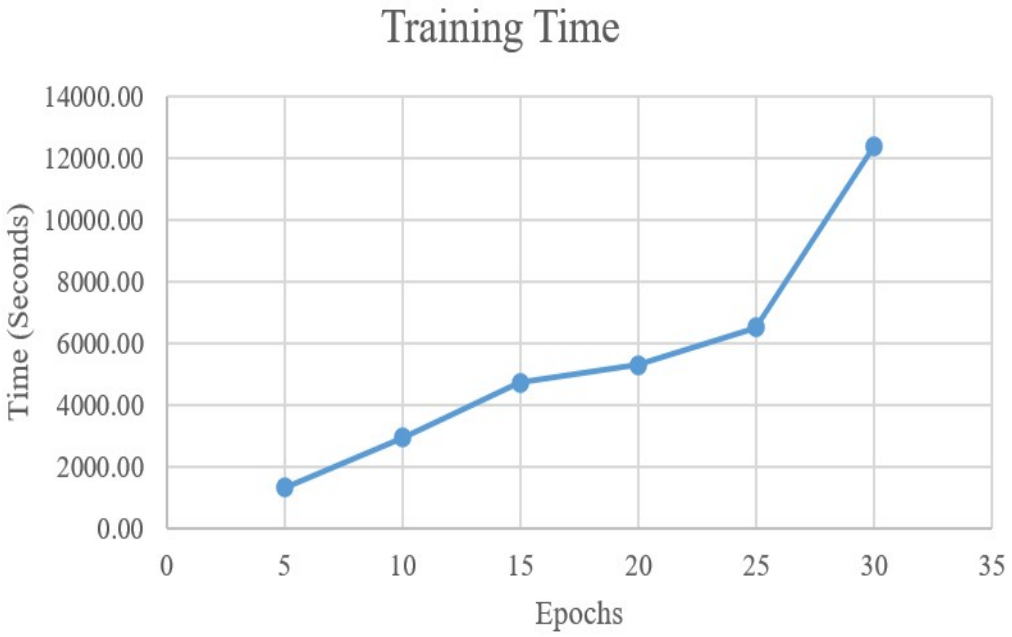


FIGURE 4.8: Time complexity of varying epochs for training time.

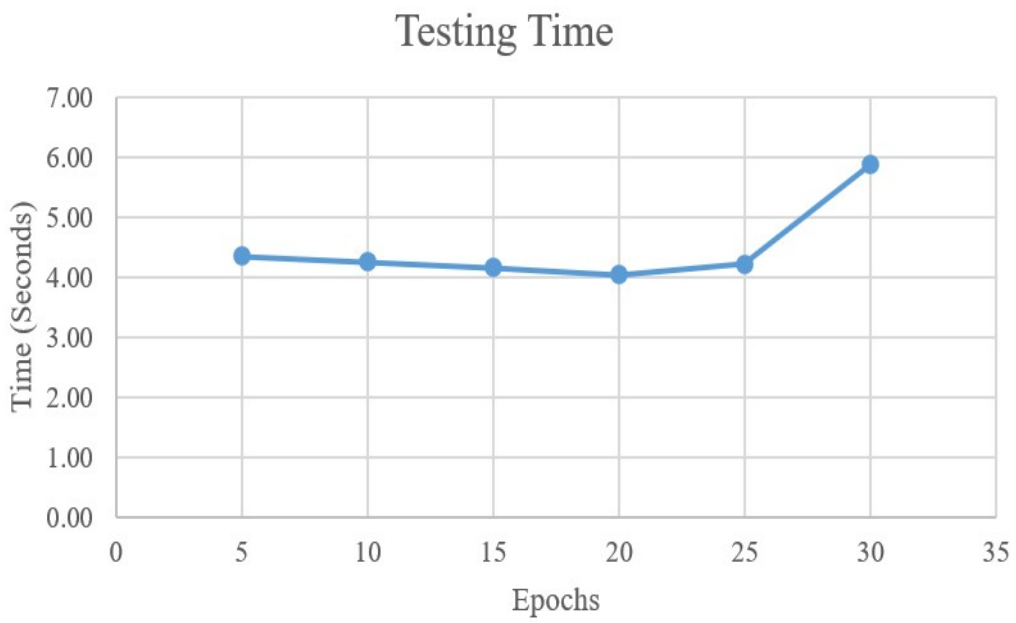


FIGURE 4.9: Time complexity of varying epochs for testing time.

training time without significantly impacting testing efficiency. Larger datasets increase both training and testing time in a near-linear manner. Epochs directly scale training time, with marginal effects on testing. These findings highlight the trade-off between computational cost and model performance in CNN-based violence incitation detection. Optimizing batch size and selecting an appropriate number of epochs can lead to substantial improvements in training efficiency without compromising testing speed.

4.12 Violence Target Group Identification

To identify target group from the violence tweets, we have used the same parameters and produced the following results using word unigram (7,973), bigrams (37,424), trigrams (47,524) and combined (1-2-3) grams (92,921) using best performing ML algorithms like Logistic Regression, Random Forest, and Support Vector Machine as shown in previous results of violence incitation detection. Among the classifiers, Gaussian Naïve Bayes (GNB), Logistic Regression (LR), Support Vector Machines (SVM), AdaBoost, and Random Forest (RF) consistently outperform others across different feature sets and target groups. These classifiers leverage different underlying algorithms and decision-making mechanisms, allow-

ing them to capture complex patterns and relationships within the data effectively. Logistic Regression stands out for its simplicity, interpretability, and robust performance across various feature sets, making it a reliable choice for violence incitation detection tasks.

Random Forest, with its ensemble-based approach, demonstrates high flexibility and resilience to overfitting, resulting in consistently strong performance across different feature sets and target groups. SVM, known for its effectiveness in high-dimensional spaces, achieves competitive performance, especially when combined with appropriate kernel functions. AdaBoost leverages the concept of boosting to iteratively improve classification accuracy by focusing on challenging instances, leading to enhanced performance, particularly for imbalanced datasets. These classifiers' ability to generalize well to different feature representations and target groups makes them suitable choices for violence incitation detection tasks across diverse contexts and datasets.

Table 4.21: Target Group Identification: Comparison of word n-gram features using three ML models.

Features	**Metrics**	**Target Group**	**LR**	**RF**	**SVM**
Word Uni-gram	Recall	1	85.66%	90.59%	91.47%
		2	85.58%	91.88%	91.56%
		3	90.95%	93.65%	93.90%
	Precision	1	87.19%	95.64%	89.81%
		2	87.70%	94.00%	90.56%
		3	81.84%	88.31%	85.90%
	Accuracy	1	86.62%	93.21%	90.56%
		2	86.76%	93.00%	91.00%
		3	85.38%	90.59%	89.26%
	F1-Score	1	86.39%	93.02%	90.60%
		2	86.58%	92.92%	91.04%

Continued on next page

Table 4.21: Target Group Identification: Comparison of word n-gram features using three ML models. (Continued)

Features	**Metrics**	**Target Group**	**LR**	**RF**	**SVM**
		3	86.11%	90.88%	89.70%
	AUC	1	86.60%	93.21%	90.54%
		2	86.77%	93.00%	90.97%
		3	85.37%	90.59%	89.26%
Word Bi-gram	Recall	1	85.55%	70.35%	89.84%
		2	85.30%	71.06%	90.03%
		3	91.23%	93.12%	90.82%
	Precision	1	93.83%	95.34%	92.63%
		2	94.46%	94.47%	94.05%
		3	79.54%	75.53%	85.79%
	Accuracy	1	90.00%	83.44%	91.38%
		2	90.18%	83.44%	92.15%
		3	83.85%	81.44%	87.88%
	F1-Score	1	89.45%	80.91%	91.18%
		2	89.62%	81.08%	91.98%
		3	84.93%	83.38%	88.19%
	AUC	1	90.00%	83.44%	91.42%
		2	90.15%	83.44%	92.16%
		3	83.89%	81.44%	87.91%
Word Tri-gram	Recall	1	82.02%	52.71%	86.08%
		2	81.02%	41.47%	86.33%
		3	91.92%	38.88%	89.47%
	Precision	1	95.53%	95.18%	95.69%

Continued on next page

Table 4.21: Target Group Identification: Comparison of word n-gram features using three ML models. (Continued)

Features	**Metrics**	**Target Group**	**LR**	**RF**	**SVM**
		2	97.95%	94.66%	95.63%
		3	72.21%	89.54%	84.84%
	Accuracy	1	89.09%	75.00%	91.09%
		2	89.65%	69.56%	91.21%
		3	78.21%	67.15%	86.79%
	F1-Score	1	88.20%	67.79%	90.59%
		2	88.65%	57.55%	90.73%
		3	80.81%	54.17%	87.07%
	AUC	1	89.10%	75.00%	91.14%
		2	89.66%	69.56%	91.19%
		3	78.26%	67.15%	86.78%
Combined Word (1-2-3) gram	Recall	1	91.10%	84.82%	90.34%
		2	91.03%	84.53%	91.38%
		3	93.26%	92.76%	91.20%
	Precision	1	91.76%	96.96%	94.98%
		2	92.81%	95.68%	95.44%
		3	85.15%	86.59%	89.57%
	Accuracy	1	91.50%	91.06%	**92.79%**
		2	91.97%	90.35%	**93.50%**
		3	88.50%	89.18%	**90.29%**
	F1-Score	1	91.40%	90.46%	92.58%
		2	91.89%	89.74%	93.35%
		3	88.98%	89.56%	90.33%

Continued on next page

Table 4.21: Target Group Identification: Comparison of word n-gram features using three ML models. (Continued)

Features	**Metrics**	**Target Group**	**LR**	**RF**	**SVM**
	AUC	1	91.56%	91.06%	92.84%
		2	92.00%	90.35%	93.52%
		3	88.50%	89.18%	90.29%

Based on the previous experiments conducted for violence incitation identification, the top three best performing machine learning classifiers (Logistic Regression, Random Forest, and Support Vector Machines) were selected for a second series of experiments focused on violence incitation target detection. Among these selected classifiers, Support Vector Machines (SVM) exhibited the highest recall values across different target groups shown in Table 4.21. Specifically, SVM achieved the highest recall of 93.90% for target group 3 and 91.47% for target group 1, whereas Random Forest (RF) outperformed for target group 2 with a recall of 91.88% using word unigram features. In terms of precision, Random Forest (RF) demonstrated superior performance for target groups 1 and 2, achieving precision values of 96.96% and 95.68%, respectively. Meanwhile, SVM achieved the highest precision of 89.57% for target group 3 using word combined (1-2-3) grams features. Accuracy, an important evaluation measure for violence incitation target detection, was highest when using SVM for all three target groups, with accuracy values of 92.79% for target group 1, 93.50% for target group 2, and 90.29% for target group 3 using word combined (1-2-3) grams features.

Additionally, SVM demonstrated the best performance in terms of F1-score, achieving the highest values for all three target groups using word combined (1-2-3) grams features. Specifically, SVM achieved F1-scores of 92.58%, 93.35%, and 90.33% for target groups 1, 2, and 3, respectively. Moreover, SVM also outperformed other classifiers in terms of Area under the Curve (AUC) values, achieving the highest AUC values for all three target group identifications using word combined (1-2-3) grams features, with values of 92.84%, 93.52%, and 90.29% for target groups 1, 2, and 3, respectively. In conclusion, while word unigrams exhibited higher recall values for all three target groups, word combined (1-2-3) grams features consis-

tently yielded the best results across other evaluation measures. SVM consistently demonstrated the best performance across all evaluation metrics, while Logistic Regression (LR) exhibited relatively lower scores.

Table 4.22: Target Group Identification Comparison of char n-gram features using three ML models.

Features	**Metrics**	**Target Group**	**LR**	**RF**	**SVM**
Char Uni-gram	Recall	1	53.00%	80.18%	59.13%
		2	58.60%	77.53%	62.90%
		3	67.84%	84.41%	73.81%
	Precision	1	64.35%	80.20%	66.91%
		2	62.24%	78.19%	63.11%
		3	61.97%	79.54%	63.98%
	Accuracy	1	61.85%	80.15%	64.94%
		2	61.50%	77.91%	63.03%
		3	63.06%	81.32%	66.06%
	F1-Score	1	58.08%	80.15%	62.65%
		2	60.31%	77.80%	62.95%
		3	64.66%	81.83%	68.45%
	AUC	1	61.80%	80.15%	65.00%
		2	61.52%	77.91%	63.06%
		3	63.09%	81.32%	66.11%
Char Bi-gram	Recall	1	80.79%	88.35%	90.94%
		2	83.27%	89.00%	91.65%
		3	85.20%	88.18%	91.30%
	Precision	1	81.35%	97.07%	88.00%
		2	80.75%	96.83%	88.22%
		3	77.87%	92.87%	85.67%

Continued on next page

Table 4.22: Target Group Identification Comparison of char n-gram features using three ML models. (Continued)

Features	**Metrics**	**Target Group**	**LR**	**RF**	**SVM**
	Accuracy	1	81.18%	92.82%	89.29%
		2	81.71%	93.03%	89.71%
		3	80.47%	90.71%	88.06%
	F1-Score	1	81.02%	92.48%	89.41%
		2	81.94%	92.73%	89.87%
		3	81.33%	90.44%	88.38%
	AUC	1	81.13%	92.82%	89.31%
		2	81.69%	93.03%	89.71%
		3	80.47%	90.71%	88.05%
Char Tri-gram	Recall	1	91.18%	91.12%	91.70%
		2	91.33%	92.59%	92.36%
		3	93.26%	90.12%	91.86%
	Precision	1	88.49%	94.32%	92.30%
		2	89.12%	93.95%	93.60%
		3	84.43%	90.87%	88.94%
	Accuracy	1	89.71%	92.79%	92.03%
		2	90.12%	93.29%	93.03%
		3	88.06%	90.53%	90.24%
	F1-Score	1	89.79%	92.67%	91.97%
		2	90.19%	93.24%	92.95%
		3	88.60%	90.48%	90.35%
	AUC	1	89.73%	92.79%	92.08%
		2	90.11%	93.29%	93.05%

Continued on next page

Table 4.22: Target Group Identification Comparison of char n-gram features using three ML models. (Continued)

Features	**Metrics**	**Target Group**	**LR**	**RF**	**SVM**
		3	88.08%	90.53%	90.26%
Combined Char (1-2-3) gram	Recall	1	91.84%	90.00%	84.86%
		2	91.43%	90.24%	85.85%
		3	93.04%	89.35%	87.67%
	Precision	1	87.67%	96.76%	96.35%
		2	87.64%	96.13%	98.11%
		3	84.61%	91.22%	92.67%
	Accuracy	1	89.47%	**93.47%**	90.79%
		2	89.26%	**93.29%**	92.09%
		3	88.06%	90.38%	**90.41%**
	F1-Score	1	89.67%	93.24%	90.17%
		2	89.46%	93.08%	91.54%
		3	88.58%	90.26%	90.09%
	AUC	1	89.48%	93.47%	90.89%
		2	89.28%	93.29%	92.10%
		3	88.09%	90.38%	90.41%

The results depicted in the graph 4.10 highlight the comparative accuracy achieved using word combined (1-2-3) grams features across three classifiers: Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM) for three distinct target groups. The analysis reveals that SVM consistently outperforms the other classifiers, achieving the highest accuracy of 93.50% specifically for target group 2. This underscores SVM's robustness in capturing the nuances of the data, making it a superior predictor for this group. Additionally, the graph illustrates that SVM has demonstrated strong performance across all three target

Accuracy			
Classifier	Target Group 1	Target Group 2	Target Group 3
Logistic Regression	91.50	91.97	88.50
Random Forest	91.06	90.35	89.18
SVM	92.79	93.50	90.29

FIGURE 4.10: Performance of word combined (1-2-3) grams features with best models in accuracy for target groups.

groups, further establishing its reliability in violence incitation detection. Furthermore, the results indicate that all three classifiers—LR, RF, and SVM—perform better for target group 1 and target group 2 compared to target group 3. This suggests that target group 3 may present unique challenges or variability in the data that complicates accurate classification. Overall, SVM's dominance in accuracy across the board reinforces its suitability for this classification task, while the performance differences among the target groups highlight areas for potential improvement and further investigation.

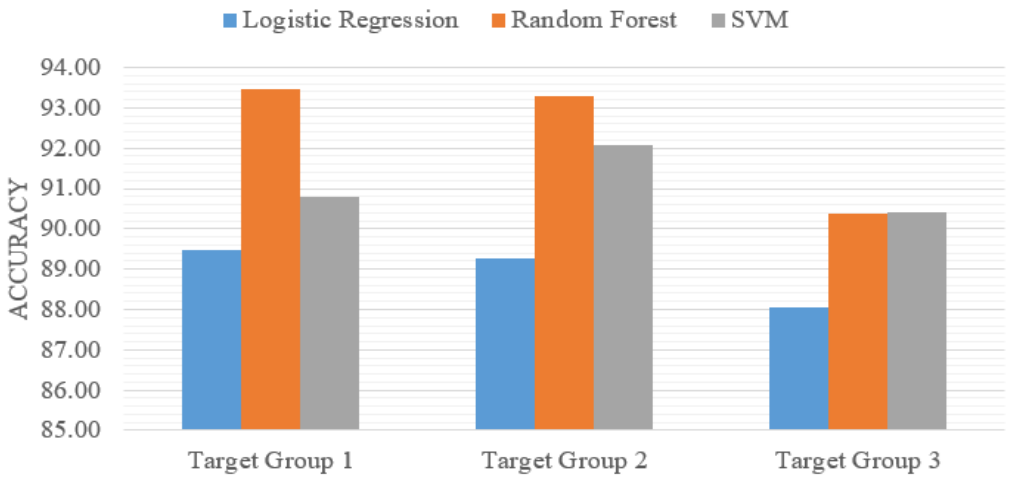
In the second set of experiments conducted on character n-grams features, char uni-gram (57), char bi-grams (1,468), char tri-gram (15,592), combined char (1-2-3) grams (17,174), the same selected machine learning classifiers (LR, RF, and SVM) were employed with identical parameters as shown in Table 4.22. Notably, the char combined (1-2-3) grams feature set emerged as the superior feature set for violence incitation target identification. Logistic Regression (LR) achieved the highest recall of 93.26% for target group 3 using char trigrams, followed closely by RF with a recall of 92.59% for target group 2, also with char trigrams. LR also performed well in achieving a recall of 91.84% for target group 1 using char combined (1-2-3) grams. For precision, Support Vector Machines (SVM) demonstrated the best

performance, achieving the highest precision of 98.11% for target group 2 with char combined (1-2-3) grams.

Random Forest (RF) secured the second position in precision, achieving values of 97.07% for target group 1 and 92.87% for target group 3 using char bigrams features. In terms of accuracy, char combined (1-2-3) grams features exhibited superior performance for identifying violence incitation target groups 1 and 2, achieving accuracies of 93.47% and 93.29%, respectively. However, for target group 1, char bigrams features yielded the best accuracy of 90.71% with RF, indicating RF's significant role in achieving the highest accuracy for all three target groups. Similarly, RF achieved the best F1-scores for all three target groups, with values of 93.24% for target group 1 with char combined (1-2-3) grams, the same for target group 2 with char trigrams, and 90.48% for target group 3 with char trigrams. RF consistently outperformed in terms of F1-score across all target groups. Furthermore, in terms of the Area Under the Curve (AUC) metric, RF again demonstrated superior performance, achieving the highest AUC values of 93.47% for target group 1 and 93.29% for target group 2 with char combined (1-2-3) grams, while char bigrams facilitated the identification of the best AUC for target group 3 with a value of 90.71%.

The results displayed in the graph 4.11 present the accuracy achieved using char combined (1-2-3) grams features across three classifiers: Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM) for three distinct target groups. The experiments show that Random Forest outperformed the other models for target group 1 and target group 2, achieving the highest accuracy of 93.47% for target group 1. This indicates that RF is particularly effective in identifying violence incitation for these groups. However, for target group 3, SVM achieved better results, showcasing its adaptability and precision in handling more complex or varied data within this group.

Additionally, the results emphasize that all three classifiers, with char combined grams features, proved to be better predictors for target group 1 and target group 2 compared to target group 3, mirroring the performance trend observed with word combined (1-2-3) grams features. This consistency suggests inherent characteristics or challenges within target group 3 that affect classification accuracy.



Accuracy			
Classifier	Target Group 1	Target Group 2	Target Group 3
Logistic Regression	89.47	89.26	88.06
Random Forest	93.47	93.29	90.38
SVM	90.79	92.09	90.41

FIGURE 4.11: Performance of char combined (1-2-3) grams features with best models in accuracy for target groups.

Furthermore, while SVM is the second-best predictor with char grams features, it remains the top predictor with word grams features, underscoring its overall robustness. These findings highlight that RF excels with char grams for specific target groups, whereas SVM demonstrates superior versatility across different feature sets.

In summary, both sets of experiments focused on identifying violence incitation targets using different feature sets and machine learning classifiers. In the first set of experiments with word n-grams features, Support Vector Machines (SVM) consistently demonstrated the best performance across various evaluation metrics, including recall, precision, accuracy, F1-score, and Area Under the Curve (AUC). SVM achieved the highest scores for most of the target groups across different features, particularly excelling with word combined (1-2-3) grams. On the other hand, the second set of experiments utilized character n-grams features, where Random Forest (RF) emerged as the top-performing classifier. RF consistently outperformed other classifiers across most evaluation metrics, demonstrating high recall, precision, accuracy, F1-score, and AUC values.

The char combined (1-2-3) grams feature set proved to be particularly effective in identifying violence incitation target groups, with RF achieving the highest scores

for all three target groups. Comparing both sets of experiments, RF performed exceptionally well in the second set with character n-grams features, while SVM stood out in the first set with word n-grams features. The choice between RF and SVM depends on the specific requirements and priorities of the task at hand. RF may be preferred for its robust performance with character n-grams features, while SVM could be chosen for its reliability and consistency with word n-grams features. Overall, both classifiers demonstrated strong performance in identifying violence incitation targets, highlighting the importance of selecting appropriate feature sets and classifiers based on the nature of the data and the task.

4.13 Threats to Validity

In any experimental research, it is essential to acknowledge the potential threats to the validity of the results. For my research on violence incitation detection from Urdu content on social media, several factors could impact the generalizability and reliability of the findings.

4.13.1 Language-Specific Constraints

The primary threat to the validity of my experimental results lies in the language specificity of the dataset. The research is focused on Urdu, a language predominantly spoken in Pakistan and parts of India. Consequently, the models and methods developed are tailored to the nuances of Urdu language, including its syntax, semantics, and idiomatic expressions. This focus inherently limits the generalizability of the results to other languages. For instance, a model trained and tested on Urdu text might not perform effectively on Arabic, English, or any other language without significant adaptation and retraining. The unique characteristics of the Urdu language mean that the techniques and models may not transfer seamlessly to texts in other languages.

4.13.2 Dataset Limitations

The dataset used in this research comprises Urdu tweets specifically collected from Pakistani Twitter handles, including those of newspapers and individuals. This

geographical and contextual limitation means the findings are representative of the specific socio-political environment of Pakistan. The dataset may not capture the full spectrum of violence incitation that could be present in other regions or contexts. Additionally, the data collection process might inadvertently introduce bias, such as over-representation of certain viewpoints or under-representation of others, thereby skewing the results. An example of over-representation in the dataset could be the disproportionate presence of tweets from mainstream media outlets or prominent political figures, as their content tends to be more frequently posted and shared. This may overshadow grassroots voices or perspectives from marginalized communities. On the other hand, under-representation may occur with tweets in regional Urdu dialects or from rural users who are less active on Twitter or have limited internet access. As a result, the dataset might not fully capture the linguistic diversity and localized narratives of violence incitation across Pakistan.

4.13.3 Data Annotation Bias

Another significant threat to the validity of the experimental results is the potential bias introduced during data annotation. Despite extensive training sessions and detailed annotation guidelines provided to the data annotators through both online and physical meetings, human biases can never be entirely eliminated. Annotators may have subconscious biases towards certain factors, such as political views, cultural norms, or personal beliefs, which could affect the consistency and accuracy of the annotations. Such biases can lead to variability in the labeled data, impacting the performance and evaluation of the models.

4.13.4 Cultural Nuances

Urdu, being a rich and culturally embedded language, contains words and expressions that can have different meanings based on context, culture, or region. Proverbs, poetry, and idiomatic expressions are particularly challenging as they often carry metaphorical or dual meanings that can be easily misinterpreted. Additionally, certain words or phrases might have specific cultural connotations that

are not universally understood. These subtleties pose a significant threat to the accuracy of violence incitation detection, as the models may struggle to discern the intended meaning behind culturally nuanced language, potentially leading to misclassification.

4.13.5 Computational Limitations

The computational resources used for conducting experiments also present a threat to the validity of the results. Due to constraints in processing power and the availability of advanced computing infrastructure, the experiments were conducted using Google Colab, a cloud-based platform that provides limited resources compared to dedicated high-performance computing environments. The limitations of Google Colab, such as restricted GPU availability, memory constraints, and potential interruptions, might affect the thoroughness and depth of the experiments. More extensive computational resources could allow for more exhaustive hyperparameter tuning, larger model architectures, and longer training periods, potentially leading to improved results.

4.13.6 Hyperparameter Tuning and Model Architecture

The process of hyperparameter tuning, including the selection of the number of hidden layers, learning rates, batch sizes, and other critical parameters, is another area that could introduce variability in the results. The models' performance is highly sensitive to these hyperparameters, and the tuning process on Google Colab might not have been as comprehensive as desired due to time and resource constraints. This could mean that the models are not fully optimized, and better performance might be achievable with more extensive tuning. While the experimental results presented in this dissertation provide valuable insights into violence incitation detection in Urdu social media content, several threats to validity must be acknowledged. These include language-specific constraints, dataset limitations, potential biases in data annotation, computational resource limitations, and the extent of hyperparameter tuning. Recognizing these threats is crucial for interpreting the results appropriately and understanding the scope and limitations of

the research. Future studies should aim to address these threats by expanding the dataset to include more diverse sources, employing more robust annotation protocols, utilizing more powerful computational resources, and performing more exhaustive hyperparameter tuning to further validate and enhance the findings.

4.14 Conclusion

The results and analysis presented in this chapter provide valuable insights into the effectiveness of various machine learning models for violence incitation detection and target group identification in Urdu text. The experiments demonstrate that feature selection plays a crucial role in model performance, with word and character n-grams contributing differently across classifiers. The findings highlight that Support Vector Machine (SVM) and Random Forest (RF) have shown superior performance across different target groups, with variations depending on the feature set used. Additionally, the analysis underscores the challenges posed by language-specific characteristics, such as cultural context, figurative expressions, and dual meanings, which can impact classification accuracy. The interpretability of the models using explainable AI techniques further provides a deeper understanding of how predictions are made, ensuring transparency in decision-making. These results contribute to the growing body of research in violence incitation detection for low-resource languages and set a foundation for future enhancements through improved feature engineering, dataset expansion, and advanced deep learning techniques.

Chapter 5

Conclusion and Future Work

In this dissertation, we have comprehensively addressed three primary research questions concerning the identification of violence incitation from Urdu content using social media platforms and the subsequent identification of targeted communities or groups.

For Research Question 1, which pertains to the development of a violence incitation dataset from Urdu content using social media platforms, we conducted a meticulous process of dataset creation. This involved the selection of Twitter as primary data source due to its advantages in authenticity, text size, and user engagement. We described the detailed steps involved in crawling data from Twitter, including the selection of relevant accounts such as newspapers, politicians, and religious scholars, as well as the development of Urdu lexicons to filter and collect relevant tweets containing violence incitation content. Moreover, we utilized Python bots and various data preprocessing techniques to clean and organize the collected data, resulting in a robust dataset ready for further analysis.

In response to Research Question 2, which focuses on the identification of violence incitation from Urdu content using textual and contextual characteristics, we employed a diverse range of machine learning and deep learning models. Through feature extraction techniques such as word n-grams, character n-grams, Latent Semantic Analysis, Word2Vec, FastText, Urdu-BERT, and Urdu-RoBERTa, we generated rich feature representations of the textual data. Subsequently, we applied classifiers such as Logistic Regression, Random Forest, Support Vector Machines, CNN, LSTM, GRU, BiLSTM, and CBiLSTM to effectively classify and

detect violence incitation instances within the dataset. By evaluating the performance of these models across various metrics such as recall, precision, accuracy, F1-score, and AUC, we provided comprehensive insights into the efficacy of different approaches for violence incitation detection in Urdu content.

Lastly, for Research Question 3, which revolves around identifying targeted communities or groups using violence incitation detection, this research extended analysis to classify and categorize the identified instances of violence incitation into distinct target groups. Leveraging the capabilities of machine learning models such as Logistic Regression, Random Forest, and Support Vector Machines, we successfully identified and categorized targeted groups based on their characteristics and affiliations. Through meticulous annotation and classification processes, we elucidated the role of violence incitation in targeting specific communities or groups within the Urdu-speaking population. In conclusion, this dissertation endeavors to provide comprehensive insights and solutions to the complex challenges surrounding the identification of violence incitation from Urdu content on social media platforms, as well as the subsequent identification of targeted communities or groups.

Through rigorous experimentation, analysis, and evaluation, we have made significant strides towards addressing each of the research questions posed, thereby contributing to the broader discourse on online extremism detection and prevention. Text classification is a substantial step to extracting knowledge and new insights from the abundance of data for several purposes such as automating business processes etc. Moreover, the need for an effective and accurate toxic detection system is highly demanding due to the plethora of unwanted texts that are constantly posted on social media. To address part of this issue, this dissertation performed a comparative analysis by evaluating the proposed violence incitation identification model on newly designed Urdu corpus. Why does violence incitation need to be addressed on a priority basis? Because it is promoting unrest and extremism in the society.

As per current knowledge, some of the effects of violence incitation are discussed. Societal Impact, the pervasive nature of social media platforms leads to the rapid dissemination of violent content, which in turn influences public opinion, increases

social tensions, and fosters a culture of fear and hostility. Political Ramifications, incitement to violence is involved in influencing political events and opinions, which is a threat to the democratic process and social stability. Identifying and mitigating such content is critical to maintaining the integrity of political discourse. Religious Extremism, instances of violence incitement can heighten tensions, escalate conflicts, and contribute to an atmosphere of mistrust in the religious community. This situation needs to be handled very carefully to promote mutual and interfaith understanding.

5.1 Research Contribution

Our research advances the current state of NLP for the low-resource language Urdu in multiple ways.

5.1.1 Violence Detection

First , most prior work on violence detection has focused on high-resource languages like English, while Urdu has very limited annotated datasets and NLP frameworks. By curating and experimenting on Urdu-specific data for violence incitation detection, we have filled this research gap.

5.1.2 Evaluate of ML and DL in Violence Incitation

Secondly, we systematically evaluate both traditional ML (Random Forest with uni-grams) and deep learning (CNN, Transformers like Urdu-BERT and Urdu-RoBERTa) approaches, providing a comparative performance benchmark for Urdu.

5.1.3 CNN Performance in Violence Incitation

Importantly, our findings show that CNN with simple uni-grams outperformed transformer-based approaches (achieving 89.8% F1), revealing valuable insights that even lightweight models can be highly effective for low-resource contexts.

5.1.4 Linguistic Challenges

Finally, by explicitly tackling linguistic challenges of Urdu (morphological richness, implicit expressions, cultural nuances), our work lays the groundwork for building robust frameworks to detect violence incitation and target-specific harms in low-resource languages.

5.2 Conclusion

In this dissertation, the critical problem of violence incitation detection in the low-resource Urdu language is systematically addressed, with a focus on both the detection of violent discourse and the identification of targeted groups within social media content. To the best of current knowledge, literature described that there is no work on violence incitation identification in under-resource languages especially in Urdu. In addition, the majority of prior work on toxic content identification dealt with high-resource languages such as English. This dissertation filled out the gap and proposed a robust violence incitation detection model for Urdu. The proposed model is tested on a newly-designed Urdu corpus (collected from Twitter accounts of Pakistan).

Although several feature engineering techniques have been explored, word unigram combined with the CNN neural model has demonstrated its effectiveness. As CNN can handle the vanishing/exploding gradient problem, it demonstrated benchmark performance. The results help researchers and readers to find insights for future research and enable them to propose a practical solution. The proposed system outperformed the baselines by showing 89.84 % accuracy and 89.80 % macro f1-score and improved the accuracy by 16.41 % and macro f1-score by 16.39 %. Thus, this proposed methodology provides a robust tool to identify violence incitation content on social media. Law enforcement and security organizations can use the findings of this research to deal with unwanted content available on social media. The contributions were guided by three core research questions.

5.2.1 Research Question 1

For Research Question 1 which asked how to systematically develop a high-quality benchmark dataset for violence incitation detection in Urdu, a novel dataset was designed and curated from Twitter by carefully selecting accounts from political leaders, religious scholars, news outlets, and general users. Given the challenges of noisy user-generated text, preprocessing steps such as cleaning, normalization, handling of code-mixing, and linguistic filtering using custom-developed Urdu lexicons were implemented. This effort resulted in the first benchmark dataset for Urdu violence incitation, addressing a major research gap for low-resource language NLP.

5.2.2 Research Question 2

For Research Question 2, which examined which NLP techniques can effectively capture the lexical, semantic, and contextual cues of violent incitement, a comparative study was conducted using traditional machine learning models (Logistic Regression, Random Forest, SVM) and advanced deep learning and transformer-based approaches (CNN, LSTM, BiLSTM, CBiLSTM, Word2Vec, FastText, Urdu-BERT, Urdu-RoBERTa). Through extensive evaluation using precision, recall, F1-score, accuracy, and AUC, it was found that contextualized embeddings coupled with deep neural models achieved the best performance, significantly outperforming traditional baselines. This demonstrates the importance of semantic and contextual modeling in violence incitation detection for Urdu.

5.2.3 Research Question 3

For Research Question 3, which focused on fine-grained identification of targeted communities or groups, the detection task was extended to classify whether violent content was directed toward political, religious, governmental, or general communities. Using machine learning and deep learning classifiers, combined with careful annotation guidelines, results demonstrated that robust NLP models can successfully distinguish between these groups despite limited annotated resources. This

provides a pathway toward more interpretable and socially meaningful categorization of harmful online discourse.

Thus this dissertation is of high research and business interest to society and the proposed system is evaluated using state-of-the-art metrics. Likewise, the result of the experiments assists in uncovering significant features that are practicable for spotting violence incitation content on social media. In addition, the findings revealed worthwhile insights for the users of social media and the society by protecting their rights and promoting peace and harmony. Furthermore, the findings encourage that the proposed methodology can be used for similar social media mining tasks. This research has several implications. The proposed automated system can enable social media platforms and law enforcement organizations to quickly identify and intervene in content that incites violence. This step certainly helps prevent online conflicts from escalating into real-world violence and crime. Furthermore, the findings of this research provide insights for public organizations to design policies and regulations aimed at preventing online violence, and creating a safer digital environment for users. Also, the identification of such unwanted content highlights the need for detoxification systems in high and low-resource languages to combat such content at an early stage. Finally, the findings of this dissertation promote further categorization of violence incitation into subcategories based on race, age, gender, religion, caste, etc.

5.2.4 Broader Contributions and Impact

5.2.4.1 Dataset Contribution

The creation of the first benchmark dataset for violence incitation in Urdu.

5.2.4.2 Methodological Contribution

A comprehensive comparison of feature engineering, embeddings, and ML/DL models for violent discourse detection in a low-resource setting.

5.2.4.3 Application Contribution

Demonstration of targeted-group classification, offering actionable insights for policymakers, law enforcement, and social media platforms.

5.3 Limitations

This dissertation has several limitations; First of all, the proposed model is trained on Nastaliq style Urdu corpus that is collected from Pakistani Twitter accounts. The results of the research cannot be generalized beyond the intended scope as the size of the dataset is 4804, which is not large enough. Additionally, the source of the dataset is Twitter and a maximum of 280 characters is allowed to write a tweet. On the other hand, Facebook and other social media platforms do not impose any restrictions on the size of comments/posts. Second, we addressed the violence incitation detection as a binary classification. However, violence incitation can be seen as direct-violence, and indirect-violence and targeted groups can be identified. Another limitation is the interpretability and explainability of the proposed framework as this model is not designed with these two criteria in mind. Finally, the proposed system is only designed for the Nastaliq Urdu style and is not directly applicable to Roman Urdu in the same settings.

5.3.1 Responsible AI

Trustworthiness and responsible AI practices, such as explainability, are critical in domains like violence incitation detection. In fact, incorporating explainable AI would certainly increase user trust, system transparency, and broader adoption. However, the primary focus of my PhD work was on developing accurate detection models and designing a framework capable of identifying incitement across targeted groups such as Political, Religious, Government, and General communities. Given the complexity of building reliable datasets, training models, and validating system performance across these categories, the scope of this research was already quite challenging. While explainability techniques such as SHAP, LIME, or attention visualization could indeed provide interpretability, integrating them would have required an additional research stream, potentially diluting the focus

from the core contribution of this thesis—which is the creation of a robust detection system with strong generalizability across multiple forms of incitement. That said, I fully recognize the importance of this point. In fact, one of the key future research directions identified in my work is to incorporate explainability mechanisms so that system decisions are not only accurate but also interpretable for policymakers, regulators, and end-users. By doing this, we can bridge the gap between performance and trustworthiness, which will be essential for real-world deployment.

5.3.2 Experimental Comparison

In our literature review, we identified several key gaps that our research aims to address.

5.3.2.1 Language Mismatch

Almost all previous studies have concentrated on English, Russian, or Arabic datasets (e.g., ISIS/terrorism detection, abusive or hate speech detection). To the best of our knowledge, no large-scale annotated dataset for Urdu incitement detection has previously existed.

5.3.2.2 Domain Differences

Prior research has primarily focused on terrorism propaganda, abusive language, or offensive speech. Our work, however, specifically targets violence incitement in Urdu, with an emphasis on political, religious, governmental, and general contexts—an area that remains largely unexplored.

5.3.2.3 Methodological Gaps

Many earlier studies relied on classical machine learning approaches (SVM, Naïve Bayes, Random Forest, KNN) with features such as TF-IDF, bag-of-words, or lexicons, often reporting strong accuracy (85–97%). In contrast, our research extends beyond these methods by incorporating advanced deep learning architectures (CNN, BLSTM, CBLSTM) and modern embeddings (Word2Vec, FastText,

Urdu-BERT, Urdu-RoBERTa), none of which had been applied to Urdu incitement detection before.

Therefore, while a direct benchmark comparison is not feasible due to differences in language, datasets, and problem definitions, our thesis contributes a comprehensive literature synthesis and demonstrates how our work fills a critical research gap for Urdu violence incitement detection.

5.4 Future Work

To the best of knowledge, this dissertation is the first attempt to identify violence incitation expressions in Urdu-a low-resource language. An automatic detection model is designed by exploring semantic, word embedding, and language models with conventional ML and state-of-the-art DL algorithms. One of the contributions is the design of an annotated Urdu corpus and the data set is balanced to avoid the problem of overfitting. Experiments showed that the word uni-gram played an important role in identifying incitement to violence. Additionally, language models (Urdu-BERT and Urdu-RoBERTa) played a moderating role in the detection of violence incitation expressions. Regarding the supervised models, the 1D-CNN model outperformed the other ML and DL models as a benchmark performance. It achieved 89.84 % accuracy and 89.80 % f1-score with word unigram, improving 16.41 % accuracy and 16.39 % f1-score compared to fine-tuned Urdu-RoBERTa and the best ML model.

Furthermore, 1D-CNN played a key role in identifying individual class instances by achieving 89.76 % for the violence class, 89.84 % for the not-violence class, and 89.80 % for the macro f1-score. Although DL models provided higher scores, we still believe that DL models can significantly improve performance. For future work, current research can be extended in the direction of further categorization of violence into direct and indirect violence incitation. Another direction can be the integration of multimodal features by exploring not only textual but also visual and audio cues for violence incitation detection. Interpretability and explainability can be added to the design of an identification system for building trust and understanding for the end users. There is a need to investigate the generalization

of violence incitation detection models across other low-resource languages beyond Urdu, which will contribute to the domain of cross-lingual learning.

5.4.1 Continuation of Research

I also have interest in continuing research in this field. The current work serves as a foundational step in building resources and models for detecting incitement and violence in Urdu, and I intend to extend it further in several directions. Future plans include:

5.4.1.1 Expanding the Dataset

Collecting larger and more diverse datasets, including code-mixed Urdu-English text, social media content, and real-time streams, to improve robustness and coverage.

5.4.1.2 Model Enhancements

Exploring more advanced deep learning architectures such as transformer-based models (BERT, XLM-R, or Urdu-specific LMs) to achieve better contextual understanding.

5.4.1.3 Explainability and Responsible AI

Integrating explainable AI (XAI) methods to make predictions more interpretable and increase user trust in the system.

5.4.1.4 Cross-Lingual and Multilingual Studies

Extending the research to cover comparisons with other languages and leveraging multilingual transfer learning for improved generalization.

5.4.1.5 Real-World Deployment

Collaborating with policymakers, media platforms, and civil society to integrate the system into moderation tools for social media and digital content, thereby enhancing social impact.

In this way, I aim to evolve the work from an academic contribution into a practical, scalable, and socially impactful AI system, ensuring long-term research continuity and relevance.

Bibliography

[1] M. Whine, "Islamist organizations on the internet," *Studies in Conflict & Terrorism*, vol. 22, no. 3, pp. 231–238, 1999.

[2] M. R. Torres, J. Jordán, and N. Horsburgh, "Analysis and evolution of the global jihadist movement propaganda," *Terrorism and Political Violence*, vol. 18, no. 3, pp. 399–421, 2006.

[3] A. H. Johnston and G. M. Weiss, "Identifying sunni extremist propaganda with deep learning," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2017, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8285623

[4] M. Thompson, "Media in conflict: Inciting violence in kosovo," *Georgetown Journal of International Affairs*, vol. 8, no. 2, pp. 125–132, 2007.

[5] S. Hamid, "Satellite sectarianisation or plain old partisanship?: Inciting violence in the arab mainstream media," *The International Communication Gazette*, vol. 71, no. 4, pp. 289–310, 2009.

[6] K. Gelber, "Inciting racial violence as sedition: a problem of definition?" *Social Identities*, vol. 13, no. 1, pp. 21–36, 2007.

[7] B. Saul, "Speaking of terror: criminalising incitement to violence." *TheUNIVERSITY OF NEW SOUTH WALES LAW JOURNAL*, vol. 28, no. 3, pp. 868–886, 2005.

[8] F. Viljoen, "Inciting violence and propagating hate through the media: Rwanda and the limits of international criminal law," *Obiter*, vol. 26, no. 1, pp. 26–40, 2005.

[9] S. Khan, M. Fazil, V. K. Sejwal, M. A. Alshara, R. M. Alotaibi, A. Kamal, and A. R. Baig, "Bichat: Bilstm with deep cnn and hierarchical attention for hate speech detection," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 7, pp. 4335–4344, 2022.

[10] G. Ali and M. S. I. Malik, "Rumour identification on twitter as a function of novel textual and language-context features," *Multimedia Tools and Applications*, vol. 82, no. 5, pp. 7017–7038, 2023.

[11] M. Lakomy, "Recruitment and incitement to violence in the islamic state's online propaganda: Comparative analysis of dabiq and rumiyah," *Studies in Conflict & Terrorism*, vol. 44, no. 7, pp. 565–580, 2021.

[12] T. Akinyetun, D. Odeyemi, and J. Alausa, "Social media and electoral violence in nigeria: Sustainable development goal 16, a panacea," *KIU Interdisciplinary Journal of Humanifies and Social Sciences*, vol. 2, no. 2, pp. 169–194, 2021.

[13] A. Aly, S. Macdonald, L. Jarvis, and T. M. Chen, "Introduction to the special issue: Terrorist online propaganda and radicalization," pp. 1–9, 2017.

[14] A. Sureka and S. Agarwal, "Learning to classify hate and extremism promoting tweets," in *2014 IEEE joint intelligence and security informatics conference*. IEEE, 2014, pp. 320–320.

[15] L. Kaati, E. Omer, N. Prucha, and A. Shrestha, "Detecting multipliers of jihadism on twitter," in *2015 IEEE international conference on data mining workshop (ICDMW)*. IEEE, 2015, pp. 954–960.

[16] M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha, "Detecting jihadist messages on twitter," in *2015 European intelligence and security informatics conference*. IEEE, 2015, pp. 161–164.

[17] S. Agarwal and A. Sureka, "Using knn and svm based one-class classifier for detecting online radicalization on twitter," in *Distributed Computing and Internet Technology: 11th International Conference, ICDCIT 2015,*

Bhubaneswar, India, February 5-8, 2015. Proceedings 11. Springer, 2015, pp. 431–442.

[18] J. R. Scanlon and M. S. Gerber, "Forecasting violent extremist cyber recruitment," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 11, pp. 2461–2470, 2015.

[19] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity," in *Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part II 8.* Springer, 2016, pp. 22–39.

[20] S. D. Bhattacharjee, B. V. Balantrapu, W. Tolone, and A. Talukder, "Identifying extremism in social media with multi-view context-aware subset optimization," in *2017 IEEE international conference on big data (big data).* IEEE, 2017, pp. 3638–3647.

[21] S. Agarwal and A. Sureka, "Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website," *arXiv preprint arXiv:1701.04931*, 2017.

[22] S. A. Azizan and I. A. Aziz, "Terrorism detection based on sentiment analysis using machine learning," *Journal of Engineering and Applied Sciences*, vol. 12, no. 3, pp. 691–698, 2017.

[23] W. Sharif, S. Mumtaz, Z. Shafiq, O. Riaz, T. Ali, M. Husnain, and G. S. Choi, "An empirical approach for extreme behavior identification through tweets using machine learning," *Applied Sciences*, vol. 9, no. 18, p. 3723, 2019.

[24] M. Nouh, J. R. Nurse, and M. Goldsmith, "Understanding the radical mind: Identifying signals to detect extremist content on twitter," in *2019 IEEE International Conference on Intelligence and Security Informatics (ISI).* IEEE, 2019, pp. 98–103.

[25] A. Kaur, J. K. Saini, and D. Bansal, "Detecting radical text over online media using deep learning," *arXiv preprint arXiv:1907.12368*, 2019.

[26] H. Alvari, S. Sarkar, and P. Shakarian, "Detection of violent extremists in social media," in *2019 2nd international conference on data intelligence and security (ICDIS)*. IEEE, 2019, pp. 43–47.

[27] T. Litvinova and O. Litvinova, "Analysis and detection of a radical extremist discourse using stylometric tools," in *The 2018 International Conference on Digital Science*. Springer, 2019, pp. 30–43.

[28] A. Johnston and A. Marku, "Identifying extremism in text using deep learning," *Development and Analysis of Deep Learning Architectures*, pp. 267–289, 2020.

[29] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian, and A. Alothaim, "Exploratory data analysis and classification of a new arabic online extremism dataset," *IEEE Access*, vol. 9, pp. 161 613–161 626, 2021.

[30] M. Gaikwad, S. Ahirrao, K. Kotecha, and A. Abraham, "Multi-ideology multi-class extremism classification using deep learning techniques," *IEEE Access*, vol. 10, pp. 104 829–104 843, 2022.

[31] R. U. Mustafa, M. S. Nawaz, J. Farzund, M. I. Lali, B. Shahzad, and P. Viger, "Early detection of controversial urdu speeches from social media," *Data Sci. Pattern Recognit.*, vol. 1, no. 2, pp. 26–42, 2017.

[32] S. Kausar, B. Tahir, and M. A. Mehmood, "Prosoul: a framework to identify propaganda from online urdu content," *IEEE access*, vol. 8, pp. 186 039–186 054, 2020.

[33] N. U. Haq, M. Ullah, R. Khan, A. Ahmad, A. Almogren, B. Hayat, and B. Shafi, "Usad: an intelligent system for slang and abusive text detection in perso-arabic-scripted urdu," *Complexity*, vol. 2020, pp. 1–7, 2020.

[34] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, and A. Gelbukh, "Urdu sentiment analysis with deep learning methods," *IEEE Access*, vol. 9, pp. 97 803–97 812, 2021.

[35] M. Das, S. Banerjee, and P. Saha, "Abusive and threatening language detection in urdu using boosting based and bert based models: A comparative approach," *arXiv preprint arXiv:2111.14830*, 2021.

[36] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, and A. Gelbukh, "Threatening language detection and target identification in urdu tweets," *IEEE Access*, vol. 9, pp. 128 302–128 313, 2021.

[37] A. Mehmood, M. S. Farooq, A. Naseem, F. Rustam, M. G. Villar, C. L. Rodríguez, and I. Ashraf, "Threatening urdu language detection from tweets using machine learning," *Applied Sciences*, vol. 12, no. 20, p. 10342, 2022.

[38] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butta, H. I. Amjad, O. Vitman, and A. Gelbukh, "Overview of abusive and threatening language detection in urdu at fire 2021," *arXiv preprint arXiv:2207.06710*, 2022.

[39] S. Hussain, M. S. I. Malik, and N. Masood, "Identification of offensive language in urdu using semantic and embedding models," *PeerJ Computer Science*, vol. 8, p. e1169, 2022.

[40] M. S. I. Malik, T. Imran, and J. M. Mamdouh, "How to detect propaganda from social media? exploitation of semantic and fine-tuned language models," *PeerJ Computer Science*, vol. 9, p. e1248, 2023.

[41] D. Kang, W. Ammar, B. Dalvi, M. Van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz, "A dataset of peer reviews (peerread): Collection, insights and nlp applications," *arXiv preprint arXiv:1804.09635*, 2018.

[42] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp," *arXiv preprint arXiv:2011.00677*, 2020.

[43] C. Pérez-Almendros, L. Espinosa-Anke, and S. Schockaert, "Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities," *arXiv preprint arXiv:2011.08320*, 2020.

[44] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "Semeval-2019 task 5: Multilingual detection of hate

speech against immigrants and women in twitter," in *Proceedings of the 13th international workshop on semantic evaluation*, 2019, pp. 54–63.

[45] Z. Wang and C. Potts, "Talkdown: A corpus for condescension detection in context," *arXiv preprint arXiv:1909.11272*, 2019.

[46] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[47] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[48] R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, "Swag: A large-scale adversarial dataset for grounded commonsense inference," *arXiv preprint arXiv:1808.05326*, 2018.

[49] S. Min, V. Zhong, L. Zettlemoyer, and H. Hajishirzi, "Multi-hop reading comprehension through question decomposition and rescoring," *arXiv preprint arXiv:1906.02916*, 2019.

[50] Y. Bisk, R. Zellers, J. Gao, Y. Choi *et al.*, "Piqa: Reasoning about physical commonsense in natural language," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 7432–7439.

[51] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *Advances in neural information processing systems*, vol. 32, 2019.

[52] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *arXiv preprint cmp-lg/9602004*, 1996.

[53] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[54] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[55] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[56] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[57] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[58] M. Z. Younas, M. S. I. Malik, and D. I. Ignatov, "Automated defect identification for cell phones using language context, linguistic and smoke-word models," *Expert Systems with Applications*, vol. 227, p. 120236, 2023.

[59] M. S. I. Malik and A. Nawaz, "Sehp: stacking-based ensemble learning on novel features for review helpfulness prediction," *Knowledge and Information Systems*, vol. 66, no. 1, pp. 653–679, 2024.

[60] S. Hussain, M. S. I. Malik, and N. Masood, "Identification of offensive language in urdu using semantic and embedding models," *PeerJ Computer Science*, vol. 8, p. e1169, 2022.

[61] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," *arXiv preprint arXiv:1712.09405*, 2017.

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

[63] T. Kumar, M. Mahrishi, and G. Sharma, "Emotion recognition in hindi text using multilingual bert transformer," *Multimedia Tools and Applications*, vol. 82, no. 27, pp. 42 373–42 394, 2023.

[64] Y. Liu, S. Wei, H. Huang, Q. Lai, M. Li, and L. Guan, "Naming entity recognition of citrus pests and diseases based on the bert-bilstm-crf model," *Expert Systems with Applications*, vol. 234, p. 121103, 2023.

[65] M. S. I. Malik, A. Nazarova, M. M. Jamjoom, and D. I. Ignatov, "Multilingual hope speech detection: A robust framework using transfer learning of fine-tuning roberta model," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 8, p. 101736, 2023.

[66] M. Rehan, M. S. I. Malik, and M. M. Jamjoom, "Fine-tuning transformer models using transfer learning for multilingual threatening text identification," *IEEE Access*, 2023.

[67] N. Mali, F. Restrepo, A. Abrahams, and P. Ractham, "Implementation of mars metrics and mars charts for evaluating classifier exclusivity: The comparative uniqueness of binary classifier predictions," *Software Impacts*, vol. 12, p. 100259, 2022.

[68] G. Ali and M. S. I. Malik, "Rumour identification on twitter as a function of novel textual and language-context features," *Multimedia Tools and Applications*, vol. 82, no. 5, pp. 7017–7038, 2023.

[69] A. Nawaz and M. S. I. Malik, "Rising stars prediction in reviewer network," *Electronic Commerce Research*, vol. 22, no. 1, pp. 53–75, 2022.

[70] Z. Duan, X. Luo, and T. Zhang, "Combining transformers with cnn for multi-focus image fusion," *Expert Systems with Applications*, vol. 235, p. 121156, 2024.

[71] M. Zhu and J. Xie, "Investigation of nearby monitoring station for hourly pm2. 5 forecasting using parallel multi-input 1d-cnn-bilstm," *Expert Systems with Applications*, vol. 211, p. 118707, 2023.

[72] D. S. Fayyaz, "Impact of violent extremism on pakistani youth," *South Asian Studies*, vol. 34, no. 2, 2020.

[73] S. A. Abbas and S. H. Syed, "Sectarian terrorism in Pakistan: Causes, impact and remedies," *Journal of Policy Modeling*, vol. 43, no. 2, pp.

350–361, 2021. [Online]. Available: https://ideas.repec.org/a/eee/jpolmo/v43y2021i2p350-361.html

[74] S. N. C. for the Study of Terrorism and R. to Terrorism), "The global terrorism database (gtd)," https://www.start.umd.edu/gtd/, 2021, accessed: 2021-09-01.

[75] S. Hanif, M. H. Ali, and F. Shaheen, "Religious extremism, religiosity and sympathy toward the taliban among students across madrassas and worldly education schools in pakistan," *Terrorism and Political Violence*, vol. 33, no. 3, pp. 489–504, 2021.

[76] U. Javaid and M. I. Chawla, "Restoring peace and adopting resilient strategies for cultural tolerance: Pakistan's efforts in countering violent religious extremism (cvre) in pakistan," *PalArch's Journal of Archaeology of Egypt/-Egyptology*, vol. 18, no. 4, pp. 7052–7063, 2021.

[77] K. Kaltenthaler and W. Miller, "Ethnicity, islam, and pakistani public opinion toward the pakistani taliban," *Studies in Conflict & Terrorism*, vol. 38, no. 11, pp. 938–957, 2015.

[78] A. Delavande and B. Zafar, "Stereotypes and madrassas: experimental evidence from pakistan," *Journal of Economic Behavior & Organization*, vol. 118, pp. 247–267, 2015.

[79] B. Hassan, A. Z. Khattak, M. S. Qureshi, and N. Iqbal, "Development and validation of extremism and violence risk identification scale," *Pakistan journal of psychological research*, vol. 36, no. 1, pp. 51–70, 2021.

[80] R. Ghosh, W. A. Chan, A. Manuel, and M. Dilimulati, "Can education counter violent religious extremism?" *Canadian Foreign Policy Journal*, vol. 23, no. 2, pp. 117–133, 2017.

[81] S. Rudinac, I. Gornishka, and M. Worring, "Multimodal classification of violent online political extremism content with graph convolutional networks," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017, pp. 245–252.

[82] W. Stephens, S. Sieckelinck, and H. Boutellier, "Preventing violent extremism: A review of the literature," *Studies in Conflict & Terrorism*, vol. 44, no. 4, pp. 346–361, 2021.

[83] I. Kfir, "Sectarian violence and social group identity in pakistan," *Studies in Conflict & Terrorism*, vol. 37, no. 6, pp. 457–472, 2014.

[84] S. E. Finkel, J. F. McCauley, M. Neureiter, and C. A. Belasco, "Community violence and support for violent extremism: Evidence from the sahel," *Political Psychology*, vol. 42, no. 1, pp. 143–161, 2021.

[85] H. S. Lillah, "Religious extremism in pakistan," Ph.D. dissertation, Monterey, California: Naval Postgraduate School, 2014.

[86] S. V. R. Nasr, "The rise of sunni militancy in pakistan: The changing role of islamism and the ulama in society and politics," *Modern Asian Studies*, vol. 34, no. 1, pp. 139–180, 2000.

[87] V. R. Nasr, "International politics, domestic imperatives, and identity mobilization: Sectarianism in pakistan, 1979-1998," *Comparative Politics*, pp. 171–190, 2000.

[88] A. Riaz, "Global jihad, sectarianism and the madrassahs in pakistan," 2005.

[89] D. N. Akhter, D. M. Amin, and D. R. Naseer, "The rise of extremism in pakistan: International dynamics," *South Asian Studies*, vol. 2, no. 35, 2021.

Appendix A

Urdu Characters

Table A.1: Examples of Urdu characters

Type	Name	Symbol
Punctuation Marks	Comma	،
	Full stop	۔
	Question mark	؟
	Exclamation mark	!
	Semicolon	؛
Quotation Marks	Double quotes	” “
	Single quotes	’ ‘
Diacritics	Fatha	ـَ
	Kasra	ـِ
	Damma	ـُ
	Sukun	ـْ
	Shadda	ـّ
	Maddah	ـٓ
	Hamza	ء
Numerical Symbols	Numerals	۹ ۸ ۷ ۶ ۵ ۴ ۳ ۲ ۱ ۰

Continued on next page

Table A.1: Examples of Urdu characters (Continued)

Type	**Name**	**Symbol**
Alphabetic Characters	Tashdid	ّ
	Tanwin	ً ٍ ٌ

Appendix B

English Characters

Table B.1: Examples of English characters

Type	**Name**	**Symbol**
Punctuation Marks	Period	(.)
	Comma	(,)
	Question mark	(?)
	Exclamation mark	(!)
	Semicolon	(;)
	Colon	(:)
	Question mark	(?)
	Exclamation mark	(!)
	Quotation marks	(” ” or ‘ ’ or “ ”)
	Apostrophe	(’)
	Parentheses	(( ))
	Square brackets	([ ])
	Curly brackets	({})
	Hyphen	(-)
	Dash	(— or –)

Continued on next page

Table B.1: Examples of English characters (Continued)

Type	**Name**	**Symbol**
	Ellipsis	(…)
	Slash	(/)
	Backslash	(\)
	At symbol	(@)
	Hash	(#)
	Dollar sign	($)
	Percent	(%)
	Ampersand	(&)
	Asterisk	(*)
	Plus	(+)
	Equals	(=)
	Pipe	(|)
	Tilde	(~)
	Underscore	(_)
	Caret	(^)
	Grave accent	(‘)
Whitespace Characters	Space	( )
	Tab	(`\t`)
	Newline	(`\n`)
	Carriage return	(`\r`)
Mathematical Symbols	Plus sign	(+)
	Minus sign	(−)
	Multiplication sign	(×)
	Division sign	(÷)

Continued on next page

Table B.1: Examples of English characters (Continued)

Type	**Name**	**Symbol**
	Equals sign	(=)
	Less than	(<)
	Greater than	(>)
	Infinity	(∞)
Currency Symbols	Dollar	($)
	Euro	(€)
	Pound	(£)
	Yen	(¥)
	Cent	(¢)
Miscellaneous Symbols	Bullet	(•)
	Degree	(°)
	Section	(§)
	Paragraph	(¶)
	Trademark	(™)
	Registered	(®)
	Copyright	(©)
Emoticons and Emoji	Basic emoticons	:) :(:D :P :O ;)
	Emoji	(Emoji)
	Handles	(@username)
	Hashtags	(hashtag)
Standard numerals	Numerals	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Diacritics and Accents	Acute	(´)
	Grave	(‘)
	Circumflex	(ˆ)

Continued on next page

Table B.1: Examples of English characters (Continued)

Type	**Name**	**Symbol**
	Tilde	(˜)
	Umlaut/Diaeresis	(¨)
	Cedilla	(¸)
Control Characters	Null character	(`\0`)
	Bell/Alert	(`\a`)
	Backspace	(`\b`)
	Form feed	(`\f`)
	Vertical tab	(`\v`)

Appendix C

Arabic Characters

Table C.1: Examples of Arabic characters

Type	Name	Symbol
Letters with Special Marks	Alef With Madda Above	آ
	Alef With Hamza Above	أ
	Alef With Hamza Below	إ
	Yeh With Hamza Above	ئ
	Waw With Hamza Above	ؤ
Arabic Symbols	Sallallahou Alayhe Wasallam	ﷺ
	Alayhe Assallam	عليه السلام
	Rahmatullah Alayhe	رحمة الله عليه
	Radi Allahou Anhu	رضي الله عنه

Appendix D

Violence History in Pakistan

D.1 Violence Incitation after 9/11

In today's world, radical and quasi-religious organizations have long used propaganda as a handmaiden of violence. Terrorist movies are disseminated online through Arab and Western media channels, because to the affordability and portability of digital technology. Some of these videos are incredibly depressing in their raw, unfiltered form, full of hateful, pathologically vicious political statements and ideologies. Brigade Media Jihad produced a movie in 2005 that portrayed several attacks against coalition Iraqi forces. The video was in English and showed enemy soldiers being burnt to death and prisoners of war being executed ritualistically. Islamic music, known as nashid, adds an energizing rhythm to the play while the western soldiers are depicted as "pigs" and painted in a blood-streaked font. The criminal law has long permitted the prosecution of characters who directly incite another person to commit a particular crime, such as terrorism. Nonetheless, this is a premature and foolish overreaction that will unavoidably criminalize important contributions to public discourse: expanding the definition of incitement to include new sedition charges and the authority to outlaw organizations. Certain incitements ought to be illegal since they pose a greater threat than others. The government announced in early 2006 that a departmental evaluation of the new offences will take place, even before they were officially passed into law in late 2005, which emphasizes how hastily they were developed. One should expect a strong and developed democracy to accept disagreeable ideas without pursuing le-

gal action against them [7]. The usefulness of Section 80.2(5) as a statute against sedition and racial vilification was examined in an article. More than most States and Territories, s. 80.2(5) is essentially a sedition offence rather than a crime of racial provocation in and of itself. First, it forbids a severe but restricted kind of racial provocation that puts Australia's constitutional government in jeopardy. It was discovered through a research that there are numerous, significant problems with s. 80.2(5). Because of its "linguistic over-inclusiveness" and "great national stress," Section 80.2(5) has the potential to severely restrict and impede free speech, and it has become a tool for racial vilification in addition to serving a more restricted purpose. Second, it is possible that s. 80.2(5) and other federal legislation cover much of the same territory. Third, the fact that the fundamental component of the crime—its seditious capacity—may be committed with no mental or fault prerequisite is unacceptable in principle. Fourth, the Attorney-General should not have been the only person with the authority to file an art. 80.2(5) prosecution because of the political element of sedition. Last but not least, the mental element of this and the overall lack of textual clarity make it challenging to determine what behaviour may be included by s. 80.2(5). This examination of s. 80.2(5) supported the Senate Committee's proposal and made a compelling argument for the immediate repeal of the other new sedition laws in addition to s. 80.2(5) [6].

It is the Arab satellite news stations' fault that throughout the Arab upheavals, sectarian violence was encouraged. An analysis was done on the way that the television networks Al-Jazeera Arabic, Al-Arabiya, and Al-Mayadeen presented important incidents involving sectarian violence in Iraq and Syria. Incitement to sectarian violence had mostly been prayed through linguistic and thematic themes that built victim narratives and legitimacy claims; abusive language and the direct encouragement of violence were uncommon in a mainstream environment. According to the study's findings, Al-Jazeera mostly supports Sunni militant groups while Al-Mayadeen supports Shifa militant groups; nevertheless, none of the networks openly cited sect as an excuse for violence on air. The television network Al-Arabiya was more selective about the voices it featured and tended to rely more on government sources. Compared to the other two stations, Al-Jazeera's

interviewing practices were more ”liberal” [5]. A research was conducted by using content and comparative analysis to close a research vacuum on the propaganda techniques used in the main online publications of the Islamic State, Dabiq and Rumiyah. The primary goal of the research was to identify strategies used by editors to persuade readers to become members of the Islamic State or to engage in violent acts of jihad against non-Muslims. It was also an effort to comprehend how the publications justified the violence against its adversaries by the ”Caliphate.” While the justification of violence and calls for violence were heavily emphasized in both magazines, it was suggested that the recruitment messaging differed significantly. Dabiq gave this issue top priority by promoting hijrah, while Rumiyah was significantly less concerned in encouraging people to solicit its ranks in the Middle East. Its main objective was to mobilize the Ummah to wage jihad against the unbelievers, particularly through lone-wolf terrorist strikes [11].

Social media has transformed how people communicate, exchange their information, and engage in political processes. Even with all of its benefits, social media has shown to be a vehicle for spreading propaganda, hate speech, and false information in order to incite electoral violence. Social media's widespread use has had a notable impact on the electoral process in Nigeria, as it has in other areas of the world. Under this situation, a research was conducted wherein it was advised to pursue Sustainable Development Goal (SDG 16), which calls for the creation of inclusive, responsible, and functional institutions as well as universal access to justice. This descriptive method argued that strong institutions are essential to achieving peace, which in turn is necessary for health, vigor, and viable development. Thus, in order to address the potential issue of electoral violence in Nigeria induced by social media, this research suggested the creation of an Electoral Peace Commission, a Justice Commission, and filtering offensive language on social media platforms.

D.2 Violence Incitation in Pakistan

Multiple research articles offer comprehensive insights into the pervasive issue of extremism and violence incitation in Pakistan, shedding light on its multifaceted

dimensions and underlying causes. Fayyaz focuses on analyzing how violent extremism has psychologically, socially, and economically affected young people in Pakistan, particularly in conflict-prone areas like Khyber Pakhtunkhwa and FATA [72]. It highlights the growing mental health issues, educational disruptions, and limited employment opportunities caused by prolonged exposure to violence. The literature also emphasizes the importance of gender empowerment and inclusive youth policies in countering extremist ideologies and promoting peacebuilding efforts. Abbas and Syed explores the root causes and far-reaching consequences of sectarian violence in Pakistan [73]. It delves into the geopolitical and internal dynamics, such as foreign policy influences, radical ideologies, and governance challenges, that fuel sectarian conflict. The authors analyze the socio-political and economic toll of terrorism on national stability and propose strategic policy measures and reforms to counteract sectarianism and restore communal harmony. Pakistan has had a range of violent episodes in its history, including insurgent, political, and sectarian violence. Here's a summary of a few noteworthy instances in Table D.1 collected from Global Terrorism Database (GTD) [74].

Table D.1: Violence history of Pakistan

Date	**Incident**	**Category**
1951	Prime Minister Liaquat Ali Khan in Rawalpindi	Murder
1958	NWFP Politician Khan Abdul Jabbar Khan (Dr. Khan Sahib) in Lahore	Murder
1975	Governor of KPK, Hayat Mohammad Hayat Khan Sherpao	Bomb Blast
1981	Chaudhary Zahoor Elahi in Lahore	Murder
1982	Jang Group journalist Zahoorul Hasan Bhopali in Karachi	Murder
1991	Martial Law Administrator of KPK, Former Governor and CM, Lt. Gen. Fazl-e- Haq in Peshawar	Murder
1993	Former Punjab Chief Minister, Ghulam Haider Wyne	Murder

Continued on next page

Table D.1: Violence history of Pakistan (Continued)

Date	**Incident**	**Category**
1996	Former Premier Zulfikar Ali Bhutto's elder son, Mir Murtaza Bhutto in Karachi	Murder
1998	Former Sindh Governor, Hakim Said in Karachi	Murder
2001	Siddiq Khan Kanju, former Minister of State for Foreign Affairs	Murder
2003	MNA Maulana Azam Tariq (Chief of the Sipah-e-Sahaba Pakistan)	Murder
2003	General Musharraf had survived an elimination attempt in Rawalpindi	Bomb Blast
2003	General Musharraf had survived in 2nd elimination attempt in Rawalpindi	Bomb Blast
2004	Prime Minister-elect Shaukat Aziz had escaped unhurt in a suicide attack	Suicide Attack
2004	Balochistan Chief Minister Jam Mohammad Yousaf had escaped an assassination	Murder
2007	Punjab Minister for SocialWelfare Zil-e-Huma Usman	Murder
2007	Interior Minister Aftab Ahmad Sherpao survived in a suicide attack	Suicide Attack
2007	General Musharraf had escaped another attempt on his life when around 36 rounds fired at his aircraft	Murder
2007	Balochistan government's spokesman Raziq Bugti was shot dead at Quetta	Murder
2007	Federal Political Affairs Minister and PML-Q Provincial President Amir Muqam, had survived suicide attack	Suicide Attack
2007	A suicide bomb blast unsuccessfully targeted Interior Minister Aftab Sherpao	Suicide Attack

Continued on next page

Table D.1: Violence history of Pakistan (Continued)

Date	**Incident**	**Category**
2007	Benazir Bhutto was assassinated in a shooting and suicide bombing at Liaquat Bagh, Rawalpindi	Suicide Attack
2007	PML-Q minister Asfandyar Amirzaib (the grandson of Wali-e-Swat) was killed by a roadside bomb in Swat	Bomb Blast
2008	ANP MPAWaqar Ahmed's brother and other family members were killed by a rocket attack at his Swat	Murder
2008	A suicide attack targeted the house of ANP leader Asfandyar Wali Khan (survived) in Walibagh, Charsadda	Suicide Attack
2009	Hussain Ali Yousafi (Chairman of the Hazara Democratic Party) was shot dead at Quetta	Murder
2009	ANP provincial lawmaker Alam Zeb Khan was killed in a remote-controlled blast at Peshawar	Bomb Blast
2009	Punjabi-born Balochistan Education Minister Shafiq Ahmed Khan, was shot dead at Quetta	Murder
2009	ANP politician Shamsher Ali Khan was killed at Swat	Murder
2010	A former provincial NWFP Education Minister Ghani-ur Rehman was killed in a roadside bomb attack	Bomb Blast
2010	Another ANP leader Aurangzeb Khan was seriously injured in a Peshawar bomb blast	Bomb Blast
2010	Former Senator Habib Jalib (a nationalist leader for the Balochistan National Party) was assassinated, Quetta	Murder
2010	Mian Rashid, the son of KPK's Information Minister, Mian Iftikhar, was shot dead at Pabbi near Nowshera	Murder
2010	An MQM MPA Raza Haider was shot dead in Karachi's Nazimabad, after 6 unidentified gunmen opened fire	Murder

Continued on next page

Table D.1: Violence history of Pakistan (Continued)

Date	**Incident**	**Category**
2010	Balochistan's provincial finance minister Asim Ali Kurd, survived a suicide car bomb attack, Quetta	Suicide Attack
2011	Punjab Governor Salman Taseer was killed by his own bodyguard in Islamabad	Murder
2011	Federal Minister for Minorities Affairs, Shahbaz Bhatti was gunned down in Islamabad	Murder
2012	Khyber Pakhtunkhwa Senior Minister, Bashir Ahmad Bilour, was assassinated in a suicide attack, Peshawar	Murder
2013	Fakhrul Islam, a Muttahida Qaumi Movement leader, was shot dead by unidentified gunmen, Sindh	Murder
2013	The son and brother of Sanaullah Zehri, the provincial chief of the PML-N killed in Balochistan's Khuzdar	Murder
2015	Punjab home minister, Shuja Khanzada, was killed in a suicide blast	Suicide Attack
2018	A Balochistan politician, Nawab Siraj Raisani, was assassinated in a suicide attack in Mastung	Suicide Attack
2018	A senior ANP leader, Barrister Haroon Bilour, killed in a suicide bomb targeted an election rally, Peshawar	Suicide Attack
2018	Ikramullah Gandapur, a former provincial minister & senior leader of the PTI, killed in a suicide bomb, DIK	Suicide Attack
2018	Maulana Samiul Haq, Darul Uloom Haqqania	Murder
2021	Johar Town, Lahore, near the house of Hafiz Saeed	Bomb Blast
2021	Chaman, Baluchistan in March	Bomb Blast

Continued on next page

Table D.1: Violence history of Pakistan (Continued)

Date	**Incident**	**Category**
2021	Chaman, Baluchistan in May	Bomb Blast
2021	Quetta, Baluchistan in December, outside a college	Bomb Blast
2022	Shia mosque in Peshawar	Suicide Attack
2022	Karachi University, Killing of Chinese Academics by a Women	Suicide Attack
2022	Police Men Killed in Islamabad	Suicide Attack
2023	Suicide Bombing in Peshawar Mosque, 84 Killed, 220 Injured	Suicide Attack
2023	Karachi Police Station Attack, 4 Killed, 16 Injured	Bomb, Gun, Suicide
2023	Bolan Suicide Bombing, 9 Killed, 13 Injured	Suicide Attack
2023	North Waziristan Suicide Bombing, 3 Killed, Several Injured	Suicide Attack
2023	Jamiat Ulema-e-Islam (F) rally in Khar, Bajaur, 63+ Killed, 200+ Injured	Suicide Attack
2023	Eid Milad-ul-Nabi near a mosque, Baluchistan, 60 Killed, 70 Injured	Suicide Attack
2023	Police Station Shooting, Suicide, 23+ Killed, 34 Injured	Bomb, Gun, Suicide

The aforementioned events in tabulated form have significantly influenced Pak-

istan's politics, society, and security environment, influencing the history and advancement of the nation. Extremism, terrorism, and violence have been major problems around the globe due to a number of variables including internal strife, regional geopolitics, and religious radicalism. Pakistan saw a number of violent episodes between 1979 and 2021, and onwards, including insurgent, political, and sectarian violence. Hanif et. al. [75] examine the influence of religious education and personal religiosity on students' attitudes toward extremist groups like the Taliban. It compares perspectives from madrassa students and those in secular institutions, highlighting varying levels of sympathy and ideological alignment. The study provides valuable insights into how educational environments shape political and religious ideologies, contributing to the broader understanding of radicalization patterns among youth in Pakistan. Javaid and Chawla [76] explores the strategic initiatives undertaken by Pakistan to combat violent religious extremism. It emphasizes the importance of promoting cultural tolerance and resilience as core components in the fight against extremism. The authors analyze state policies, institutional reforms, and community-level interventions aimed at restoring peace and building societal resilience, presenting a comprehensive overview of Pakistan's counter-extremism framework.

Kaltenthaler and Miller investigates how ethnic identity and interpretations of Islam shape public opinion regarding the Pakistani Taliban [77]. The research uses survey data to assess the relationship between religious beliefs, ethnic affiliations, and levels of support or opposition to the Taliban. The authors highlight the complexity of public sentiment in Pakistan, revealing that both religious and ethnic factors significantly influence attitudes toward extremist groups, thereby offering valuable insights for counter-terrorism and policy development. Delavande and Zafar [78] explores public perceptions and stereotypes associated with madrassa students in Pakistan. Using experimental methods, the research examines how individuals from different educational backgrounds are viewed in terms of skills, behavior, and social integration. The findings reveal that madrassa students often face negative stereotypes, which may not align with their actual characteristics or performance, highlighting the impact of biased perceptions on social cohesion and educational policymaking.

Hassan et. al. [79] focuses on creating a reliable psychometric tool to assess the risk of extremism and violence among individuals. The study presents the design, development, and empirical validation of the scale within the Pakistani context, aiming to support early identification and prevention efforts. The scale incorporates psychological, social, and behavioral indicators, offering a structured approach for researchers and practitioners to evaluate susceptibility to extremist tendencies. Ghosh et. al. [80] explores the role of education as a preventive tool against violent religious extremism. It examines how inclusive, critical, and peace-oriented educational frameworks can challenge extremist ideologies and promote resilience among youth. The authors highlight case studies and theoretical insights that demonstrate how education fosters tolerance, critical thinking, and democratic values, ultimately contributing to counter-extremism strategies. Rudinac et. al. [81] explore the application of graph convolutional networks (GCNs) for the multimodal classification of violent online political extremism content. Their findings integrates both textual and visual features extracted from social media content to improve detection accuracy. By modeling content relationships through graphs, their approach effectively captures contextual connections and enhances the classification of extremist material. The research contributes to the development of intelligent systems aimed at identifying and mitigating the spread of violent political extremism online.

Stephens et. al. [82] provide a comprehensive review of literature focused on preventing violent extremism (PVE), analyzing the conceptual foundations, strategies, and effectiveness of various PVE interventions. The article critically examines existing approaches in policy and practice, highlighting the need for context-sensitive, inclusive, and community-driven methods. It also emphasizes the importance of integrating education, dialogue, and social engagement to counter radicalization processes. This review serves as a valuable resource for understanding the multifaceted nature of PVE efforts across different socio-political contexts. Kfir [83] explores the dynamics of sectarian violence and its relationship with social group identity in Pakistan. Drawing on empirical research and theoretical insights, the author examines the role of sectarian identities in fueling inter-group conflicts and violence within Pakistani society. Through a nuanced analysis of

historical, cultural, and political factors, Kfir elucidates the complex dynamics that underpin sectarian violence in Pakistan, shedding light on the motivations, grievances, and ideologies driving such conflicts. By situating sectarian violence within the broader context of social group identity formation, the article offers valuable insights into the root causes and dynamics of inter-group tensions in Pakistan, thereby contributing to understanding of the complexities of identity-based conflicts.

Finkel et. al. [84] examine the relationship between community violence and support for violent extremism in the Sahel region. Drawing on empirical data and theoretical frameworks, the authors explore the impact of endemic violence on individuals' attitudes, perceptions, and behaviors regarding violent extremism. Through a rigorous analysis of survey data collected from communities across the Sahel, the article provides empirical evidence of the linkages between exposure to violence and support for extremist ideologies and groups. Moreover, the authors investigate the role of socio-economic factors, political grievances, and social networks in shaping individuals' susceptibility to extremist propaganda and recruitment. By elucidating the complex interplay between community violence and support for violent extremism, this article contributes to understanding of the drivers and dynamics of extremism in conflict-affected regions. Lillah's dissertation [85] examines the phenomenon of religious extremism in Pakistan. Through a comprehensive analysis of historical, socio-cultural, and political factors, the author investigates the roots and manifestations of religious extremism within Pakistani society. By drawing on qualitative research methods and empirical data, Lillah offers insights into the ideologies, narratives, and networks that perpetuate extremist ideologies and fuel acts of violence in Pakistan. Moreover, the dissertation explores the role of state policies, international dynamics, and societal factors in shaping the trajectory of religious extremism in the country. Through its interdisciplinary approach and in-depth analysis, this dissertation provides a valuable contribution to the scholarly literature on religious extremism and radicalization in Pakistan.

Nasr [86] explore the rise of Sunni militancy in Pakistan and the evolving role of Islamism and the Ulama in society and politics. Through a historical analysis

of religious movements, political dynamics, and socio-economic factors, the author traces the emergence and consolidation of Sunni militant groups in Pakistan. By examining the intersections between religion, politics, and identity, Nasr sheds light on the complex dynamics that have contributed to the radicalization of Sunni communities in Pakistan. Moreover, the article analyzes the role of state policies, regional dynamics, and international influences in shaping the trajectory of Sunni militancy in the country. Through its nuanced analysis and empirical insights, this article provides valuable insights into the drivers and dynamics of Sunni militancy in Pakistan. Nasr [87] examines the dynamics of sectarianism in Pakistan and its relationship with international politics, domestic imperatives, and identity mobilization Through a comparative analysis of historical events, political processes, and social movements, the author elucidates the role of external factors and internal dynamics in fueling sectarian tensions and violence in Pakistan. By exploring the interplay between state policies, societal cleavages, and regional conflicts, Nasr provides valuable insights into the complexities of sectarian politics in Pakistan. Moreover, the article discusses the impact of identity mobilization, political competition, and economic grievances on the escalation of sectarian violence and conflict. Through its interdisciplinary approach and empirical analysis, this article contributes to understanding of the drivers and dynamics of sectarianism in Pakistan.

Riaz [88] explore the nexus between global jihad, sectarianism, and madrassahs in Pakistan. Through an analysis of historical developments, ideological trends, and institutional dynamics, the author examines the role of madrassahs in promoting extremist ideologies and sectarian violence in Pakistan. By tracing the evolution of madrassah networks and their connections to transnational jihadist movements, Riaz sheds light on the mechanisms through which extremist ideologies are disseminated and propagated within Pakistani society. Moreover, the article discusses the impact of state policies, international interventions, and societal factors on the proliferation of radical ideologies and violent extremism in Pakistan. Through its critical analysis and empirical insights, this article provides valuable insights into the complex dynamics of global jihad, sectarianism, and madrassahs in Pakistan. Akhter et. al. [89] examine the rise of extremism in Pakistan and its

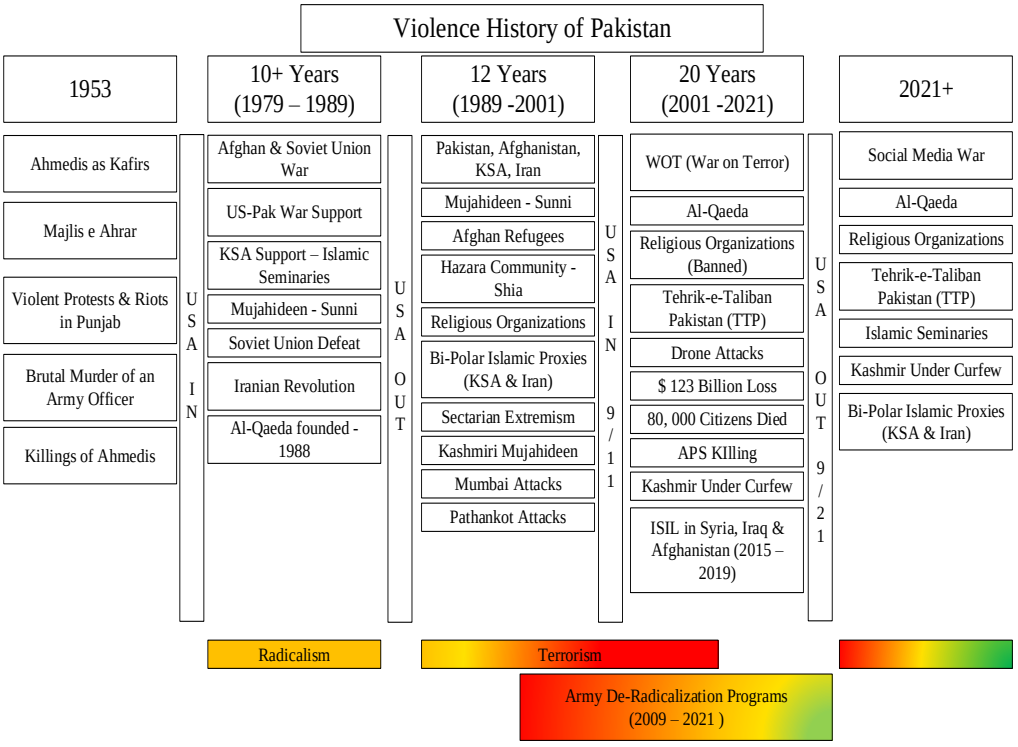


FIGURE D.1: Violence history in Pakistan

relationship with international dynamics. Through a comprehensive analysis of historical events, political developments, and socio-economic factors, the authors explore the drivers and manifestations of extremism within Pakistani society. By examining the role of external actors, regional conflicts, and global trends in shaping the trajectory of extremism in Pakistan, the article provides valuable insights into the complex dynamics of radicalization and violent extremism in the country. Moreover, the authors discuss the impact of state policies, societal grievances, and ideological factors on the proliferation of extremist ideologies and the emergence of extremist groups in Pakistan. Through its interdisciplinary approach and empirical analysis, this article contributes to understanding of the multifaceted nature of extremism and radicalization in Pakistan.

Figure D.1 illustrates key violent incidents in Pakistan from 1953 through the end of the War on Terror (WOT), as reported in the Global Terrorism Database (GTD) [74]. It also highlights the influence of the bi-polar Islamic proxy dynam-

ics between Iran and Saudi Arabia within Pakistan [73], the impact of the 9/11 attacks and the subsequent WOT on Pakistani youth [72], the ideological alignment and support of religious seminaries (madrassas) for Tehrik-e-Taliban Pakistan (TTP) [75], and the counter-radicalization initiatives led by the Pakistani Army [76]. The aforementioned challenges have significantly impacted Pakistan's security, governance, and society, underscoring the intricate and diverse characteristics of extremism, terrorism, and violence within the nation. The color scheme used in the Figure D.1 is grounded in the historical context of violence incitation and its evolving phases. Yellow represents radicalism, which gained significant momentum during the Afghan-Soviet war era. This period saw the proliferation of extremist ideologies, later transitioning into terrorism—denoted in red—following the end of the war. The spread of terrorism triggered two decades of the War on Terror (WOT), primarily centered in Afghanistan. Toward the conclusion of this prolonged conflict, efforts toward deradicalization began to take shape, marking a shift toward peace and stability. This phase is illustrated in green, symbolizing initiatives to counter extremist narratives and promote reconciliation. The color-coded representation thus encapsulates the chronological and ideological journey from radicalism to terrorism and, ultimately, toward deradicalization and peace.

Appendix E

Data Annotation Guidelines

E.1 Guidelines for Violence Incitation

Data Annotation Protocol – Violence Incitation

This data annotation protocol comprises crucial information provided to human annotators, covering categories like violence-incitation and no-violence-incitation. It includes definitions and examples to guide annotators on how to annotate a tweet.

For violence incitation tweet, tag it as follows

• Violence Incitation Tweet: If tweet contains violence incitation which to encouraging, condoning, justifying, or supporting the commission of a violent act against a group or groups (whether distinguished by race, religion, nationality, ideology or political opinion) to achieve political, ideological, religious, social, or economic goals.

• Non-Violence Incitation Tweet: If tweet does not belong to YES class, it will be classified as "No" class member, even tweet have violent words used for purpose of incident reporting or for the general purpose.

Important Notes:

• Use the comment section, if any of the given tweets aren't appropriate

• Use the comment section if additional information is needed

• Please, be careful while labeling the tweets referring to incidents only especially news channels. This isn't a violence incitation text.

Examples of Tweets & Class Labels

Examples 1:

"نواز شریف کو جوتا مارنا صحیح عمل تھا ؟ جو انکی حرکتیں اس لحاظ سے تو ٹھیک ہے یہ کیوں نکالا کا جواب ہے ، سب کچھ پتا ہونے کے باوجود دیکھئے جوتا سیاست پر عوامی رائے ۔ جوتا نہیں پھانسی دینا چاہیے اس گنجر کو اس نے ھمارے پیسوں کی ساتھ ساتھ ختم نبوت پہ بھی ڈاکہ ڈالہ ہے"

How to tag?

Is this tweet inciting for violence? Yes this part of text is inciting for violence (جوتا نہیں پھانسی دینا چاہیے)

Examples 2:

"مینار پاکستان میں ایک ٹک ٹاکر لڑکی کے ساتھ زیادتی پر سارے پاکستان والے، میڈیا والے اور حکومت نے سخت نوٹس لیا. لیکن کراچی میں ایک گھر کے 14 افراد قتل ہونے پر سب پاکستانی، میڈیا والے اور حکومت خاموش ہیں. کیوں؟ "

How to tag?

Is this tweet inciting for violence? No it is using violent words but it is not doing violence incitation.

Here, examples for violence incitation and for non-violence incitation are added in Table E.1.

Table E.1: A few samples (Yes/No-Class) for data annotators

Tweet	**Violence Incitation**
رانا ثناءاللہ بجا فرمایا ! آپ صرف انہیں محفلوں میں بیٹھتے رہے ہیں، جن میں کس کو کب، کہاں اور کیسے مارنا ہے، اس کی منصوبہ بندی ہو رہی ہو۔ ملیحہ ہاشمی کا ردعمل یار جو لعنت اللہ کی طرف سے ھے وہ تو رھے گی تم ثبت دو اور اس لعنتی کو لعنت کے ساتھ ھی پھانسی چڑھا دو جس نے بھی کیا تھا بہت اچھا کیا تھا --تم جسے نمک خرام کو سیدھی گولی مارنا چاھیے تھا	Yes

Continued on next page

Table E.1: A few samples (Yes/No-Class) for data annotators (Continued)

Tweet	**Violence Incitation**
سانحہ سیالکوٹ اور اس جیسے دیگر چاروں سانحات میں ایک بات مشترک ہے ۔ ہجوم اکٹھا کرنے کیلئے مساجد میں اعلانات کئے گئے اور دکھ بھری بات یہ ہے کہ ان اعلان کرنے والے فسادی مولویوں کو نہیں پکڑا گیا حالانکہ اصل بڑے مجرم یہ تھے ریاست کو اگر زندہ رہنا ہے تو فسادیوں کو مارنا ہو گا !	Yes
اور جو عورت جسم کی نمائش کرے تنگ پاجامہ اور شرٹ سے جسم کا انگ انگ ظاہر کرے اور لوگوں کے جذبات ابھارے تو ایسی عورت کے نمائش والے اعضا بھی کاٹ دینے چاہئیں	Yes
مریم اورنگزیب ہر حملہ اور بدتمیزی ۔ جس نے بھی بوٹ مارا ھے ۔ ھم اس مجاھد کوخراج تحسین پیش کرتے ہیں۔	Yes
اسے اور مارو۔ یہ ہے ہی اسی قابل ۔ تم بھی مارو، میں بھی اسے مارنا چاہتا ہوں ۔ ہر راہ گیر کو چاہئے، اس کو پتھر مارے ۔ سودائی، اسی سلوک کے تو مستحق ہوتے ہیں۔ یہ مڈل کلاسیا وکیل سینیٹر کیا بنا اپنی اوقات ہی بھول گیا ۔ بڑے چھوٹے کا فرق ہی کرناچھوڑ دیا ۔ نہال ہاشمی بہت بڑا احمق ہے، مجھے موقع ملا تو میں بھی نہال ہاشمی کو پتھر ماروں گا ۔ سہیل وڑائچ	Yes
نقیب قتل کیس: راؤ انوار کا نام ای سی ایل میں شامل رکھنے کا حکم ۔ راو انوار کع پھانسی دی جائے تو پاکستان زندہ باد ۔ اگر راو انوار کو پھانسی نہیں ہوئی تو مردہ آباد کا نعرہ لگے گا ۔ بھائی پہلے اسکا وکیل کو پھانسی دینا چاہیے کراچی میں مجھے شک ہے کیوں کہ اسلام آباد میں اسکو پھانسی دینا چاہیے کیونکہ پاکستان درالخلافہ ہے کراچی میں زرداری ھرامی سپورٹ کرتا ہے	Yes
میں تشدد کا قائل نہیں ہوں لیکن جب تک حسن نثار جیسے لوگوں کے منہ کالے کرکے گدھوں پر نہیں بٹھایا جائے گا یہ ملک ٹھیک نہیں ہوگا یہ بےشرم لوگ عوام کو گولیاں مارنا چاہتے ہیں ایسے بےشرم اور بےغیرت آدمیوں کو سیٹھ چینلوں پر بٹھانے کا کیا مطلب ہے؟	Yes

Continued on next page

Table E.1: A few samples (Yes/No-Class) for data annotators (Continued)

Tweet	**Violence Incitation**
عین اس وقت جب بھارت پاکستان پر حملہ کرنے کے لیے تیار ہے، ہم اپنے بدترین دشمنوں اور جہنم کے کتوں خوارج کو کیوں رہا کرنا چاہتے ہیں؟ یہ خوارج اپنے تمام کمانڈروں اور دہشتگردوں کی رہائی مانگ رہے ہیں ۔ یہ وقت تو ان کو کچلنے کا تھا، جب افغانستان میں اسلامی حکومت قائم ہو چکی ہے۔	Yes
سلام الشجاع علیک یاامیر المجاہدینؒ ۔ اب بتائیں حضور صلی اللہُ علیک وسلم کے خاکے بنا کر سرکاری عمارتوں پر لگانے والوں کو ایٹم بم نہیں مارنا چاہیے، لبیک کی حکومت آنے دیں ہم نے پہلے دن ہی کھینچ دینا ہے ۔ آنی ہے لبیک کی حکومت تم نے پریشان نہیں ہونا	Yes
بے نظیر کو جس دن پنڈی میں مارا گیا اس دن کراچی میں پنڈی واقعے کا بدلہ کچھ اس طرح سے لیا گیا جیسے کراچی والو نے اس بے نظیر کو مارا ہو جبکہ مارنا جلانا تو زرداری اینڈ کمپنی والو چاہیے تھا	Yes
مرتضی وہاب کو جوتا اتار کر گلزاری لال نندا کے منہ پر مارنا چاہیئے تھا۔ اس ملک نے زندہ رہنا ہے تو عوام کو جج جرنیل جرنلسٹ اتحاد کی نس بندی کرنی پڑے گی۔۔	Yes
اگر کوئی آپ کو iPhone 13 گفٹ کرے اور بولے کہ نواز شریف کو اسکے بدلے ایک چماٹ مارنی ہے تو کیا آپ یہ آفر قبول کریں گے؟	Yes
بچپن میں استاد محترم سے سنا تھا کہ شاگروں کو مار نا ایسا ھوتا ھیں جیسے کھیت میں فصل کو پانی دینا۔ اگر فصل کو زیادہ پانی دیاجائے تو فصل خراب۔ نہ دیا تو بھی خراب اچھا فصل وہ ھوتا ھیں جس کو وقت کے مناسبت سے پانی دیا جائے۔	Yes

Continued on next page

Table E.1: A few samples (Yes/No-Class) for data annotators (Continued)

Tweet	**Violence Incitation**
جب کوئی چور ڈاکو بن جاتا ہے ان کا استقبال پھولوں سے نہیں پتھروں سے ہو نا چاہیے تاکہ ائندہ کوئی چوری کرنے سے پہلے سو بار سوچھے کہ عدالت میں پیشی کے علاوہ عوام سے کیسے اور کہاں چپ جاوں ۔ اگر ایسا سارے عوام کا رویہ بن جائے تو مجال ہے کہ کوئی چوری کرے یا رشوت لے یا دے	Yes
زیادتی کرنے والے مجرم کا پٹواری کاٹ کر اسے بلاول بنا دیا جائے گا	Yes
یہودیوں امریکیوں اور بھارتیوں سے فنڈنگ لے کر ملک کی سلامتی کے خلاف سازشیں کرنے والوں کی گردنیں مار دینی چاہئیں ایسے سانپوں کا سر کچل دینے سے بہت سے معصوموں کی زندگیاں بچائی جا سکتی ہیں۔۔ پاکستان زندہ آباد	Yes
'میں نے منع کیا تو ڈاکٹر نے بالوں سے پکڑ کے مارنا شروع کر دیا' نرس کا FIR میں موقف ۔ اس ڈاکٹر کی بھی چھتر پریڈ ہونی چاہئیے اور اس کے بعد اس کو نوکری سے نکال دیا جائے ۔ اس پین چ کو الٹا لٹکا کر مارنا چاھے یہ ڈاکٹر ھے یہ جانور اسکے گھر میں ماں بہنیں نہیں ھے کیا ان جیسے جاہل لوگوں کی وجہ سے یہ ملک تباہ ھے	Yes
الحمد اللہ جوتا بھی سب سے پہلے نوازشریف نے کھایا تھا احسن اقبال	No
جو جوتا نواز شریف نے کھایا میں نے بھی کھایا ۔ ہم اپنے لیڈر کے نقش قدم پر چلتے رہیں گے ۔ احسن اقبال	No
مینار پاکستان میں ایک ٹک ٹاکر لڑکی کے ساتھ زیادتی پر سارے پاکستان والے، میڈیا والے اور حکومت نے سخت نوٹس لیا. لیکن کراچی میں ایک گھر کے 14 افراد قتل ہونے پر سب پاکستانی، میڈیا والے اور خکومت خاموش ہیں. کیوں؟	No
آج گودار میں قادو حجام کا مجسمہ دستی بم سے اڑایا گیا کوئی پرزہ باقی نہ رہا۔	No

Continued on next page

Table E.1: A few samples (Yes/No-Class) for data annotators (Continued)

Tweet	**Violence Incitation**
محمد علی مرزا پر گزشتہ روز جہلم میں ہفتہ وار لیکچر کے دوران حملہ آور کیا گیا۔ نہیں یہ اردو کہاں سے آئی باقی باتوں کو گولی مارو حملہ آور کیا گیا	No
جتنے لوگوں کے عضو کاٹنے ہیں کاٹ لیں لیکن یہ کیس تب ہی رکیں گے جب اللہ اور اس کے رسول کی بتائی سزائیں نافذ کروگے	No
جوتا گولی مارنا عوام کے جذبات کی عکاسی ہے ابرارالحق تو اسکا مطلب ہوا کہ شبلی فرار پر پتھراؤ ٹھیک تھا	No

E.2 Guidelines for Targeted Groups

Followings are definitions used for data annotation.

E.2.0.1 Group 1 (Government and Religious)

Group 1, identified as government/religious, encompasses entities and individuals targeted based on their association with governmental institutions or religious beliefs. This category includes government buildings, institutions, and officials, as well as symbols of statehood such as flags, passports, and constitutions. Additionally, it encompasses religious organizations and places of worship that may be targeted due to political or religious motivations. Individuals with strong religious convictions or those who express anti-religious sentiments are also included in this group, along with those who incite religious hatred through their actions or writings.

E.2.0.2 Group 2 (Political)

Group 2, categorized as political, consists of individuals or organizations targeted due to their political affiliations, views, or activities. This group includes journalists, activists, and members of political organizations, both current and former. It encompasses individuals who hold strong opinions or grievances against political

TABLE E.2: Violence incitation target dataset

Class Description	Class Label	Total
Government & Religious	1	832
Political	2	768
General	3	802

parties and may have advocated or incited violence through their communications, characterized by extreme language and viewpoints.

E.2.0.3 Group 3 (General)

Group 3, designated as the general public, encompasses instances of violence incitement targeting various segments of society. This includes women, children, members of the LGBTQ+ community facing persecution based on gender identity or sexual orientation, refugees, migrants, healthcare workers, and human rights defenders. It also covers individuals involved in criminal activities such as rape, sexual abuse, or property damage. Violence incited against people based on race, gender, nationality, caste, culture, or color falls within this category.

E.2.1 Violence Incitation Target Dataset

The dataset resulting from data annotation for violence incitement targets comprises three distinct classes. The first class, labeled as "Government & Religious" with a count of 832 instances, encompasses targets associated with governmental entities, religious institutions, and individuals known for their strong religious convictions. The second class, labeled "Political," consists of 768 instances and represents targets related to political figures, journalists, activists, and political organizations.

Lastly, the "General" class, comprising 802 instances, encompasses a diverse range of targets such as women, children, members of the LGBTQ+ community, refugees, migrants, healthcare workers, and human rights defenders, along with individuals involved in criminal activities. This dataset provides a comprehensive overview of the different groups targeted by violence incitement, allowing for in-depth analysis and understanding of the phenomenon as shown in table E.2.