ISE-Hate: A benchmark corpus for Inter-faith, Sectarian, and Ethnic hatred detection on social media in Urdu

Article · January 2023

CITATIONS
0

READS
21

3 authors, including:

Muhammad Hammad Akram
National University of Computer and Emerging Sciences
4 PUBLICATIONS **1** CITATION
SEE PROFILE

Khurram Shahzad
University of the Punjab
92 PUBLICATIONS **564** CITATIONS
SEE PROFILE

ISE-Hate: A benchmark corpus for Inter-faith, Sectarian, and Ethnic hatred detection on social media in Urdu

Muhammad Hammad Akram [2a], Khurram Shahzad [2b], Maryam Bashir[2a,*]

[a]FAST School of Computing, National University of Computer and Emerging Sciences, Lahore, Pakistan
[b]Department of Data Science, University of the Punjab, Lahore, Pakistan

Social media has become the most popular platform for free speech. This freedom of speech has given opportunities to the oppressed to raise their voice against injustices, but on the other hand, this has led to a disturbing trend of spreading hateful content of various kinds. Pakistan has been dealing with the issue of sectarian and ethnic violence for the last three decades and now due to freedom of speech, there is a growing trend of disturbing content about religion, sect, and ethnicity on social media. This necessitates the need for an automated system for the detection of controversial content on social media in Urdu which is the national language of Pakistan. The biggest hurdle that has thwarted the Urdu language processing is the scarcity of language resources, annotated datasets, and pretrained language models. In this study, we have addressed the problem of detecting Interfaith, Sectarian, and Ethnic hatred on social media in Urdu language using machine learning and deep learning techniques. In particular, we have: 1) developed and presented guidelines for annotating Urdu text with appropriate labels for two levels of classification, 2) developed a large dataset of 21,759 tweets using the developed guidelines and made it publicly available, and (3) conducted experiments to compare the performance of eight supervised machine learning and deep learning techniques, for the automated identification of hateful content. In the first step, experiments are performed for the hateful content detection as a binary classification task, and in the second step, the classification of Interfaith, Sectarian and Ethnic hatred detection is performed as a multiclass classification task. Overall, Bidirectional Encoder Representation from Transformers (BERT) proved to be the most effective technique for hateful content identification in Urdu tweets.

Keywords: `Hateful content detection, Urdu, corpus generation, sectarian, BERT, ethnic, hatred`

1. Introduction

The advent of social media and microblogging services has increased global connectivity. Twitter is a popular and influential social media that is also used by heads of state, business tycoons, celebrities, and groups of influencers. It has an active audience base that swiftly consumes and transmits the content published online [1]. Today, Twitter trends have a high influence on our society, and Twitter trends are widely recognized as a public opinion. It is well documented that social media, in particular Twitter, played a significant key role in mobilizing the protesters during the Arab revolution [2]. Furthermore, the role of Twitter in the presidential elections of various countries, including the USA and Europe, has been widely studied [3]. Recognizing the importance of Twitter, a plethora of studies have been conducted on measuring the influence of the Twitter network [4].

The arena of social media has a great sense of openness as these platforms, as well as the governments, refrain from controlling free speech on these platforms. Hence, billions of users benefit from these services

[2]Authors contributed equally

[*]Corresponding author
Email address: `maryam.bashir@nu.edu.pk` (Maryam Bashir)

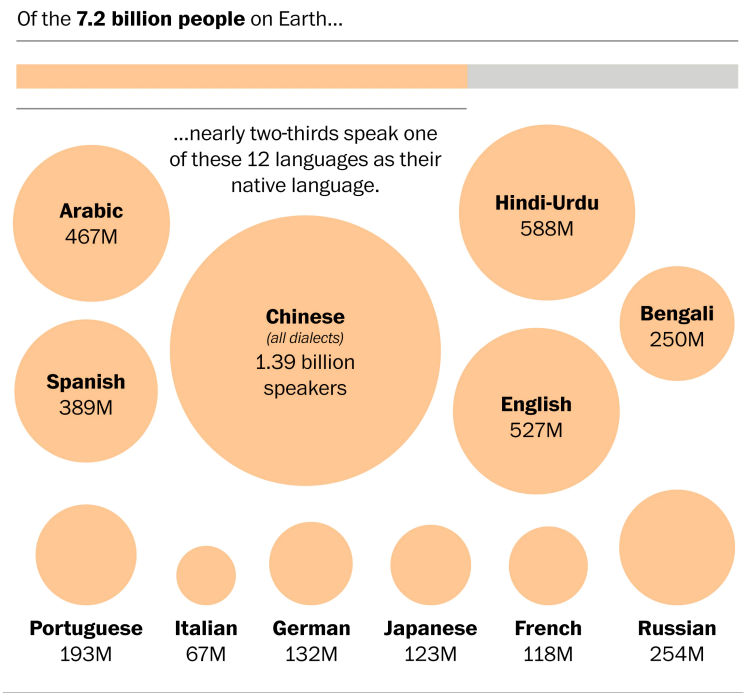

Figure 1: Sources: Ulrich Ammon University of Dusseldorf population reference bureau. Note: Total for languages including bilingual speakers. The Washington Post [15]

to freely exchange their ideas, post their opinions in large numbers, and exchange messages with each other. On one hand, the freedom of speech allows the oppressed to raise voices and the civil society to reinforce such a voice, but on the other hand, some drawbacks have also emerged. For instance, in virtual interaction, people tend to use more aggressive and hateful language as they feel physically safer [5]. Furthermore, social media has provided an opportunity for terrorist organizations to recruit workers for spreading hate and hence radicalize the society [6, 7].

The potential of the social media content to become rapidly viral and the less stringent scrutiny of the platforms poses serious threats to the stability of our society. However, due to the volume and velocity of the Twitter stream, it is cumbersome to manually verify each tweet ahead of publishing it. This necessitates the development of automated systems that can detect and eradicate hateful content from Twitter. To address this issue, a plethora of studies have been conducted on hateful content detection in Twitter [8, 9] and several Semantic Evaluation (SemEval) tasks have been dedicated to the detection of such content [10–13]. However, a large majority of the studies and tasks have focused on Western and other European languages, whereas, little attention has been paid to the detection of such content in Urdu, despite the following facts: a) Urdu is one of the 33 languages that are used in Twitter, and b) spreading hate is a legislated crime according to the Cybercrime Act of Pakistan [14].

1.1. Research Objectives & Contributions

Sectarian and ethnic violence has escalated in recent years, which has resulted in thousands of deaths in Pakistan. Believers of a particular sect began organizing in 1970s and 1980s which resulted in a directed violence against another sect. Similarly, during the last decade, people from a particular ethnicity were also targeted on many occasions in Pakistan. The issue of the spread of hateful content on social media is increasing day by day in Pakistan among various religious sects and ethnic groups. Pakistan is a Muslim majority country with a low literacy rate. Furthermore, religion is topic of discussion among masses. Some clerics and preachers exploit religious believes to spread hateful content about other sects and ethnic groups through social media to increase their followers using their native language, Urdu.

Urdu is indigenous to 170 million people in the Asian subcontinent. Figure 1 shows statistics of the most popular languages spoken in the world. It can be observed that there are more than 300 million Urdu speakers in the world [16]. Urdu is Pakistan's national language and it is also spoken by millions of people in India, USA, UK, and Canada. It follows Nastaliq writing script which is inspired by Arabic. Urdu is recognized as a low-resource language as several essential resources and accurate text processing toolkits are not available for Urdu [17]. Furthermore, to the best of our knowledge, an annotated dataset for hateful content detection and the pertinent language models are scarcely available for Urdu, which has thwarted the development of automated systems for the detection and eradication of such content. Hence, millions of Urdu speakers are at the risk of being exposed to hateful content on Twitter. To that end, the research objectives of this study are the following:

- To develop a fine-grained hateful content detection corpus on social media for Urdu - a low-resource language.
- To develop data annotation guidelines for distinguishing between hateful and neutral content and further distinguishing between interfaith, sectarian, and ethnic hatred detection.
- To evaluate and compare the effectiveness of machine learning techniques, and state-of-the-art deep learning techniques, for detecting hateful content on social media using binary and fine-grained multiclass classification.

To fulfill these objectives, this study has made the following key contributions:

- ISE-Hate corpus: This study has developed a fine-grained hateful content detection corpus (ISE-Hate), for the Urdu language. The corpus is composed of 21,759 tweets that are classified at two levels. The first level distinguishes between hateful and neutral tweets, whereas the second level further classifies hateful tweets into Interfaith, Sectarian, Ethnic, and Other hateful tweets. The first level contains 8,652 and 13,107 tweets, respectively. And the second level contains 1,537 Interfaith, 1,590 Sectarian, 167 Ethnic, and 5,421 Other tweets. As the corpus is made publicly available [1], therefore it will be useful to commence the already belated research on hateful content detection in the Urdu language.
- Generic guidelines: A systematic approach is used to develop a set of guidelines in an iterative fashion for each class of the two levels. A key benefit of using the guideline is that they can disambiguate the doubtful tweets. The effectiveness of the disambiguation guidelines is evaluated by using the Kappa score on 10% randomly selected tweets of the raw corpus. As the guidelines included in this paper are language-independent, therefore they can be used to enhance our developed corpus and develop new corpora for Urdu, as well as other languages.
- Effectiveness of machine learning techniques: Experiments are performed using four classical machine learning techniques and four state-of-the-art deep learning techniques including BERT (Bidirectional Encoder Representation from Transformers) to identify the most effective techniques for the identification of Interfaith, Sectarian and Ethnic hatred using binary and multiclass classification.

The rest of the paper is organized as follows: Section 2 provides the background and motivation for conducting a study on hateful content detection in Urdu. An overview of the existing studies is presented in Section 3. The process used for the development of our ISE-Hate corpus and the specifications of the corpus are presented in Section 4. The detailed setup of the experimentation is presented in Section 5. The results are presented and discussed in Section 6. Finally, Section 7 presents conclusion, and Section 8 presents ethical considerations.

[1] https://github.com/hammad7007/ISE_dataset

2. Background and Motivation

This section provides an overview of the Cybercrime Act of Pakistan 2016 which prohibits the use of social media for spreading controversial content. It also discusses the magnitude of Urdu language use on Twitter in Pakistan. Finally, the motivation for hateful content detection on Twitter is discussed.

2.1. Cybercrime Act of Pakistan

The Cybercrime Act of Pakistan was introduced in 2016 to make the prevision against electronic crimes [14]. Overall the act identifies twenty four types of offenses. Nine offenses are about unauthorized access to information systems, data, infrastructure, interference with these resources, and six offenses are about tempering a device, obtaining a counterfeited device, illegal use of identity, or developing a code or website with dishonest intentions. The remaining nine offenses cover generating, producing, and disseminating inappropriate or controversial content like videos, images, or text. While seven of these nine offenses are about videos and images, two types of offenses (cyberterrorism and hateful content) are of particular interest in the context of social media and microblogging services, such as Twitter. Both of these offenses discuss "preparing or disseminating information to advance Interfaith, Sectarian or Ethnic hatred", which is the focus of this study.

Twitter is a forum of public conversation and it has developed rules about the content that can be posted on the platform [18]. It uses automated approaches for the enforcement of its rules and it also offers a reporting mechanism against any violation to make certain that everyone can participate freely and safely. A recent study [19] has highlighted that during the second half of the year 2019, Twitter officially received 219 legal demands for removal of content from Pakistan and it responded with a 35.2% compliance rate. Given that Interfaith, Sectarian and Ethnic hatred is declared a crime, thus the prevention of disseminating such content in Pakistan is desirable in English, Urdu, as well as any other local language, for a promising and stable Pakistan.

2.2. Urdu language on Twitter

A comprehensive investigation of 100 million tweets revealed that English is the most prominent language of conversation using Twitter [20], i.e. 31.8% of all tweets are in English [20, 21]. Urdu is also among the 33 languages supported by the Twitter [22]. Recent data has shown that there are over 3.1 million Twitter users in Pakistan [23, 24]. As Urdu is widely spoken in the sub-continent which is a population of 1.7 billion and given that Urdu speakers can be found worldwide, the number of users who can read Urdu tweets and use Urdu for a conversation on Twitter is in large numbers. The scale of the presence of Urdu language on Twitter can be judged from the fact that a recent study [25] scarped Urdu tweets over two months to create a dataset of 1,140,824 tweets. We, therefore, contend that the presence of any hateful content on Twitter has the potential to affect a large audience.

2.3. Motivation

As discussed earlier, preparing or disseminating hateful content is recognized as a particular form of offense. Spreading or exchanging hateful content can create tension between communities, and it can also cause actual conflict or escalate an existing conflict. Pakistan has been suffering from the issue of sectarian and ethnic violence for many years and the loss of thousands of lives is attributed to such violence.

Social media has attracted more users than a physical medium, a website, or any digital media platform. Furthermore, compared to the other platforms where a single source transmits and the majority receives, in social media platforms, all the users can serve as receiver and transmitter at the same time. Also, since users feel safer due to virtual interaction they tend to be more aggressive in transmission of hateful content. Consequently, any content posted on social media can become viral in a short period of time.

There are cases where it is desired to transmit content to a large audience, such as help requests during disasters. However, the transmission of hateful content to a large audience is not desired. Furthermore, due to the limited editorial control in the presence of millions of potential transmitters, it is desired to detect and omit any hateful content that can strain relations between groups or escalate a tense situation. As discussed

earlier, Interfaith, sectarian, and ethnic hatred is recognized as cybercrime in Pakistan, the identification of such content in Urdu text is arguably more challenging than in Western languages due to several reasons. These reasons are: Urdu is widely recognized as a low-resource language with scarce computation resources. In particular, human-annotated dataset for Interfaith, Sectarian and Ethnic hatred detection in the Urdu language is not available for training supervised learning techniques. Therefore, it is necessary to develop a benchmark dataset for the detection of such hateful content on social media in Urdu language.

2.4. Disposition from the existing work

This study significantly differs from the two existing studies, [26] and [27], that appear to be related to this work. The primary difference is that both the existing studies have been conducted for the Roman Urdu text, whereas this study is conducted for the classical Nastaliq script. The two scripts are entirely different as the Roman Urdu script contains Latin alphabets, whereas the classical script is modified from the Persian script. Furthermore, it is widely acknowledged that any computation resources developed for the Roman Urdu script are not useful for the classical script. Similarly, the machine learning techniques optimized for the Roman Urdu script may not be equally effective for the Nastaliq script. Another difference is the content itself as the users who write in Nastaliq Urdu are different from the users who write in Roman Urdu. Roman Urdu datasets are not able to capture the content being created by users writing in Nastaliq Urdu.

Secondly, [27] has developed a coarse-grained dataset as it merely distinguishes between Offensive and Hate speech content. In contrast, this study has developed a fine-grained dataset in which hateful content is further classified into interfaith, sectarian, and ethnic hatred. A notable disposition from [26] is that it has randomly selected four classes, Abusive, Sexism, Religious hate and Profane, for the fine-grained classification. In contrast, the choice of our three classes, interfaith, sectarian, and ethnic hatred, for the fine-grained classification stems from the Cybercrime Act of Pakistan 2016, having a direct application.

Finally, the detailed guidelines for the corpus developed by [26] are not presented. Consequently, any extensions to the developed corpus may not be consistent with the existing annotations. In contrast, we have employed an iterative approach to develop annotation guidelines that have achieved a high inter-annotator agreement and made these guidelines publicly available. These guidelines can be used to enhance the ISE-Hate corpus, as well as to develop new corpora.

3. Related Work

A plethora of studies have been conducted on various types of inappropriate or controversial content detection on social media. It includes offensive language detection, abusive language detection, toxic content detection, and cyberbullying, etc. Although a majority of the work has been done on the detection of such content in English language, several studies have also been conducted for the detection of such content in Western and Asian languages. The presence of such a large number of studies requires a dedicated survey that can provide a comprehensive review of the existing studies. Therefore, it is appropriate to restrict the discussion of this related work section to the detection of such content in the Urdu language.

To identify the related work in Urdu language, a comprehensive search was performed using a two-phase process. In the first phase, several keywords were used in digital libraries and search engines to find an initial set of related studies. The keywords used for searching include offensive, abusive, threat, hateful content, violence, sexism, hate speech, cyberbullying, and toxic content detection in Urdu. These keywords were used to search through Google Scholar and digital libraries including, ACM Digital Library, Springerlink, IEEE Xplore, and Sciencedirect. The search was restricted to the studies published as conference or workshop papers, journal articles, technical reports, and theses, which were conducted during the last two decades. The choice of the timeline stems from the fact that social media came into existence during this time frame. In the second phase of the literature search, snowballing was used to identify the relevant studies that may have been missed in the first phase. In particular, forward and backward tracing was used to finally identify thirteen studies that are included in this paper.

Table 1 presents the summaries of the studies identified as a result of the two-phase search process. In the table, the studies are grouped by the scripts of Urdu, the classical Nastaliq script, and the Roman

Ref	Year	Script	Size	Classes	Source
[28]	2020	U	NF	NF	NF
[29]	2020	U	2171	Offensive - Non-offensive	Youtube
[21]	2020	U	6420	Abusive - Non-abusive	Twitter
[30]	2020	U	8000	Controversial - Non-controversial	Twitter
[31]	2020	U	11574	Propaganda - Non-Propaganda	Zenodo
[32]	2021	U	2171	Abusive - Non-abusive	Youtube
[33]	2021	U	3500	Abusive - Non-abusive	Twitter
[33]	2021	U	3564	Threat - Non-threat	Twitter
[34]	2021	U	5000	Toxic - Non-toxic	
[35]	2021	U	16000	Highly Offensive - Offensive - Neutral - Positive - Highly Positive	Twitter
[29]	2020	RU	10000	Offensive - Non-offensive	Youtube
[26]	2020	RU	10012	Offensive - Sexism - Religious Hate - Profane - Normal	Twitter
[27]	2021	RU	1547	Hate Speech - Offensive	Twitter
[36]	2021	RU	3000	Certain bullying - Certainly non-bulling - Indeterminate	Twitter
[27]	2021	RU	3570	Simple Hostile - Complex Hostile	Twitter
[27]	2021	RU	5000	Neutral - Hostile	Twitter
[32]	2021	RU	10000	Abusive - Non-abusive	Youtube
[37]	2021	RU	72771	Toxic - Non-toxic	Facebook, Twitter, Youtube

U = Classical Urdu script, RU = Roman Urdu script, NF = Not found

Table 1: Summary of the existing studies

Urdu (RU) script. Within each script, the studies are sorted in chronological order, and within each year the studies are sorted according to the increasing size of the corpus. The datasets whose specifications are neither available in the paper, nor could be found from the personal pages of the researchers are marked as Not Found (NF). During the synthesis of the literature, it is observed that some studies, such as [27] and [33], have developed multiple datasets and performed separate experiments on these datasets. For each dataset and each classification task performed in these studies, a separate record is added to the table. The notable observations about the identified studies are the following:

Commencement of interest in Urdu language. It can be observed from the table that a small number of studies have been conducted on the detection of controversial content on social media. This represents that this Natural Language Processing (NLP) task has received little attention from the research community. It can also be observed that all these studies have been conducted during the last two years, 2020 - 2021. This represents that the interest in the detection of controversial content for the Urdu language has commenced recently.

Comparable focus on the two Urdu scripts. It can also be observed from Table 1 that studies are conducted on both scripts of Urdu language, the classical Nastaliq script, and Roman Urdu (RU) script. RU is the script in which English alphabets are used to write Urdu text. Another notable observation from Table 1 is that a comparable number of studies (10 and 8 studies) have been conducted on the two scripts, which represents that both scripts have received equal attention of researchers.

Small datasets for the classical Urdu script. As indicated in Table 1, a majority of the datasets developed for Urdu Nastaliq script are small in size with less than 5,000 sentences. Among the remaining four datasets

merely two datasets contain greater than 10,000 sentences, and the largest dataset has 16,000 sentences [35]. In contrast, half of the RU datasets are composed of more than 10,000 sentences, and the largest dataset has 72,771 sentences [37]. One possible reason for the development of larger datasets for RU script stems from the fact that this script is widely acknowledged as a prominent language of the internet in South Asia due to the fluency of users with English keyboard [38]. This justifies the need for the development of a large dataset for the Urdu Nastaliq script which is the focus of this study.

Diverse classification. Finally, the existing studies have performed classification tasks that are different in two ways. Firstly, the studies have used different classes and labels. For instance, [30] has distinguished between controversial and non-controversial tweets, [31] has distinguished between propaganda or non-propaganda text, and [21] has distinguished between abusive and non-abusive tweets. Secondly, several studies have performed binary classification, whereas fewer studies have performed non-binary classification. For instance, [32] has classified sentences into abusive & non-abusive sentences. In contrast, [35] has used five classes, highly offensive, offensive, neutral, positive, and highly positive. Another notable observation is that all the studies have performed one-level classification except [27] in which three-level classifications are performed. From the discussion, we conclude that no study has been conducted on interfaith, sectarian, and ethnic hatred detection, which is recognized as a crime according to the cybercrime act of Pakistan.

Ref	Year	Classical techniques	Deep Learning techniques
[28]	2020	-	-
[29]	2020	DT‡, KNN, LR‡, NB, RF, SVM, LogitBoost	-
[21]	2020	-	-
[30]	2020	LR, NB, SVM	-
[31]	2020	LR	CNN
[32]	2021	IBK, NB, SVM, LR, JRip‡	Bi-LSTM, CLSTM, CNN, LSTM
[33]	2021	AdaBoost, LR, MLP, RF, SVM	CNN, LSTM
[33]	2021	AdaBoost, LR, MLP, RF, SVM	CNN, LSTM
[34]	2021	-	-
[35]	2021	NB, SVM	-
[29]	2020	DT‡, LR‡, LogitBoost, KNN, RF, SVM	-
[26]	2020	Ensemble	LSTM†, GBDT†, CNN†, Bi-LSTM† XLM-RoBERTa+CNN-gram
[27]	2021	LR, NB, RF, SVM	CNN
[36]	2021	-	-
[27]	2021	LR, NB, RF, SVM	CNN
[27]	2021	LR, NB, RF, SVM	CNN
[32]	2021	NB, IBK, SVM, LR, JRip‡	Bi-LSTM, CLSTM, CNN, LSTM
[37]	2021	LR, NB, RF, SVM	BGRU‡, Bi-LSTM, CNN‡, CNN-Tweaked

‡ represents that multiple variants of the technique are used.
DT = Decision Tree, KNN = K Nearest Neighbors, NB = Naive Bayes, LR = Logistic Regression,
RF = Random Forest, AdaBoost = Adaptive Boost, MLP = Multi Layer Preceptron,
IBK = Instance Based Learner, SVM = Support Vector Machine, CNN = Convolutional Neural Network,
LSTM = Long Short Term Memory, Bi-LSTM = Bi-Directional Long Short Term Memory,
CLSTM = Contextual Long Short Term Memory, BGRU = Bidirectional Gated Recurrent Unit,
GBDT = Gradient Boosting Decision Tree, XLM = Cross-Lingual Language Model,
RoBERTa =Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach

Table 2: Summary of the techniques for automatic classification

Table 2 presents the techniques used by the studies for the experimentation. The order of Table 2 is

consistent with Table 1 (each row in Table 2 is an extension of its corresponding row in Table 1). Similar to Table 1, the upper half of the table presents the studies on the Nastaliq script and the lower half of the table presents the RU studies. In Table 2, the studies that focus on the development of resources or techniques without performing experiments are marked with the '-' sign. For instance, [28] and [21] have proposed lexicon-based approaches for anti-social behavior and abusive text detection, respectively. However, these studies have performed experiments using their proposed techniques, i.e. these studies have neither performed experiments using classical techniques nor deep learning techniques. In contrast, the two studies, [34] and [36] have merely focused on the development of datasets and no experiments are performed. The notable observations about the techniques used for the classification task are the followings:

Growing use of deep learning techniques. It can be observed from Table 2 that a large majority of the studies have used classical machine learning techniques, whereas the use of deep learning techniques is less frequent. Furthermore, a majority of the studies that use deep learning technique are published in the year 2021, which shows that the Urdu language processing community is swiftly embracing the deep learning paradigm. Therefore, it is desirable to evaluate the effectiveness of deep learning techniques for the identification of Interfaith, Sectarian and Ethnic hatred, which is the focus of this research.

Differences in the choice of techniques for the two Urdu scripts. From the comparison of the techniques used in the Nastaliq script Urdu and RU texts we observe the following. Firstly, a majority of the studies on RU text have used classical, as well as deep learning techniques. In contrast, deep learning techniques are less frequently used for the Nastaliq script. The second observation is that the studies on RU script use the techniques that are widely used for various NLP tasks, including Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM). In contrast, the studies on the Nastaliq script use less frequently used techniques, such as Instance-Bases Learning with parameter k (IBK) and LogitBoost. A deep analysis of all the studies revealed that the studies on RU text are published in reputed journals and conferences, including Language Resource and Evaluation (LRE) journal and International Conference on Empirical Methods in Natural Language Processing (EMNLP). We contend that there is a need for conducting high-quality studies for the detection of controversial content in Nastaliq script for Urdu.

4. ISE-Hate Corpus for Urdu

An overview of the protocol used to generate our ISE-Hate is presented in Figure 2. It can be observed from the figure that the protocol is composed of four major steps, data scraping, data cleansing, guidelines development, and data annotation. The first two steps of the protocol aim to prepare Urdu text for annotation, whereas, the two subsequent steps aim to annotate the collected text. The details of these steps are given in the following.

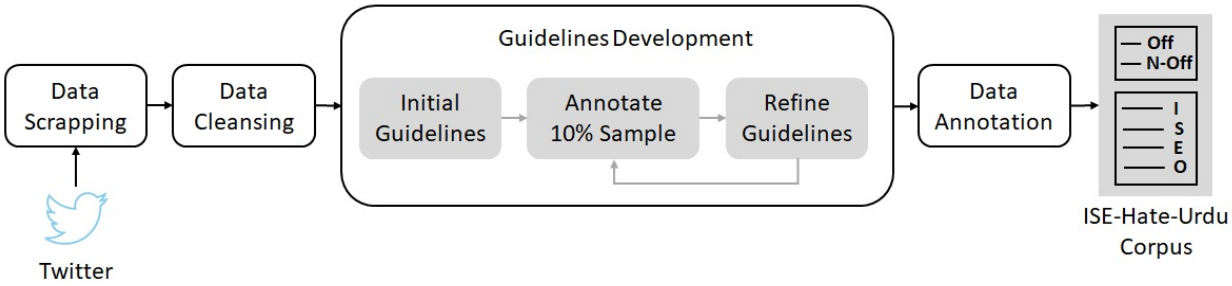

Figure 2: Overview of the protocol

4.1. Data Scraping

The first step of the protocol is to scrap Urdu in Nastaliq script in digital form. This study has used Twitter for scraping Urdu text due to the following reasons. Firstly, Twitter offers a large collection of text that can be used for research and development. In contrast, the content of the other social media networks, such as Facebook and Youtube, does not allow free data scraping. The second reason for choosing Twitter is that typically, tweets are publicly available with limited or no propriety issues associated with them. Therefore, a dataset generated from tweets can be released for use by the academic community. Thirdly, Twitter trends are very influential in our society. A notable example of the influence of Twitter can be observed from its role in the Arab Spring that changed the world. Fourthly, Twitter is a rich source of Urdu text in digital form, and a significant part of Urdu text is not available in digital form. Finally, a plethora of diverse studies have been conducted using Twitter data.

For scraping Urdu text from Twitter, a third-party python library, SNScrape [2], is used. The library allows scraping tweets using hashtags. Keeping in view the objective of the study, several hashtags were identified that potentially contain Interfaith, Sectarian, and Ethnic hate content. A subset of the hashtags used for data scraping of each class are presented in Table 3. These hashtags were chosen from [39] during March 2020 and May 2020, as they provide the top trends of hashtags in Pakistan. The reason for choosing Pakistan as a source is that Urdu is the national language of Pakistan and a large number Urdu speakers can be found in Pakistan. Hence, there is a high probability of finding Urdu tweets using Pakistan as a source. In total, Twitter Scraper was used to scrap nearly 100,000 tweets.

4.2. Data cleansing

From initial screening, it was observed that the collected data contained garbage content, such as images, timestamps, URLs, hashtags, and emojis. The presence of garbage makes it necessary to clean the text before processing it further. Therefore, the data was cleaned and pre-processed. For data cleansing, a python script was written that omits URLs, special characters, hashtags, and emojis. Furthermore, the script identifies the tweets that possibly contained code-switched text, a mixed vocabulary of Urdu and English, using a lookup-based approach. The identified code-switched tweets were manually checked and removed. In the second step, the duplicate tweets were omitted, and remaining 21,759 tweets were used for this study.

4.3. Guidelines Development

The quality of training data plays a pivotal role in the effectiveness of machine learning techniques. That is, high-quality training data can substantially improve the accuracy of machine learning techniques, whereas low-quality training data can impede the effectiveness of the technique. The last two steps of the protocol, guidelines development and data annotation, ascertain the development of a high-quality dataset. That is, the systematically developed guidelines are used for a consistent annotation, whereas, the annotation procedure, presented in the subsequent section guarantees the correct implementation of the guidelines. The details of the guideline development are discussed in the following.

Figure 3 depicts the two levels of classification that are used in this study. It can be observed from the figure that the first classification level distinguishes between hateful and neutral tweets and the second level further classifies the hateful tweets into interfaith, sectarian, and ethnic hatred, whereas the tweets that do not belong to any of the three classes are categorized as Others.

To develop the guidelines for the first-level classification, 10 % sample was randomly selected from the cleaned corpus, and two annotators were asked to independently assign a hateful or neutral label to each tweet. Furthermore, the annotators were asked to develop an initial set of guidelines based on the annotations they performed. The results of the annotations were compared, the differences in the tagged data sample were discussed and the guidelines were combined. Also, the inter-annotator agreement was computed using the Kappa statistic. The process discussed in the preceding paragraph was repeated multiple times to develop guidelines in such a way that a higher inter-annotator-agreement of 0.821 was achieved. As a result

[2] https://github.com/JustAnotherArchivist/snscrape

Nastaleeq Urdu	Roman Urdu translation	English translation
Potentially Ethnic hashtags		
#جاگ_پنجابی_جاگ	#Jaag_Punjabi_jaag	#Wake_up_Punjabi_Wake_up
#جاگ_مہاجر_جاگ	#Jaag_muhajir_jaag	#Wake_up_immigrants_Wake_up
#جاہل_سندھی	#Jahil_sindhi	#Illiterate_Sindhi
#بھٹو_زندہ_عوام_مردہ	#Bhutto_zinda_awam_murda	#Bhutto_alive_the_people_dead
#بلاول_بھکاری_مانگے_پیسہ	#Bilawal_bhikaree_paisa_mangey	#Bilawal_beggar_asks_for_money
Potentially Interfaith hashtags		
#مساجد_پر_پابندی_نامنظور	#masajid_per_pabandi_namanzoor	#Restriction_on_Mosques_disapproved
#سلام_مفتی_منیب_الرحمن	#Salam_Mufti_MuneeburRehman	#Salam_Mufti_MuneeburRehman
#آؤ_توبہ_استغفار_کریں	#Ao_toba_astaghfar_karein	#lets_repent_on_sins
#قادیانیت_سے_پاک_پاکستان	#Qadiyaniat_se_pak_Pakistan	#Pakistan_free_from_Qadiyaniat
#قادیانی_اقلیت_نہیں_غدار	#Qadiyani_aqliyat_nhn_ghaddar	#Qadiaynis_are_traitor_not_minority
#قادیانی_زندیق_لعنتی_کافر	#Qadiyani_zindeeq_laantee_kafir	#Qadiyani_zandaqa_cursed_infidel
Potentially Sectarian hashtags		
#مساجد_پر_پابندی_نامنظور	#masajid_per_pabandi_namanzoor	#Restriction_on_Mosques_disapproved
#سلام_مفتی_منیب_الرحمن	#Salam_Mufti_MuneeburRehman	#Salam_Mufti_MuneeburRehman
#زلفی_بخاری_قومی_مجرم	#Zulfi_Bukhari_Qaumi_mugrim	#Zulfi_Bukhari_national_culprit
#تبلیغی_امت_کے_محسن	#Tablighee_umaat_ke_mohsin	#Tablighee_nation's_mohsin
#گستاخ_معاوی_آزاد_کیوں	#Gustakh_Muawiya_azad_kiyon	#Disrespectful_Muawiya_Why_Free
#گستاخوں_کو_پھانسی_دو	#gustakhon_ko_phansi_do	#Hang_the_disrespectful
Potentially Others hashtags		
#شہداء_میلادالنبی_نشتر_پارک	#Suhdaemillad_ulnabi_nishtar_park	#Martyr_of_milad_nabi_Nishtar_park
#بھٹو_زندہ_عوام_مردہ	#Bhutto_zinda_awam_murda	#Bhutto_alive_the_people_dead
#بلاول_بھکاری_مانگے_پیسہ	#Bilawal_bhikaree_paisa_mangey	#Bilawal_beggar_asks_for_money
#بلاول_زرداری_شرم_کرو	#Bilawal_zardari_sharam_kero	#Shame_on_you_Bilawal_Zardari

Table 3: Examples of hashtags of potentially interfaith, sectarian, ethnic, and others

of the iterative process, guidelines were developed for the first level of annotation. These guidelines are presented in Table 4.

For the second level annotation, 10 % of the hateful tweets were randomly selected, and two independent annotators were asked to independently assign one of the four labels to each hateful tweet, interfaith, sectarian, ethnic hatred, or others. Similar to the first-level annotation, the iterative process was repeated multiple times to mature the guidelines for the second level annotation and to achieve the inter-annotator agreement of 0.82. This higher value of the Kappa score represents that the use of the guidelines can lead to consistent labeling of tweets. The guidelines generated for the second level of classification are presented in Table 5 - Table 8.

We contend that the development of guidelines is a significant contribution as they can have manifold benefits. Firstly, the guidelines ascertain a consistent annotation of the corpus, hence, improving the effectiveness of the machine learning technique. Secondly, the guidelines can be used by other researchers to enhance the size of the corpus without compromising the quality of the annotation. Finally, the guidelines are not language-specific, therefore, they can be used to develop similar corpora for Western, as well as Asian languages.

4.4. Data Annotation

In addition to the annotation guidelines, the annotation process also plays a pivotal role in generating a quality dataset. The annotation process defines the systematic steps that should be used for manually

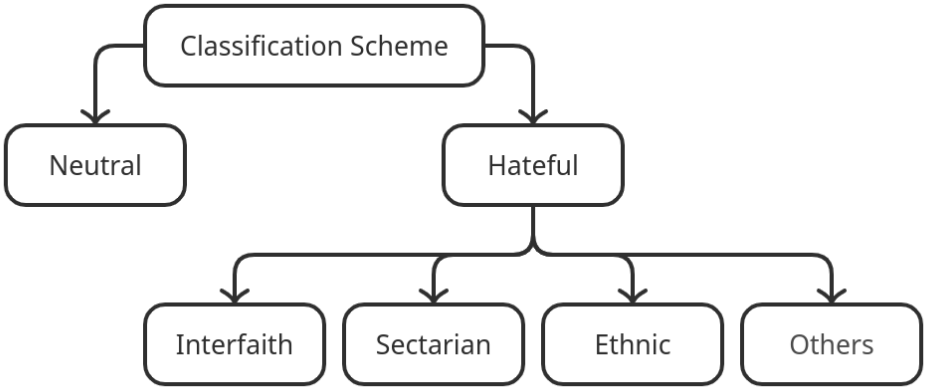

Figure 3: Classification of Urdu tweets

Guidelines for Hateful
1. Associating humans with animals. 2. Blaming someone for being a coward or liar. 3. Questioning or criticizing practicing of religion. 4. Criticizing media for following a specific agenda. 5. Blaming some religion or belief for unethical acts. 6. Blaming a country or city administration for any act. 7. A tweet that announces a decision in offensive terms. 8. Criticism on religious books, clerics, or religious places. 9. Unsolicited or inappropriate information like offensive/threatening messages. 10. An expression that provokes someone to violate the law of the land or government policy. 11. Blaming the government for participating in a global conspiracy or acting as an agent of others. 12. The local context should be taken into account for deciding whether a tweet is offensive or not. 13. Presenting a negative notion of government or religious scholars due to a decision that involves limiting religious activities.
Guidelines for Neutral tweets
1. Poetry or lyrics. 2. Misquote of religious scriptures. 3. The conversational statements. 4. Demand to arrest an individual without being abusive. 5. Any lightweight criticism on a decision using appropriate words. 6. Blaming natural resources of a specific region for causing trouble. 7. The quotation of religious scriptures or quotation of a natural or religious leaders.

Table 4: Guidelines for Hateful and Neutral tweets

assigning appropriate labels to each tweet in the corpus. That is, a complex and multi-facet annotation process is more error-pruned than a simple iterative process. The details of the annotation process used in this study are discussed in the following.

- As a starting point, two annotators who are native Urdu speaker independently examined all 21,759 tweets manually and assigned appropriate labels of hateful or neutral to each tweet. Subsequently, the decisions of both annotators were compared and the conflicts were identified. Each conflict was discussed by the two annotators to develop consensus in the light of the respective guidelines. However,

Guidelines for Interfaith
1. A tweet is annotated as Interfaith only if the target group faith is clear. 2. Sarcasm on religion or faith. 3. Attack on religion is interfaith. 4. Relating religious group with an animal. 5. Blaming a faith for actions against other communities. 6. Curse on a group of people with or without any religion. 7. Expression or demand of refusing fundamental rights to a religious group. 8. Expression of blaming a religious group for being not loyal to the state. 9. Comparing individuals of one religion for their acts with other religious acts. 10. Relating the target to an individual or a group which is widely pronounced as against a religion. 11. If the expression is in the second or third person without specifying who they are it is not interfaith. 12. Any comment that is about hurting, threatening a specific faith, or provoking someone to take some action. 13. Associating something that has a negative notion in the general public, with some religion will be interfaith. 14. Blaming individuals of a religious belief for facilitating one group, being sympathetic to a group or depriving another group. 15. An offensive expression about religious beliefs of an individual or a group of people, or about their habits of practicing religion. 16. A hateful expression about a religious person, preachers, or its activities that are not closely bonded with a religion.

Table 5: Guidelines for Interfaith

Guidelines for Sectarian
1. A tweet is annotated as Sectarian only if the target sect is identifiable. 2. Sarcasm on a sect. 3. Questioning believes of a sectarian group. 4. Blaming a sect for collusion with enemies. 5. An expression that promotes sectarian feelings. 6. Relating an individual's offensive act with a sect. 7. Encouraging others or provoking someone for violence against a sect. 8. Being decisive about the believes of a sect or blaming a sect for blasphemy. 9. An abusive or offensive comment about a religious act that is associated with a specific sect. 10. Blaming the leaders that are strongly associated with a sect for some act against state or community. 11. A specific characteristic or a specific practice of a sect is mentioned that helps identification of sect.

Table 6: Guidelines for Sectarian

in some cases a third annotator was also involved and the decisions were taken by the majority vote.

- The tweets that were labeled as hateful were independently reviewed by the two annotators to determine whether each tweet is interfaith, sectarian, ethnic, or other. It is important to note that the annotators used multiple rounds of annotations for the application of the three sets of guidelines. In the first round, the annotators reviewed each tweet and applied Interfaith guidelines to determine if the tweet is an Interfaith tweet or not. Subsequently, in the second round, the annotator reviewed the tweets to ascertain if each tweet is Sectarian or not following the sectarian guidelines. In the

Guidelines for Ethnic
1. A tweet is annotated as Ethnic only if the target ethnicity is identifiable. 2. Blame of a violence on an ethnic group. 3. Hateful or abusive comment about an ethnicity. 4. Being decisive about the actions of an ethnic group. 5. Associating an act of an individual with an ethnic group. 6. Blaming ethnic groups or their leaders to use their position for benefits. 7. Use of a nickname for an ethnicity with which negative sentiments are associated. 8. The tweet in which the ethnicity is associated with extremely negative expression. 9. A negative expression about certain activity that is associated with a specific ethnic group.

Table 7: Guidelines for Ethnic

Guidelines for Other hateful tweets
1. A hateful tweet with unclear target faith, sect or ethnicity. 2. The absence of a target community is the primary criteria for this class. 3. Insulting remarks towards an individual or a group of people or a specific company or government department, without commenting on their belief system, sect or ethnicity. 4. A hateful expression pointing towards a community or group of people based on any characteristic other than faith, sect or ethnicity. 5. All hateful comments that do not fulfil the criteria of the three classes, Interfaith, Sectarian or Ethnicity.

Table 8: Guidelines for Other hateful tweets

final round, the annotator applied Ethnic guidelines to ascertain whether each tweet is ethnic or not and labeled the non-ethnic tweets as Others. Similar to the first level, the decisions of annotators were compared and conflicts were identified. Finally, each conflict was discussed to develop consensus, whereas in the remaining cases the third annotator was involved and the labels were assigned by the majority vote.

4.5. Specification of ISE-Hate Corpus

The specifications of our developed ISE-Hate corpus are presented in Table 9. The corpus is composed of 21,759 tweets that are annotated at two levels. The first level annotation is composed of 8,589 hateful tweets and 13,107 neutral tweets. For the second level annotation, the hateful tweets are labeled with the three classes, i.e. the 8,689 Hateful tweets are further classified into 1,537 interfaith, 1,590 sectarian, 167 ethnic tweets, and 5,421 others. However, it was observed that 63 tweets belonged to two classes. These 63 tweets are duplicated, one class name was assigned to first copy whereas the second class name was assigned to second copy. Out of these 63 Hateful tweets, 54 tweets are both interfaith and sectarian, 8 tweets are ethnic and sectarian, whereas one tweet is interfaith as well as ethnic.

Category	Total Tweets
Hateful	8,652
Neutral	13,107
Total	21,759

Table 9: Distribution of tweets at the first classification level

4.6. Example tweets from each category

Some examples of tweets are shown in Figure 11. It can be observed from the examples that in the Ethnic tweets, hate is targeted towards specific groups, for example Sindhi (Sindhi represents residents of a

Category	Total Tweets
Interfaith	1,537
Sectarian	1,590
Ethnic	167
Other	5,421
Unique tweets	8,589
Grand Total	8,652

Table 10: Distribution of tweets at the second level classification

province of Pakistan called Sindh). In the Sectarian tweets, hate is targeted towards a specific sect, such as Shia and Bralvi. In contrast, the tweets in the Others group are those that express hate but do not lie in a specific category of Interfaith, Ethnic, or Sectarian.

5. Experimentation

This section presents the details of the experiments that are performed to evaluate and compare the performance of classical machine learning and deep learning techniques. Figure 4 presents the workflow of a hateful content detection approach that is used in this study. The details of each step are given in the following.

5.1. Pre-processing

People use informal language in form of unstructured text for expressing their opinions, and feelings on Twitter. Therefore, it is necessary to apply some pre-processing to the corpus before employing any technique for learning and prediction. The pre-processing is used to omit the garbage and non-substantive content from the raw text. Figure 5 shows pre-processing steps of our system. The steps involved in pre-processing of our dataset are listed below:

- Data cleaning: This step includes the removal of URLs (http:// or https://), and the removal of words that do not belong to Urdu language. The tags for persons are also removed in order to anonymize for ethical considerations.
- Removal of Punctuation: The punctuations like commas, question marks, etc., are removed as they do not give any information about the topic of the tweet.
- Tokenization: Tokenization is the process of breaking the string into separate sentences and words. Each language has a different tokenizer according to its script. For example, in English, the sentence ”NLP is fun and exciting” will be tokenized into ”NLP”, ”is”, ”fun”, ”and”, ”exciting”.

5.2. Features

The effectiveness of supervised learning techniques is dependent on the features given as input to them. This study has used two types of features, the classical bag of words model and state-of-the-art representations called word embeddings.

The bag of words model is a common method for feature representation in machine learning models. In this representation, we only keep the information about the count of the words and lose the information about word sequence and their relative position. This is a very simple way of representing text features and gives a comparable result to the more complex feature representation methods. We have used the unigram model. Unigram features are simple unigrams (single words) collected from the training set.

Word embedding is a dense vector representation method used for the representation of words for deep learning models. The first layer of the deep learning model is the embedding layer, which learns a representation of words from the training data. Each word is represented by its neighboring words. If two

Interfaith		
1	Nastaliq Urdu	قادیانی کل بھی کافر تھے آج بھی کافر ہیں کل بھی کافر ہی مریں گے
	RU Translation	Qadiyani kal bhi kafir they, aj bhi kafir hein, kal bhi kafir hi marein ge
	EN Translation	Qadianis were infidels yesterday, they are infidels today, they will die as infidels tomorrow
2	Nastaliq Urdu	وئیرس آ یاسیلاب یاپھرطوفان لیکن شیعہ کل بھی کافرتھاشیعہ آج بھی کافر ہے اور وئیرس ختم ہونےکےبعد بھی کافر رہےگا انی دیو لاک ڈاؤن میں حالات بدلتے ہیں عقیدے نہیں
	RU Translation	Virus aye ya saillab ya toofan lakin shia kal bhi kafir tha shia aj bhi kafir ha aur virus khatam honey ke baad bhi kafir rahey ga. Anee deo lock down me hallat badaltey hein aqeedey nhn
	EN Translation	Whether there is virus, floods, or hurricanes, Shia was infidel yesterday, Shia is infidel today, and he will stay infidel after the virus ends. Lockdown changes circumstances not beliefs.
Ethnic		
1	Nastaliq Urdu	جنوبی پنجاب کھپے مگر کالاباغ ڈیم نہ کھپے. سندھی ڈاکو کا دوغلاپن
	RU Translation	Janoobi Punjab khapey magar kalabagh dam na khapey. Sindhi dakoo ka doghlapun
	EN Translation	South Punjab is accepted but Kalabagh Dam is not accepted. Hypocrisy of Sindhi dacoit
2	Nastaliq Urdu	کتے کے بچو 2002 کا ایکسپائرڈ راشن دینے سے تو بہتر تھا سندھیوں کو زہر دے دیتے، کم از کم کورونا اور بھٹو مافیا حکومت کے دہرے عذاب سے تو جان چھوٹتی انکی
	RU Translation	Kuttey ke bacho, 2002 ka expired rashan deney se to behter tha sindhiyon ko zehar de dettey. Kum az kum corona aur Bhutto ke dohrey azab se to jaan chootatee inki
	EN Translation	Sons of d*g (B*stards), It would have been better to poison the Sindhis than to give the expired rations of 2002, at least they could escape from the double torment of Corona and Bhutto mafia government.
Sectarian		
1	Nastaliq Urdu	شیعہ غلیظ کافر ہے
	RU Translation	Shia ghaleez kfir ha
	EN Translation	Shias are filthy infidels
2	Nastaliq Urdu	وہ پوچھنا تھا نیشنل ایکشن پلان کہاں ہے یا صرف سنیوں کے خلاف بنایا گیا ہے
	RU Translation	Wo poochna tha national action plan kahan ha ya sirf sunniyon ke khilaf banaya gia ha
	EN Translation	I want to ask, where is the National Action Plan or is it made only against Sunnis
3	Nastaliq Urdu	اگر ایران کے دو ہزار شیعہ زائرین روک لیتے تو مساجد کو زبردستی نہ بند کرنا پڑتا
	RU Translation	Agar iran ke do hazar shia zaereen rok lettey to masajid ko zabardasti band na kerna perrta
	EN Translation	If 2,000 Shiite pilgrims from Iran were stopped, mosques would not have to be closed forcibly
Other		
1	Nastaliq Urdu	عمران خان 4 کروڑ 02 لاکھ اس لئیے بھونکنے والے کتوں پر خرچ کرے
	RU Translation	Imran khan 4 corore bees lakhiss liye bhonkney waley kutton pe kharch kerey
	EN Translation	Imran Khan should spend Rs 42 million on barking d*gs
2	Nastaliq Urdu	کیا نیازی کے باپ کا پیسہ ہے بے شرم انسان کس چکر میں قومی خزانہ لٹا رہا ہے
	RU Translation	Kia niazi ke baap ka paisa ha be sharam insan kis chakar me qaumi khazana lutta raha ha?
	EN Translation	Is it Niazi's father's money? For what purpose, is the shameless man squandering the national treasury?

Table 11: Example tweets for each class

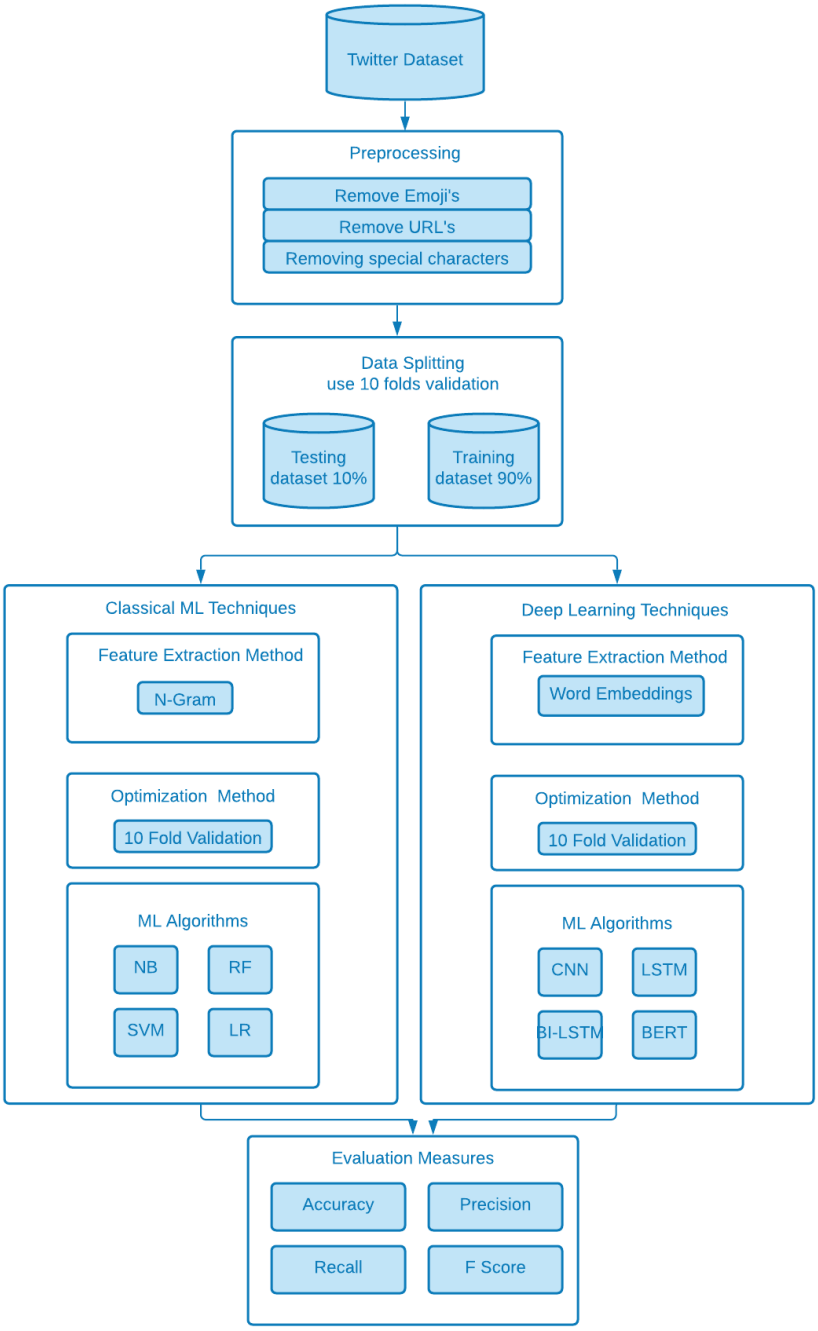

Figure 4: High level system architecture

words appear in similar contexts then their dense vector representation is very similar. Word embeddings can capture the semantics of words based on their environment. All of our deep learning models use the embedding layer for the representation of words.

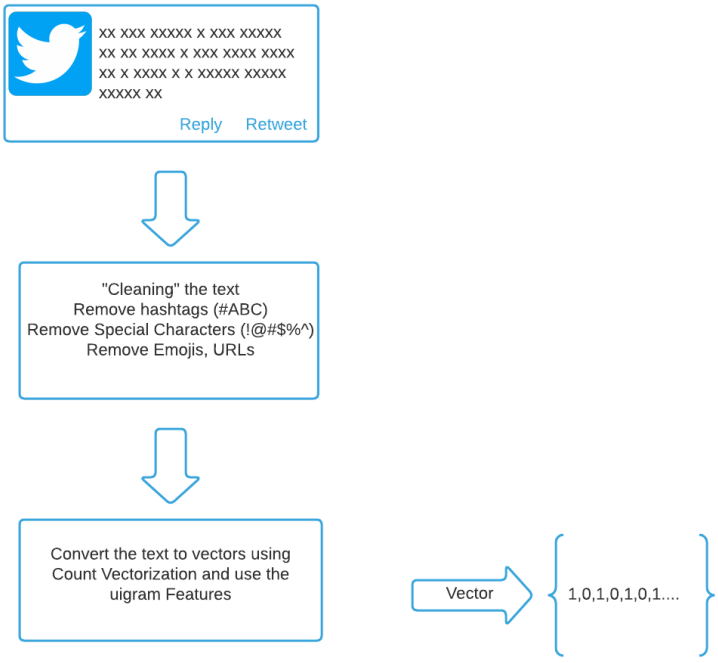

Figure 5: Overview of the pre-processing of Tweets

5.3. Classification techniques

In this section, we present the traditional supervised machine learning techniques and state-of-the-art deep learning techniques for the detection of interfaith, sectarian, and ethnic hatred from ISE-corpus. The machine learning algorithms used in this study are Support Vector Machine (SVM) [40], Random Forest [41], Naïve Bayes [42], and Logistic Regression [43]. The deep learning techniques used in our experiments include Convolution Neural Network (CNN) [44], Long Short Term Memory (LSTM) [45], Bidirectional LSTM (Bi-LSTM) [46, 47], and Bidirectional Encoder Representations from Transformers (BERT) [48].

Sequence-based deep learning models have shown state-of-the-art performance for text classification techniques in past literature. LSTM and Bi-LSTM have been popular sequence-based deep learning models for text classification. LSTM [45] is a type of Recurrent Neural Network (RNN) that lie in the category of multi-layer deep neural networks with connections between different hidden states. The Bidirectional LSTM (Bi-LSTM) [46, 47] is a type of LSTM that uses information from both the past and future. Words have a dependency not only on previous words but also on the next words so Bi-LSTM uses information from both directions to represent the meaning of a word. The connections between hidden states are in both directions. Bidirectional Encoder Representations from Transformers (BERT) [48] is a state-of-the-art model for text classification proposed in 2018. It outperformed LSTM and BiLSTM on many tasks with sequence-based inputs [48]. It replaced the sequence-based architecture of RNN and LSTM with the feed-forward neural network using a multi-head attention mechanism. In this study, we have trained BERT Base Multilingual Uncased model for the ISE-corpus and used it for the classification of tweets. For LSTM, BiLTM and CMM, the learning rate of Adam Optimizer (learning rate = 0.001) is used, whereas for the BERT the learning rate of 0.00001 and CLS token is used and for the BERT. Figure 6 shows the architecture of one encoder state of BERT. BERT processes sequence-based information by using a multi-head attention mechanism. The self-attention layer takes the dot product of the input word with all query vectors of all other words in the sequence. The final output is the vector representation of input words such that the representation of each word has information about surrounding words baked into it.

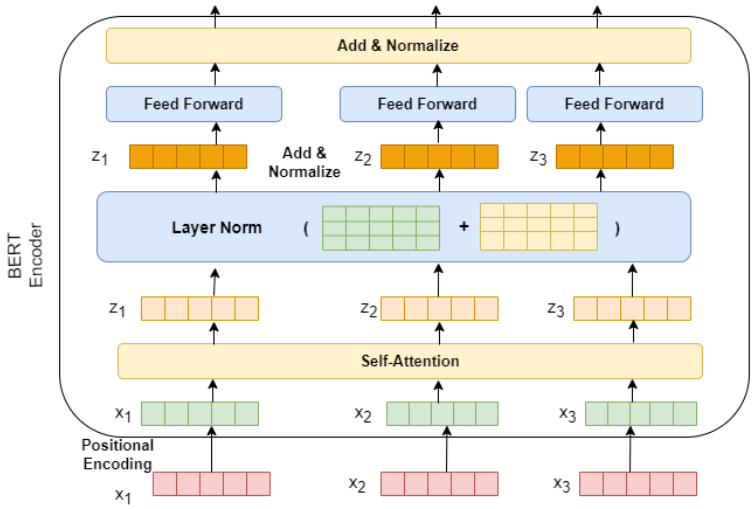

Figure 6: Architecture of one encoder of BERT

5.4. Evaluation measures

We have used benchmark evaluation measures for text classification in this study for the evaluation of our classification models. These measures are Precision, Recall, and F1 score. Where, Precision also known as positive predictive value is the fraction of relevant instances among the retrieved instances. Recall is the fraction of the total number of relevant instances that were retrieved. Recall deals with finding relevant results from the corpus. And, F1 score takes both precision and recall into account. It is the harmonic mean of Precision and Recall. Both of these values should be high for F1 score to be higher.

		CNN			LSTM			BiLSTM			BERT
	Layers →	1	2	3	1	2	3	1	2	3	
L1	Max features	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	
	Max length	400	400	400							140
	Embedding dimension	50	50	50	128	128	128	128	128	128	
	Dropout	20%	40%	20%	20%	40%	20%	20%	20%	20%	30%
	Filters	250	250	250							
	Kernel Size	3	3	3							
	Hidden Dim	250	250	250							
	Epochs	2	2	3	2	2	3	2	2	3	3
L2	Max features	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	
	Max length	400	400	400							140
	Embedding dimension	50	50	50	128	128	128	128	128	128	
	Dropout		40%	20%	20%	40%	40%	20%	20%	20%	30%
	Filters	250	250	250							
	Kernel Size	3	3	3							
	Hidden Dim	256	250	250							
	Epochs	10	2	3	10	2	3	10	2	3	3

L1 = Level 1, L2 = Level 2

Table 12: Hyper-parameters settings for deep learning models

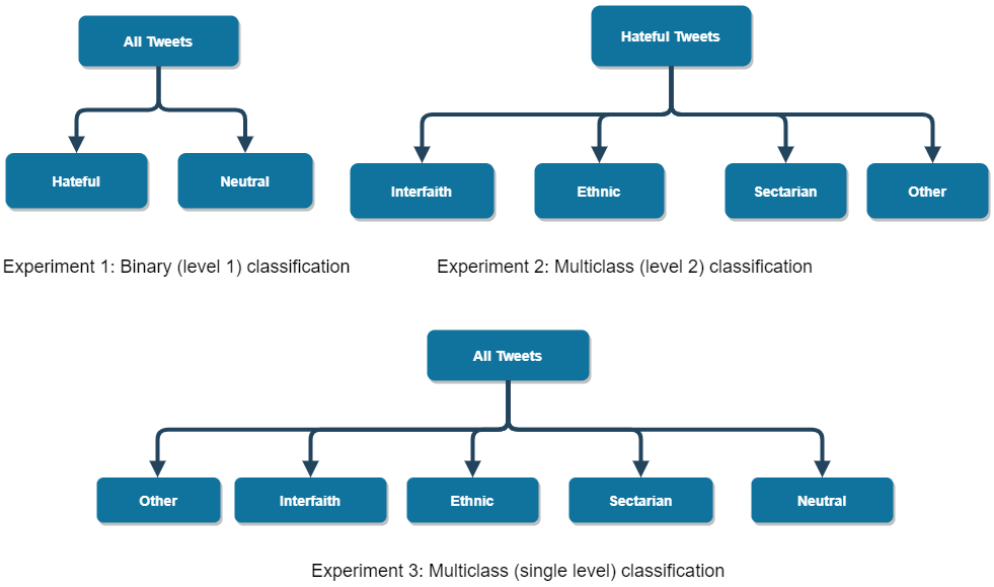

Figure 7: Three sets of classification experiments

5.5. Experimental setup

The ISE-Hate corpus contains two levels of classification, including a binary and a multi-class classification, so the experiments were also performed for binary and multiclass classification. Three sets of experiments were performed. In the first set of experiments, the tweets were classified as hateful or neutral using binary (level 1) classification. In the second set of experiments, fine-grained multiclass (level 2) classification is performed on hateful tweets. Multiclass classification is done to classify the hateful tweets into one of the four categories, interfaith, sectarian, ethnic, or other. The third set of experiments were performed where all tweets are classified in a single-level into one of the five categories: neutral, ethnic, sectarian, interfaith, or other. Figure 7 shows three sets of classification experiments.

K-fold cross-validation was used in all sets of experiments because it is widely acknowledged as a standard way of optimizing hyper-parameters and generating reliable results. According to this approach, the data was divided into k partitions, where k -1 partitions were used for training and one partition was used for testing. The experiments were repeated k times using each partition for testing and the remaining partitions for training. In this study, each experiment used 10-fold cross-validation with $k = 10$, which means that for each fold 90% of the corpus was used for training and 10% was used for testing.

Experiments were performed using the four classical machine learning techniques, NB, RF, LR, and SVM, and four deep learning techniques, CNN, LSTM, Bi-LSTM, and BERT. For CNN, LSTM, and Bi-LSTM experiments are performed using one, two, and, three layers to evaluate their effectiveness. The parameters tuned for deep learning techniques are presented in Table 12. Batch size for all models is 32. For the binary classification binary the cross entropy loss function and sigmoid activation function is used. In contrast, for the for the non-binary classification the categorical cross entropy and softmax activation function is used. BERT large with 12 encoders 12 attention heads was used in this study. An overview of all the techniques is presented in Section 5.3.

6. Results and Discussion

In this section, the results of the three sets of experiments are discussed.

Type	Technique	Layers	Binary (First-level)			Multiclass (Second-level)		
			Precision	Recall	F1 Score	Precision	Recall	F1 Score
Classical	NB		0.607	0.589	0.531	0.458	0.468	0.456
	RF		0.833	0.781	0.792	0.842	0.721	0.752
	LR		0.828	0.810	0.816	0.842	0.740	0.774
	SVM		0.814	0.806	0.809	0.833	0.770	0.794
Deep learning	CNN	1	0.827	0.823	0.825	0.789	0.788	0.784
		2	0.825	0.822	0.823	0.680	0.684	0.681
		3	0.805	0.793	0.798	0.661	0.669	0.664
	LSTM	1	0.824	0.817	0.819	0.771	0.723	0.737
		2	0.823	0.812	0.816	0.663	0.669	0.665
		3	0.788	0.798	0.792	0.656	0.667	0.660
	Bi-LSTM	1	0.824	0.815	0.818	0.786	0.750	0.762
		2	0.776	0.735	0.724	0.656	0.661	0.657
		3	0.799	0.747	0.748	0.619	0.610	0.606
	BERT		0.839	0.831	0.834	0.831	0.846	0.832

Table 13: Results of the experiments for the two level classification

6.1. Binary (Level 1) classification

Table 13 presents the macro average results of the experiments that are performed for the binary classification. This set of experiments takes a Tweet as input and classifies it as either neutral or hateful. It can be observed from the results that NB is the worst performing technique as it merely achieved an F1 score of 0.531. In contrast, BERT is the most effective techniques as it achieved an F1 score of 0.834. The performance of Random Forest, Logistic regression and SVM is also comparable with deep learning models. By comparing the results of the classical and deep learning techniques it is observed that all the deep learning techniques achieve a higher F1 score than the classical techniques in at least one setting. However, the difference in the performance is not significant, with an exception of Naive Bayes which is the worst-performing technique. It can also be observed from the table that the precision and recall scores are comparable.

6.2. Fine-grained multiclass (Level 2) classification

In this set of experiments, the classifier is given a hateful tweet as input, it classifies it into one of the four categories, interfaith, sectarian, ethnic, or other. The macro average F1 scores of the fine-grained multiclass classification are also presented in Table 13. It can be observed from the table that similar to the binary classification, NB is also the worst performing for the second level classification as it merely achieved an F1 score of 0.456. In contrast, it can be observed that BERT achieved the best F1 scores of around 0.832.

By comparing the results of the binary and multiclass classification, a notable observation is that the F1 score achieved by NB for multiclass classification is substantially lower than that of the binary classification ($0.468 < 0.531$). Furthermore, the F1 scores achieved by all the techniques for the multiclass classification are lower than the corresponding binary classification. This decrease in performance is due to two reasons. First, binary classification is widely recognized as an easier task than multiclass classification. Second, the size of the training dataset for the multiclass classification is lower than that of the binary classification, which is impeding the learning ability of the machine learning techniques.

6.3. Single level Multiclass Classification

In third set of experiments, a tweet is given to a classifier, and it classifies the tweet into one of the five classes, neutral, interfaith, sectarian, ethnic, or other. Table 14 presents the macro average results of the single-level multiclass classification. It can be observed from the table that NB is the least effective technique with F1 score of 0.324, whereas BERT is the most effective technique with F1 score of 0.740.

The best F-scores of the third set of experiments in Table 14 are less than the best F1 scores from the second set of experiments where a Hateful tweet is classified into one of the four hateful categories as shown in Table 13. The reason is that in the second set of experiments, the classifier is only given Hateful tweets, whereas, in the third set of experiments, the classifier is given all types of tweets. In this set of experiments, the classifier is doing both types of classification simultaneously. It is classifying Hateful tweets from Neutral tweets and then also doing fine-grained classification of hateful tweets into one of the four categories. This set of experiments is more close to real settings since the real dataset will contain all types of tweets.

It is observed from results that BERT has outperformed other deep learning and machine learning models with a substantial margin in both single level multiclass classification (Table 14) and two level fine grained classification (Table 13). Single level multiclass classification is more challenging for the classifiers and superior performance of BERT in these experiments shows the power of self attention mechanism used in BERT. Attention mechanism has been popular for deep learning models in image processing tasks but during last few years it has also gained much attention from NLP community due to its effectiveness in Natural Language Processing (NLP) tasks. BERT uses multi-layered self attention mechanism which models associations between neighbouring words to produce sequence based word representations [48]. These representations capture contextual dependencies for creating meaningful representations of long sequences of text. The dot-product attention plays an important role in improving text classification results [49]. Meaning of some words is dependant on the context in which they are used and attention mechanism uses information from all surrounding words to build better representations. The large difference in performance of BERT as compared to other models proves that creating meaningful representations of text using attention mechanism makes the prediction model more robust.

Type	Technique	Layers	Precision	Recall	F1 score
Classical	NB		0.325	0.377	0.324
	RF		0.693	0.565	0.597
	LR		0.724	0.640	0.671
	SVM		0.717	0.664	0.685
Deep learning	CNN	1	0.686	0.681	0.681
		2	0.684	0.691	0.686
		3	0.687	0.677	0.680
	LSTM	1	0.678	0.628	0.645
		2	0.678	0.657	0.663
		3	0.685	0.640	0.650
	Bi-LSTM	1	0.687	0.646	0.662
		2	0.670	0.651	0.657
		3	0.672	0.656	0.661
	BERT		0.727	0.760	0.740

Table 14: Results of the single level multiclass classification

	Naive Bayes		BERT	
	Hateful	Neutral	Hateful	Neutral
Hateful	7,223	1,429	6,687	1,965
Neutral	8,586	4,521	1,444	11,663

Table 15: Confusion matrix of the least and most effective technique for binary (Level 1) classification

6.4. Error Analysis

To identify the causes of the misclassification, the incorrect predictions of the best and the worst performing techniques were analyzed. In particular, both quantitative and qualitative approaches were used

	Naive Bayes				BERT			
	Interfaith	Sectarian	Ethnic	Other	Interfaith	Sectarian	Ethnic	Other
Interfaith	848	165	80	444	1411	36	1	82
Sectarian	309	814	65	402	88	1,423	18	55
Ethnic	9	25	18	115	5	20	104	38
Other	819	489	185	3,928	136	80	71	5,134

Table 16: Confusion matrix of the least and most effective technique for multiclass (Level 2) classification

for the analysis. The quantitative approach relies on the confusion matrix which gives an overview of the distribution of misclassified tweets, while giving no regard to the content of the tweets. In contrast, the qualitative approach relies on the content of the misclassified tweets. Table 15 presents the confusion matrix of Naive Bayes and BERT, the worst and best-performing techniques, for the binary classification. It can be observed from the table that the number of hateful tweets that are correctly classified by Naive Bayes is higher than that of BERT (7,223 > 6,687). In contrast, the number of neutral tweets that are correctly classified by BERT is significantly higher than that of Naive Bayes (11,663 > 4,521). Furthermore, the number of neutral tweets that are misclassified by Naive Bayes is much higher than BERT (8,586 > 1,444). These values represent the overfitting problem of Naive Bayes, i.e. in most cases Naive Bayes incorrectly classifies neutral tweets as hateful tweets.

From the confusion matrix, it can be observed that 6,687 hateful and 11,663 neutral tweets are correctly classified by BERT, whereas a smaller number of tweets are misclassified. Furthermore, it can be observed that the number of incorrectly classified neutral tweets (1,444) is comparable with the number of incorrectly classified hateful tweets (1,965) which represents that the overfitting problem does not exist for BERT. To further understand the causes of misclassification by BERT, a qualitative analysis of the misclassified tweets is performed. That is, the misclassified tweets are extracted and separated into incorrectly classified hateful tweets and incorrectly classified neutral tweets. Subsequently, we manually reviewed these tweets to identify that several tweets contain words that are typically used in hateful tweets, however, the overall meaning of the tweet is not hateful. For instance, a majority of tweets that use the word 'Qadiani' or its inflectional forms are hateful, however there are several tweets that use the word 'Qadiani' but they do not reflect hate. Similarly, there are numerous misclassified tweets that contains multiple words that are commonly used in hateful tweets but the overall meaning of the tweet is not hateful.

For the qualitative analysis, Figure 8 presents the word cloud of the neutral tweets that were misclassified as hateful, whereas Figure 9 presents the word cloud of the hateful tweets that were misclassified as neutral. It is observed that there is a significant overlap of vocabulary between the misclassified hateful and neutral tweets, i.e. the word 'Sindh' appears 135 times in the misclassified neutral tweets and 294 times in the misclassified hateful tweets. Similarly, 'Pakistan', 'haqoomat' (government), 'Bhutto', and 'Corona' are the other words that frequently occur in the misclassified neutral and hateful tweets. This analysis shows that these words are frequently used in both contexts.

Similar to the binary classification, an error analysis is also performed for the multiclass classification using both qualitative and quantitative approaches. Table 16 presents the confusion matrix of the best and the worst performing techniques, i.e. Naive Bayes and BERT. It can be observed from the table that Naive Bayes misclassified 819 other types of tweets as Interfaith and another 489 as Sectarian, representing the under-fitting problem. From the confusion matrix of BERT, it can be observed that the two classes of tweets that were misclassified are Interfaith and Others, which represents that BERT does not overfit.

To further understand the causes of misclassification by BERT, we performed a qualitative analysis of all the misclassified tweets. That is, the Interfaith tweets that were misclassified as Other and the Other tweets that were misclassified as Interfaith were separated and a word cloud of the vocabulary of these tweets is generated. Figure 10 presents the word cloud of the misclassified Interfaith tweets, whereas Figure 11 presents the word cloud of the other tweets that were misclassified as Interfaith. It can be observed from the two figures that there are fewer words that are common between the two classes. Furthermore, it can be observed from Figure 11 that the Other tweets that are often misclassified as Interfaith contain several

Figure 8: Word cloud of misclassified tweets by BERT Layer 1. Actual class is neutral and predicted is hateful

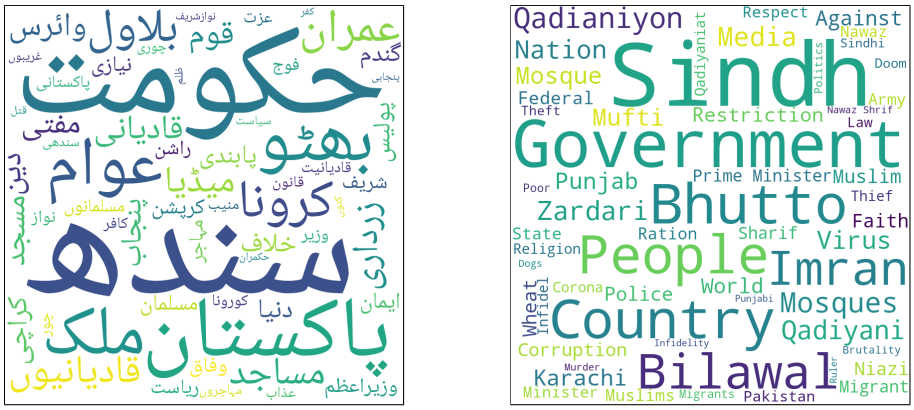

Figure 9: Word cloud of misclassified tweets by BERT Layer 1. Actual class is hateful and predicted is neutral

Figure 10: Word cloud of misclassified tweets by BERT Layer 2 classifier. Actual class is Interfaith and predicted class is Other

words or their inflectional morphemes that are frequently found in Interfaith tweets. For a further analysis, we manually examined the incorrectly classified tweets to identify the types of tweets that are more difficult to classify. The types of tweets are the following, whereas the examples of each type are presented in Table 17.

- Interfaith misclassified as Others. The tweets that blames a person or a group for abandoning religious

Figure 11: Word cloud of misclassified tweets by BERT Layer 2 classifier. Actual class is Other and predicted class is Interfaith

or not following the practices of religion in true spirit without using hateful words were mostly misclassified as Others. Table 17 presents the three examples (E1 - E3) of such tweets that are misclassified as Others.

- Interfaith misclassified as Sectarian. This is a unique case where most of the Sectarian comments that were identified as Interfaith were the ones with multiple labels. In the table, E4 and E5 are the example tweets having multiple labels. Another notable observation is that the several misclassified Interfaith tweets either did not include an abusive word or the name of the target group was not mentioned in the tweet, i.e. their associated believes, region or ritual was used to identify them. E6 is an example of a tweet in which the target group can be identified by their place.

- Sectarian misclassified as Interfaith. A careful examination of the misclassified Sectarian tweets revealed that most of the Sectarian comments that were identified as Interfaith were the ones having multiple annotations. E7 - E9 are the examples of the tweets which were Interfaith as well as Sectarian.

- Ethnic misclassified as Others. It was observed from the analysis that in a majority of the misclassified tweets the twisted name of the targeted ethnicity was used to refer to the community which makes it harder to classify such sentences. E13 - E15 are the examples of such tweets in which the twisted name of the ethnic group is used.

7. Conclusion and Future Work

Sectarian and ethnic violence is an ongoing issue in Pakistan where Urdu is the most widely spoken language. Various religious and ethnic groups use social media to spread hatred for other sects and ethnic groups in the Urdu language. Interfaith, sectarian, and ethnic hatred on social media is an offensive crime. Several studies have been conducted for the detection of hateful content in the English language, however, no study has been conducted for interfaith, sectarian, and ethnic hatred in the Urdu language which is widely used in Pakistan. Furthermore, no benchmark labeled dataset exists in the literature for this problem in the Urdu language. This study contributes to existing work on cybercrime detection on social media in the following two ways. The first contribution is the generation of the first cybercrime Urdu Corpus, formally ISE-Corpus, for interfaith, sectarian, and ethnic hate detection in Urdu. The corpus was labeled manually by two assessor at two levels. The first level assigns binary labels of hateful and neutral, whereas, the second level assigns multiclass labels of "Interfaith", "Sectarian", "Ethnic", and "Other" to each tweet in the dataset. The dataset was annotated based on guidelines developed iteratively. Separate guidelines were developed for both types of labels (binary and multiclass). There are two advantages of developing these guidelines. First, the guidelines helped in increasing assessor agreement and subsequently improved Kappa score. Second, these guidelines can be used in the future to extend the dataset.

Misclassified	ID	Misclassified example
Interfaith → Others	E1	One reason for not doing Hujj by Imran Khan is that in the application he demanded that he will perform all the rituals except he will not stone the devil.
		حج نہ کرنے کی اک وجہ ویزادرخواست میں عمران خان کا یہ مطالبہ بھی تھا کہ وہ تمام مناسک اداکریگامگرشیطان کوپتھر نہیں مارے گا
	E2	He would be someone like you who has abandoned Islam.
		وہ تم جیسا کوئی ہوگا جس نے اسلام کو چھوڑ دیا ہو۔
	E3	They praise Imran Khan. Imran Niazi is addict. His children are Jews. They eat Ham.
		وہ عمران خان کی تعریف کرتے ہیں۔ عمران نیازی نشے کا عادی ہے۔ اس کے بچے یہودی ہیں۔ وہ ہام کھاتے ہیں۔
Interfaith → Sectarian	E4	Clean from Shia and Qadiyani.
		شیعہ اور قادیانی سے پاک
	E5	Shia and Qadiyani are the two sides of one coin.
		شیعہ اور قادیانی ایک سکے کے دو رخ ہیں۔
	E6	A dog from Rabwa.
		ربوہ کا ایک کتا
Sectarian → Interfaith	E7	Shia and Quadiyani have the same believes. The ruling elite is requested that the Constitution of Pakistan should be amended and Shia should also be declared as non-muslins.
		شیعہ اور قادیانی کا عقیدہ ایک ہی ہے۔ حکمران طبقہ سے گزارش ہے کہ پاکستان کے آئین میں ترمیم کی جائے اور شیعہ کو بھی غیر ململ قرار دیا جائے۔
	E8	Quadiyani and Rafaizi are the worst and dirtiest non-muslims and those who do not recognize them ans non-muslims are also non-muslims.
		قادیانی اور شیعہ دونوں کائنات کے بدترین کافر ہیں جو نا مانے وہ بھی کافر
	E9	In my view Qadiyani and Shia are two sides of the same coin and both have abandoned and non-muslims. And anyone having has softcorner he is also non-muslim. If any Qadiyani and Shia wants to have debate I am available.
		میرے خیال میں قایانی اور شیعہ ایک ہی سکے کے دو روخ ہیں دونوں مرتد زندیق اور کافر ہیں جو ان کے بارے میں دل میں نرم گوشہ رکھے وہ بھی کافر اگر کسی قادیانی اور شیعہ نے مناظرہ کرنا ہے تو میں حاضر ہوں
Ethnic → Other	E13	Job opportunities in Sindh and Domicile Punjabistani is necessary. It necessary to get freedom from Non-Punjabistan.
		ملازمت کے مواقع سندھ میں اور ڈومیسائل پنجابستان کہ ضروری ہے اس ناپنجابستان سے آزادی ضروری ہے
	E14	Sindhi Sindhri Billorani you have to answer Karachi is the capital of which province.
		سندھی سڈھی بلورانی جواب تو دینا ھوگا کراچی کونسے صوبہ کا دارالخلافہ ھے.
	E15	A Baloch is installed on Punjabis. This is the curse of God on Hunjabis.
		پنجابیوں پر ایک بلوچ لگا ہوا ہے۔ یہ پنجابیوںس پر خدا کی لعنت ہے۔

Table 17: The misclassified example for the second level classification

The second contribution is the application of state-of-the-art machine learning and deep learning models for the automatic classification of tweets. We have applied different data cleaning and text preprocessing steps to the collected dataset. Afterward, four machine learning techniques four deep learning techniques were applied for the identification of hateful and subsequently classifying it into Interfaith, sectarian and ethnic hate identification. The results of the experiments show that BERT is the most effective technique as it achieved a higher F1 score for the fine-grained multiclass classification and a comparable F1 score for the binary classification.

We firmly believe that this study will inspire other researchers to work on this critical issue of hateful content detection on social media for Urdu language. There are several future directions for this study. One direction is to increase the size of the labeled dataset using guidelines and to use data augmentation techniques for generating a balanced dataset to evaluate the effectiveness of techniques. One possibility is employing undersampling techniques to reduce the sample size of the majority class. Whereas, another approach is to use an oversampling technique which proposes to increase the sample size of the minority class. Furthermore, there could be multiple strategies for increasing the sample size of the minority class. For instance, to scrap more data and search for Ethnic tweets. Another way could be to augment the existing examples and generate synthetic examples. Another direction is to try different linguistic, syntactic, and semantic features and emoticons for the automatic classification of tweets. Similar guidelines can be used for other low-resource languages for developing benchmark datasets in those languages.

8. Ethical Considerations

We have considered the ethical and legal issues related to collecting and processing Twitter data [50]. Furthermore, the data is gathered in compliance with the legal and ethical standards recognized by the scientific community. In particular, the tweets are collected by using a scraper from Twitter and the released dataset is made anonymous. It is widely established that transformer-based models like BERT learn biases from the text when pre-trained on large amounts of text [51]. Examples of biases include gender bias, such as associating females with certain professions or role. Typically, these biases occur due to pre-training of BERT on large amounts of raw text. However, in our study, BERT is only trained on our developed corpus. Due to that reason, the learned word meaning representations of BERT have a low chance of biases.

As discussed earlier, the developed ISE-hate dataset and the working code used in this study is released for the public. These released resources will serve as catalyst for advancement of hateful content detection in the Urdu tweets. However, the released of these resources may also be misused. For instance, the offenders may rephrase the controversial content to dodge a systems that is developed using the datasets or the trained models. This can be addressed by continuously developing a pipeline to enhance the training dataset and optimizing the machine learning model used for the detection of the hateful content.

References

[1] I. Anger, C. Kittl, Measuring influence on Twitter, in: Proceedings of the 11th international conference on knowledge management and knowledge technologies, 2011, pp. 1–4.

[2] A. Nasrallah, N. Sarkis, The role of social media during the Arab spring, in: Business and Social Media in the Middle East, Springer, 2020, pp. 121–136.

[3] A. Bovet, H. A. Makse, Influence of fake news in twitter during the 2016 US presidential election, Nature communications 10 (1) (2019) 1–14.

[4] F. Riquelme, P. González-Cantergiani, Measuring user influence on Twitter: A survey, Information processing & management 52 (5) (2016) 949–975.

[5] H. Watanabe, M. Bouazizi, T. Ohtsuki, Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection, IEEE Access 6 (2018) 13825–13835.

[6] M. Barhamgi, A. Masmoudi, R. Lara-Cabrera, D. Camacho, Social networks data analysis with semantics: Application to the radicalization problem, Journal of Ambient Intelligence and Humanized Computing (2018) 1–15.

[7] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, I. Awan, Detection and classification of social media-based extremist affiliations using sentiment analysis techniques, Human-centric Computing and Information Sciences 9 (1) (2019) 1–23.

[8] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2) (2021) 477–523.

[9] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, PloS one 14 (8) (2019) e0221152.

[10] Ò. G. i Orts, Multilingual detection of hate speech against immigrants and women in twitter at semeval-2019 task 5: Frequency analysis interpolation for hate in speech detection, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 460–463.

[11] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 75–86.

[12] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020), in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1425–1447.
[13] J. Pavlopoulos, J. Sorensen, L. Laugier, I. Androutsopoulos, Semeval-2021 task 5: Toxic spans detection, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 59–69.
[14] B. Dean, Cybercrime act of Pakistan, http://www.nr3c.gov.pk/law.html, accessed: 2021-10-08 (2016).
[15] The world's languages, in 7 maps and charts, https://www.washingtonpost.com/news/worldviews/wp/2015/04/23/the-worlds-languages-in-7-maps-and-charts/.
[16] S. Kanwal, K. Malik, K. Shahzad, F. Aslam, Z. Nawaz, Urdu Named Entity Recognition: Corpus generation and deep learning applications, ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 19 (1) (2019) 1–13.
[17] A. Nawaz, M. Bakhtyar, J. Baber, I. Ullah, W. Noor, A. Basit, Extractive text summarization models for Urdu language, Information Processing & Management 57 (6) (2020) 102383.
[18] Twitter, Twitter Rules and Policies, https://help.twitter.com/en/rules-and-policies#twitter-rules, accessed: 2020-12-29.
[19] R. Jahangir, Twitter complied with 35pc legal requests from Pakistan in 2019, https://www.dawn.com/news/1577311, accessed: 2020-12-29.
[20] Research on 100 million Tweets 2018, https://www.vicinitas.io/blog/twitter-social-media-strategy-2018-research-100-million-tweets, accessed: 2020-12-29.
[21] N. U. Haq, M. Ullah, R. Khan, A. Ahmad, A. Almogren, B. Hayat, B. Shafi, USAD: An intelligent system for slang and abusive text detection in PERSO-Arabic-scripted Urdu, Complexity 2020.
[22] Twitter, Supported languages and browsers, https://developer.twitter.com/en/docs/twitter-for-websites/supported-languages, accessed: 2020-12-29.
[23] A. Pro, Pakistan Social Media Stats 2018, http://alphapro.pk/pakistan-social-media-stats-2018/, accessed: 2020-12-29.
[24] T. R. Soomro, S. M. Ghulam, Current status of Urdu on Twitter, Sukkur IBA Journal of Computing and Mathematical Sciences 3 (1) (2019) 28–33.
[25] R. Batra, Z. Kastrati, A. S. Imran, S. M. Daudpota, A. Ghafoor, A large-scale tweet dataset for Urdu text sentiment analysis.
[26] H. Rizwan, M. H. Shakeel, A. Karim, Hate-speech and offensive language detection in Roman Urdu, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 2512–2522.
[27] M. M. Khan, K. Shahzad, M. K. Malik, Hate speech detection in Roman Urdu, ACM Transactions on Asian and Low-Resource Language Information Processing 20 (1) (2021) 1–19.
[28] M. Sohail, A. Imran, H. U. Rehman, M. Salman, Anti-social behavior detection in Urdu language posts of social media, in: 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), IEEE, 2020, pp. 1–7.
[29] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, M. T. Sadiq, Automatic detection of offensive language for urdu and Roman Urdu, IEEE Access 8 (2020) 91213–91226.
[30] R. U. Mustafa, M. S. Nawaz, J. Farzund, M. Lali, B. Shahzad, P. Viger, Early detection of controversial Urdu speeches from social media, Data Sci. Pattern Recognit. 1 (2) (2017) 26–42.
[31] S. Kausar, B. Tahir, M. A. Mehmood, ProSOUL: a framework to identify propaganda from online Urdu content, IEEE Access 8 (2020) 186039–186054.
[32] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. AbdelMajeed, T. Zia, Abusive language detection from social media comments using conventional machine learning and deep learning approaches, Multimedia Systems (2021) 1–16.
[33] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, A. Gelbukh, Threatening language detecting and threatening target identification in Urdu tweets, IEEE Access.
[34] A. Aslam, H. Tariq, Modeling toxicity in social media for Urdu text (2018).
[35] M. Z. Ali, A. Ehsan-ul Haq, S. Rauf, K. Javed, S. Hussain, Improving hate speech detection of Urdu tweets using sentiment analysis, IEEE Access.
[36] A. Dewani, M. A. Memon, S. Bhatti, Development of computational linguistic resources for automated detection of textual cyberbullying threats in roman urdu language, 3C TIC. Cuadernos de desarrollo aplicados a las TIC (2021) 101–121.
[37] H. H. Saeed, M. H. Ashraf, F. Kamiran, A. Karim, T. Calders, Roman Urdu toxic comment classification, Language Resources and Evaluation (2021) 1–26.
[38] Z. Ansari, S. Ali, F. Khan, Use of Roman script for writing Urdu language, International Journal of Linguistics and Culture 1 (2) (2020) 165–178.
[39] Top hashtags in Pakistan, https://getdaytrends.com/pakistan/top/longest/year/, accessed: 2020-09-30 (2020).
[40] W. S. Noble, What is a Support Vector Machine?, Nature Biotechnology 24 (12) (2006) 1565–1567.
[41] A. A. Chandio, M. Pickering, K. Shafi, Character classification and recognition for Urdu texts in natural scene images, in: International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), IEEE, 2018, pp. 1–6.
[42] D. Phuc, N. T. K. Phung, Using naïve Bayes model and natural language processing for classifying messages on online forum, in: IEEE International Conference on Research, Innovation and Vision for the Future, IEEE, 2007, pp. 247–252.
[43] D. HOSMER JR, S. Lemeshow, R. STUDIVART, Applied logistic regression 2nd ed (1987).
[44] H. Peng, J. Li, Y. He, Y. Liu, M. Bao, L. Wang, Y. Song, Q. Yang, Large-scale hierarchical text classification with recursively regularized deep Graph-CNN, in: Proceedings of the International Conference on World Wide Web, 2018, pp. 1063–1072.
[45] D. Tang, B. Qin, X. Feng, T. Liu, Effective lstms for target-dependent sentiment classification, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 3298–3307.

[46] Papers with code, https://paperswithcode.com/method/bilstm, accessed: 2020-09-30 (2020).
[47] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, B. Xu, Text classification improved by integrating bidirectional lstm with two-dimensional max pooling, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 3485–3495.
[48] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[49] X. Sun, W. Lu, Understanding attention for text classification, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3418–3428.
[50] F. Rangel, P. Rosso, On the implications of the general data protection regulation on the organisation of evaluation tasks, Language and Law/Linguagem e Direito 5 (2) (2019) 95–117.
[51] P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, K. Kersting, Large pre-trained language models contain human-like biases of what is right and wrong to do, Nature Machine Intelligence 4 (3) (2022) 258–268.