IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 July 2024, accepted 12 August 2024, date of publication 16 August 2024, date of current version 2 September 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3444188*

RESEARCH ARTICLE

# Hate Speech and Target Community Detection in Nastaliq Urdu Using Transfer Learning Techniques

**MUHAMMAD SHAHID IQBAL MALIK[1,2], AFTAB NAWAZ[3], AND MONA MAMDOUH JAMJOOM[4]**

[1]Department of Computer Science, University of Wah, Wah Cantt 47040, Pakistan
[2]Department of Computer Science, National Research University Higher School of Economics, 109028 Moscow, Russia
[3]Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock 43600, Pakistan
[4]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia

Corresponding author: Muhammad Shahid Iqbal Malik (mumalik@hse.ru)

This work was supported by Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia, through the Princess Nourah Bint Abdulrahman University Researchers Supporting Project under Grant PNURSP2024R104.

**ABSTRACT** Freedom of expression on social media has provided oppressed people with many opportunities to raise their voices against violence and injustice, but this freedom is being misused to spread various forms of hate speech. Several studies have been conducted to identify hate speech in high-resource languages, however, work on under-resource languages is very limited, especially for Nastaliq Urdu. Pakistan has been dealing with the issue of hateful and violence incitation content for the last two decades. Therefore, this study handled the problem of detecting hate speech and fine-grained multi-class target community identification in Nastaliq Urdu. Using the transfer learning paradigm, two benchmark Urdu transformer models are explored with fine-tuning. A Nastaliq Urdu Hate Speech and Target Community (HSTC) corpus is designed by collecting posts from Pakistani Facebook accounts. In particular, the strengths of the Urdu Robustly Optimized BERT Pre-Training Approach (Urdu-RoBERTa) and Urdu Distillated Bidirectional Encoder Representations from Transformers (Urdu-DistilBERT) are explored to design an automated system instead of hand-crafted features. The proposed framework consists of four steps: 1) data cleaning and preprocessing; 2) data transformation; 3) utilization of Grid search for fine-tuning process; and 4) classification (binary and multi-class). The results on the Nastaliq Urdu corpus showed that the proposed system achieved benchmark performance for binary classification task (hate speech) and target community detection (multi-class classification) on hateful Facebook posts. In particular, fine-tuned DistilBERT achieved 86.58% accuracy and 86.52% f1-score for binary classification and outperformed sixteen baselines. Furthermore, it demonstrated 84.17% accuracy and 83.91% f1-score for target community (religious, political, and gender-based) identification and outperformed all baselines. The findings of this study can be beneficial in detecting and filtering out hate speech in Nastaliq Urdu on the Facebook platform.

**INDEX TERMS** Nastaliq Urdu, target community, hate speech, DistilBERT, fine-tunning, Facebook.

## I. INTRODUCTION

Social media platforms have provided many opportunities to online users such as communication with friends and family, and expressing/discussing ideas freely. However, these platforms are often used for spreading hate speech and toxic comments [1]. Several definitions of hate speech exist in the literature [2], [3], [4]. However, the majority of researchers have a consensus on the following definition: "Hate speech is a language used to attack/target an individual or a group on the basis of ethnicity, race, gender, or religion" [5], [6]. The interactive nature of social media platforms allows the

The associate editor coordinating the review of this manuscript and approving it for publication was Sawyer Duane Campbell.

dissemination of hateful content, resulting in organization of hate crimes which is quite difficult to handle by traditional law enforcement organizations. In addition, manually identifying such content requires a lot of human effort. Hence, there is a need for automated identification systems, which is also a challenge due to the dynamic nature of hate speech. In particular, hateful content evolves and is highly subjective [7].

Social media organizations (Facebook and Twitter, etc) are under public and political pressures to deal with hateful content due to rise in online hate speech. The recent prevention measures adopted by social media are not effective in detecting such unwanted material and protecting social media users from harm. The failures are due to the following reasons: 1) it is laborious efforts to tag a large collection of online content as hate/neutral, which also results in discrimination due to subjective views and personal biases, 2) to classify online content as hate/neutral is challenging in the presence of a diversity of languages, 3) it is an un-reliable approach to expect from users to report online harmful material fully, resulting in un-identified a lot of material, and 4) Developing an automated system for detecting hateful content using the NLP paradigm is quite challenging due to the presence of linguistic nuances in context capture. Furthermore, hundreds of languages are spoken in different regions of the world and it is not possible to design a single classification system for two or more languages.

In literature, several studies attempted to develop systems for identifying hate speech in high-resource languages such as English [8], and French [9], etc. However, hundreds of languages are spoken in different regions of the world, but research on automated systems for hate speech detection in low-resource languages, especially in Nastaliq Urdu is limited. Some studies addressed this task by proposing detection models for under-resource languages such as Arabic [10], Roman Urdu [11], and Nastaliq Urdu [12]. Urdu is the national language of Pakistan with a population size of approximately 242 million. In addition, more than 300 million people speak Urdu in other regions of Asia, Europe, and the USA such as India, UK, USA, and Canada. It has two writing styles; Nastaliq and Roman. As hate speech is increasing rapidly, there is a high demand to develop automated and accurate models for identifying hate speech in Urdu. The study [13] developed a model to classify hate speech in Nastaliq Urdu tweets. They used BOW features with SVM and NB models but did not handle target community detection. Later, the study [14] explored hate speech in Urdu at the first level and categorized it into three classes (hate, simple-complex, or neutral). TF-IDF and word2vec techniques are used with SVM and RF models. Another study [15] introduced a model for hate speech detection in Urdu. They explored word unigram, and embedding model for feature generation but missed target community detection and did not make their dataset public. Recently, a study [16] designed an identification model for hate speech in Urdu but did not handle target community detection. They used word and char n-grams with LR and CNN models and achieved significant performance. We found the following issues in the prior studies; First, the majority of the approaches in Nastaliq Urdu used hand-crafted (linguistic, semantic, or frequency-based) features for hateful content identification. Second, vulnerable target community identification on the hateful content is not addressed by the prior studies. Third, to the best of our knowledge, no public dataset is available in Nastaliq Urdu that supports two levels of hate speech classification. To address these issues, this study devised the following research objectives:

1. To develop an automated and high-performance detection framework to classify Nastaliq Urdu Facebook posts into hate or not-hate classes.
2. Considering hateful posts; design an accurate and automated system to detect the vulnerable community being targeted based on political, religious, or gender-related conflicts.
3. To explore the strengths of Urdu-RoBERTa and Urdu-DistilBERT with fine-tuning to design a high-performance multi-level (coarse and fine-grained) classification system.

To achieve these objectives, we propose a completely automated end-to-end robust framework for binary (coarse-grained) hateful content identification and fine-grained multi-class target community detection in Nastaliq Urdu on the Facebook platform. The potentials of two state-of-the-art Urdu transformers (RoBERTa and DistilBERT) are explored with fine-tuning in the design of an automated model. To evaluate the efficacy and robustness of the proposed framework, we designed an annotated Nastaliq Urdu corpus by collecting posts from Pakistani Facebook accounts. Furthermore, sixteen SOTA baselines are used to compare the effectiveness of the proposed framework for the said task. The salient contributions of this study are presented below:

1) This study introduced a detection system (first attempt) that used contextual embedding models with fine-tuning to classify two levels of Nastaliq Urdu hate speech on Facebook.
2) The potentials of Urdu-RoBERTa and Urdu-DistilBERT are explored with fine-tuning on the newly designed corpus to handle the complexity and ambiguity concerns of Nastaliq Urdu.
3) The fine-tuning is accomplished with the grid search technique for choosing optimal values of eight hyper-parameters for both levels of classification tasks.
4) The results showed that the fine-tuned Urdu-DistilBERT exhibited benchmark performance and outperformed sixteen SOTA baselines including Urdu-RoBERTa.
5) In particular, the improvement achieved by the proposed framework for detecting hateful class at the first level and political and religious communities at the second level is promising.

The remaining parts of the article are described as; related work is presented in section II, followed by section III, which describes the process of corpus construction for coarse

and fine-grained hate speech detection. Section IV presents the proposed methodology and experimental setup in detail. Section V provides the results and Section VI presents the discussion and limitations of the study. Section VII describes the conclusion and future suggestions.

## II. RELATED WORKS

Various studies are conducted to address the tasks of controversial or unwanted text identification using NLP techniques. The tasks include the identification of abusive, offensive, controversial, toxic, and hate speech content. The majority of previous work has dealt with resource-rich languages such as English and some European languages, but work in low-resource languages has been limited. In this section, we restrict the discussion of related studies to Roman Urdu and Urdu languages.

### A. HATE SPEECH CLASSIFICATION

In 2017, a study gained the attention of researchers and proposed a detection model for controversial content on Twitter [17]. The TF-IDF weighting method with the traditional ML model is used and their model presented significant performance. Later, a study [18] developed an Urdu corpus for spotting propaganda in news articles. Various word embedding, semantic, and lexicon models are explored and word n-gram delivered the highest performance (91%).

**Glossary of Key Terms**

| | |
|---|---|
| MLP | Multilayer Perceptron |
| ML | Machine Learning |
| NLP | Natural Language Processing |
| BERT | Bi-directional |
| HSTC | Hate Speech and Target Community |
| DL | Deep Learning |
| SOTA | State-of-the-art |
| CNN | Convolutional Neural Network |
| RoBERTa | Robustly Optimized BERT Pre-training Approach |
| BERT | Bidirectional Encoder Representations from Transformers |
| DistilBERT | Distilled-BERT |
| HPC | High-Performance Computing |
| RF | Random Forest |
| SVM | Support Vector Machine |
| LR | Logistic Regression |
| TL | Transfer Learning |
| Bi-LSTM | Bidirectional Long Short Term Memory |
| BOW | Bag of words |
| NB | Naïve Bayes |

Likewise, another study [19] introduced the detection model for offensive content in Roman and Nastaliq Urdu. The word and char n-grams are used with seven ML models and their framework achieved promising performance. For slang and abusive word identification, the study [20] proposed a custom model for the detection of unwanted content. They designed a lexicon for feature generation and introduced a custom classification model. All these studies developed classification systems for controversial content but missed further categorization of unwanted content.

Then, a study [21] developed an anti-social behavior detection model for Nastaliq Urdu. The authors explored lexicon-based features and compared their model with baselines. Likewise, five types of abusive content are categorized by [22] and the proposed model explored various embedding methods to design an effective solution. The offensive, sexist, profane, religious hate, and normal are the classes. Their model achieved 82% performance. Later in 2021, a study proposed a model for sentiment analysis based on hate content identification [13] for Twitter. The authors used several techniques such as dynamic stop word filtering to handle the sparsity, and data imbalance issues. Their model achieved significant performance. Likewise, another study designed a detection model in the Roman Urdu language for multi-class hate speech detection [14]. TF-IDF and word2vec models are used with five ML models and their model obtained significant performance. A recent study introduced a model for identifying cyber bullying in Urdu tweets [23]. They made a public dataset and their model demonstrated significant performance with FastText and LSTM model. Likewise, another study proposed a detection model for cyber bullying in English and Urdu comments [24]. They explored several ML and DL models and obtained best performance with SVM model.

Some studies in the literature addressed threatening content detection task such as an article [25] proposed a system for threatening text and target identification in tweets but their dataset has annotation issues. The proposed model presented the best performance with MLP + word n-gram. Then, another study developed a detection model for abusive and threatening content identification in Nastaliq Urdu [26]. The BERT model is utilized for classification task but they did not address the community identification task. The study [27] proposed a system for threatening content identification in Urdu. The word and char n-grams are used with four DL models including the ensemble model. The top performance is achieved by the ensemble model. Likewise, a recent study proposed a system to address the issue of threatening content and target identification on Twitter [6]. The authors used semantic features and the BERT model with fine-tuning. Their model outperformed the baselines. Recently, a study proposed a multi-lingual framework for threatening content identification in Nastaliq Urdu and English languages using TL [28]. They explored transformer models with fine-tuning and presented 89% f1-score as the top performance. Another recent study proposed a detection model for threatening content in the Urdu language [29]. Their model used a deep sequential approach and achieved an accuracy of 82%.

Some studies addressed this issue as a toxic content identification framework such as a study that introduced a comparatively big corpus for toxic content identification [30]. Their model used various word embedding models with an

ensemble deep learning model and achieved benchmark performance. For Facebook posts, authors [31] introduced an offensive content detection model for Nastaliq Urdu. Various semantic models are explored with traditional ML models and their model outperformed the baseline but they did not handle community detection task. Likewise, this study [32] developed an Urdu dataset in Nastaliq script and proposed a binary classification model for offensive content detection. Their study explored word and char n-grams with various ML and DL models and demonstrated benchmark performance. Recently, a study introduced a spotting model [15] for hateful content and its categorization based on Ethnic, Sectarian, and Interfaith opinions. Their model obtained a macro f1-score of 83.9% as the top performance. Multiple levels of classification for offensive and hate speech are performed by the study [16]. It is the first study that categorized hate speech on the basis of severity. Extensive experiments reveal that their model outperformed the baselines. Recently, the study [33] introduced a framework for violence incitation detection in Nastaliq Urdu and used the Twitter platform for data collection. They used the CNN model with word uni-gram and obtained the highest performance compared to baselines. Likewise, a recent study proposed a model for binary abusive content identification in Urdu [34]. The CNN and Bi-LSTM are utilized with TF-IDF, word2vec, and FastText embeddings. The best performance is achieved with Bi-LSTM and word2vec. Another work [35] developed an explainable fine-grained hate speech detection model for roman Urdu. They used the LSTM model and designed explainable architecture. Their model achieved state-of-the-art performance for coarse and fine-grained hate speech detection.

### B. LIMITATIONS OF THE EXISTING RESEARCH

From the comparison of prior studies proposed for Roman and Nastaliq Urdu, we observe the following limitations:

- **Lack of research on target community detection:** To the best of our knowledge, no work has been conducted on targeted community detection of hateful speech in Nastaliq Urdu.
- **Lack of annotated corpus:** According to our knowledge, we did not find any annotated dataset for two-level categorization (hate speech and target community detection)
- **Lack of automated classification methods:** The majority of prior approaches derived classification systems using hand-crafted features.
- **Lack of comparison of supervised methods:** Limited studies performed comparisons between traditional ML and DL models to define the best model.

To handle these issues, we designed a framework for detecting hate speech at the first level and targeted community detection at the second level. A Nastaliq Urdu corpus is developed that consists of two levels of annotations. The strengths of two benchmark transformer models are explored to design a robust automated classification system. Extensive comparisons of conventional ML and DL models are conducted to describe the best model.

## III. HSTC CORPUS FOR NASTALIQ URDU

In this section, the process of building the annotated corpus for hate speech and target community detection is described. An overview of the process used to create an HSTC corpus is presented in **Fig. 1**. It is evident that the process is composed of six steps as described in the figure. The first three steps aim to collect relevant data from Facebook pages. The details of these steps are given in the following.

### A. DATA SCRAPING AND SELECTING PAGES

The scraping of Urdu Nastaliq content is conducted in the first step. The Facebook platform has been selected. All available offensive/hate speech Urdu datasets are developed by collecting data from Twitter and YouTube platforms but the Facebook platform is ignored. There is no restriction on the length of Facebook posts that's why it offers a rich source of publicly available data for research and development. Furthermore, there are no propriety concerns or limitations enforced by the Facebook platform. The Facebook Application Programming Interface (API) is available to scrap the data from the Facebook pages of Pakistan. Why we chose Pakistani Facebook pages because the Urdu language is spoken in Pakistan as the national language, therefore we have an opportunity of a large number of Urdu speakers.

To build a representative dataset, various newsgroups of Pakistan on Facebook are shortlisted and these newsgroups consist of popular political, religious, and blogger pages. The selected Facebook pages are diverse (not related to one community), which makes our corpus representative of Pakistani society. We follow a certain criterion to select an Urdu Facebook page, which is given below:

- The Facebook page should use Urdu language frequently for sharing posts and comments.
- The page should be followed and liked by at least thirty thousand users. This restriction enables us to select more active public pages on the Facebook community.

We applied the above criteria and the selection process ended with the nineteen public pages. The detail of the Facebook pages is described in **Table 2**. The posts and comments are collected for the period of 28 months ranging from February 01, 2021, to June 30, 2023. There was political instability in Pakistan in this time-period and there were protests demonstrated by various political parties. The data collection process led to 20,000 Facebook posts. After that, we need to apply cleaning and filtering steps to finalize the corpus for the annotation step.

### B. DATA CLEANING & FILTERING

The initial screening reveals that the crawled data contained garbage material, that needs to be filtered out to proceed further with the annotation step.

**TABLE 1.** Summary of related studies in roman and nastaliq urdu language.

| Year [ref] | Platform [Language] | Classes | Features | Techniques |
|---|---|---|---|---|
| 2017 [17] | Twitter [U] | Controversial, not-controversial | TF-IDF | SVM, LR, NB |
| 2020 [18] | News [U] | Propaganda, not-propaganda | LIWC, NELA, word2vec, BERT | CNN |
| 2020 [20] | Twitter [U] | Abusive, not abusive | Lexicon-based | Custom |
| 2020 [19] | YouTube [RU, U] | Offensive, not-offensive | Word and char n-grams | RF, SVM, NB, KNN |
| 2020 [21] | NF [U] | Sentiment Analysis | Lexicon-based | Customized Algorithm |
| 2020 [22] | Twitter [U] | Offensive, Sexism, Religious-hate, Profane, Normal | Embedding models (FastText, BERT) | CNN, LSTM, SVM |
| 2021 [13] | Twitter [U] | Hate, not-hate (sentiment analysis) | Bag of words | SVM, Multinomial NB |
| 2021 [25] | Twitter [U] | Threatening, not-threatening; individual, group | Word and char n-grams, FastText | SVM, MLP, LR, 1D-CNN, LSTM |
| 2021 [30] | Several [U] | Toxic, not-toxic | Word embedding models | CNN, Bi-LSTM, BGRU |
| 2021 [14] | Twitter [RU] | Hate, simple-complex, neutral | TF-IDF, word2vec | NB, LR, SVM, Bagging, Boosting |
| 2021 [26] | Twitter [U] | Abusive, not-abusive; threatening, not-threatening | BERT transformer | BERT classifier |
| 2022 [27] | Twitter [U] | Threatening, not-threatening | Char and word n-grams | CNN, LSTM, GRU, Ensemble Model |
| 2022 [32] | YouTube [RU, U] | Offensive, not-offensive | Word and char n-grams | NB, RF, LR, SVM |
| 2022 [31] | Facebook [U] | Offensive, not offensive | TF-IDF, word2vec | LR, RF, SVM, Ensemble |
| 2023 [6] | Twitter [U] | Threatening, not-threatening; individual, group | BERT, TF-IDF, word, and char n-grams | NB, LR, RF, BERT classifier |
| 2023 [15] | Twitter [U] | Hateful, not-hateful; Ethnic, Sectarian, Interfaith | Word embedding, word uni-gram | NB, RF, SVM, LR, CNN, LSTM |
| 2023 [28] | Twitter [U, E] | Threatening, not-threatening | RoBERTa, MuRIL models | Fine-tuning transformers |
| 2023 [16] | Twitter [U] | Offensive, hate speech, neutral; symbolization, insult, and attribution | Word and char n-grams, word2vec | LR, SVM, CNN, LSTM |
| 2024 [23] | Twitter [U] | Cyberbullying, neutral | TF-IDF, word2vec, FastText | SVM, NB, LR, LSTM |
| 2024 [24] | Multiple [E, RU] | Cyberbullying, neutral | N-grams, TF-IDF | NB, LR, SVM |
| 2024 [33] | Twitter [U] | Violence incitation | Word n-grams, BERT, RoBERTa | CNN, SVM, RF, bi-LSTM |
| 2024 [29] | Twitter [U] | Threatening, not threatening | Word n-gram, TF | LSTM based model |
| 2024 [34] | Twitter [U] | Abusive, neutral | Word2vec, FastText, TF-IDF | RF, LR, SVM, CNN, Bi-LSTM |
| 2024 [35] | YouTube [RU] | Coarse and fine-grained hate speech | Word embeddings | ULMFIT-LSTM |
| *Proposed | Facebook [U] | Hate, Not-hate; community detection | ---------- | Fine-tuning BERT, DistilBERT, RoBERTa, MuRIL |

RU: Roman Urdu; U: Urdu; E: English; NB: Naïve Bayes; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; KNN: K-nearest neighbor; CNN: Convolutional Neural Network; MLP: Multilayer Perceptron; LSTM: Long Short Term Memory; BGRU: Bilingual Gated Recurrent Unit; NF: Not Found; GRU: Gated Recurrent Unit; ULMFIT: Universal Language Model Fine-tunning

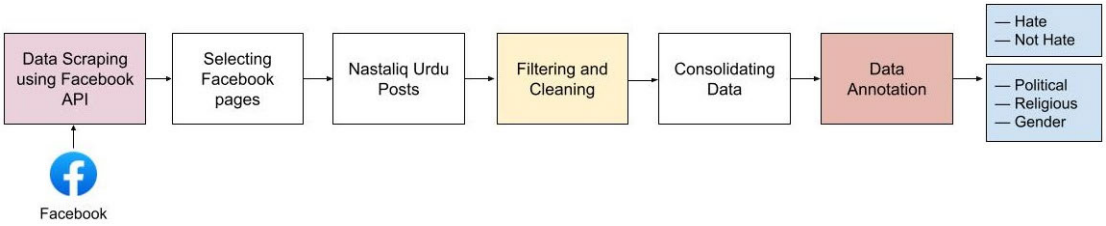



**FIGURE 1.** Overview of data collection process.

**TABLE 2.** Chosen facebook pages.

| S # | Facebook pages | S # | Facebook Pages |
|---|---|---|---|
| 1 | Dunya News | 11 | PTI Official |
| 2 | Samma TV | 12 | PPP Official |
| 3 | Express News | 13 | PMLN Official |
| 4 | Bol News | 14 | Pervaiz Hood |
| 5 | ARY News | 15 | JIP Official |
| 6 | GeoUrduNews | 16 | JUI PK |
| 7 | Urdu Point | 17 | Chingari |
| 8 | Daily Pakistan | 18 | marvisirmed |
| 9 | BBC Urdu | 19 | Liberal2016 |
| 10 | Zem Tv | | |

Therefore, the following cleaning steps are considered to clean the corpus:

1. The empty and/or duplicate posts are removed.
2. The hashtags, numbers, mentions, HTML tags, URLs, and punctuations are removed.
3. Translation of some English words/Abbreviations into Nastaliq Urdu.
4. Missing words/characters are filled in the sentences.

After applying the cleaning steps, we have 9771 Facebook posts/comments in the corpus. The demonstration of some cleaning steps is presented in **Table 3**. Now dataset is ready for the next step of preparation.

### C. DATA ANNOTATION

Here the process of data annotation is described. Three annotators are hired to perform this task. As our dataset needs categorization on two levels, therefore proper annotation protocol/guidelines are designed to guide the annotators in labeling. In the first step, annotators have to categorize the Facebook posts into hate speech or not-hate after understanding the context of the posts. Some examples are presented in **Table 4**. In the second step, the posts tagged with hate speech are further considered for the categorization of community, i.e. religious, political, or gender-based community. The description of both types of classification tasks is presented in **Fig. 2**. The posts which are tagged as not-hate, are not considered for the second level of categorization.

To select the qualified annotators, we made certain criteria to conclude the annotation process appropriately. The criteria are: 1) he/she must be a native Urdu speaker, 2) should have prior experience in data annotation, and 3) his/her education must be at least a master's degree. Thus, many annotators are interviewed and finally, three are selected.

### D. SPECIFICATION OF CORPUS

The prepared corpus consists of 9771 samples (Facebook posts) in total that are labeled in two levels. The first categorization consists of 5289 hate speech and 4682 not-hate instances. For the second level of annotation, only hate speech instances are considered. The corpus contains an equal percentage of instances for the second level, i.e. 33% gender, 33% political, and 34% religious-based community.

## IV. PROPOSED METHODOLOGY AND EXPERIMENTAL SETUP

This section presents the description of the proposed methodology. The workflow for the hate speech and target community detection framework is presented in **Fig. 2**. The sequence is as follows; First, the dataset is constructed, and the process is explained in section III. Then, the dataset is encountered with the pre-processing steps to prepare it for further processing. Before applying transformers, the dataset should be tokenized and normalized to represent it in a uniform format (representation step). After that, fine-tuning of two Urdu transformers is described with grid search technique. The first level of classification is binary and the second level of classification is multi-class. The classifiers are evaluated using five state-of-the-art metrics. The description of baselines is also presented. The tweets are classified into hate or not-hate in the first step. Then hateful tweets are further categorized based on the targeted community (political, religious, or gender based).

### A. PRE-PROCESSING

Online users usually use informal language to express their feelings and opinions on social media platforms. This information is in unstructured form; therefore, pre-processing is necessary before applying any ML/DL algorithm for learning/prediction. The following pre-processing steps are employed:

1. Urdu stop words are removed (not for transformer models).

**TABLE 3.** Cleaning steps demonstration on nastaliq urdu corpus.

| Urdu Text Preprocessing | |
|---|---|
| **1. Punctuation Removal** | |
| **Before** | **After** |
| مین, شیخ ٹلی پر لعنت کرتا ہون..... | مین شیخ ٹلی پر لعنت کرتا ہون |
| **2. Stop word Removal** | |
| **Before** | **After** |
| اس خبیث کو مفتی نہ کھنا | خبیث مفتی نہ کھنا |
| **3. Replace Emoji with Corresponding Text** | |
| **Before** | **After** |
| موجودہ دنیا کا سب سے بڑا منافق 😩 | موجودہ دنیا کا سب سے بڑا منافق ,ناراض چہرہ |
| **4. Hashtag, Numbers, and Link Removal** | |
| **Before** | **After** |
| بڑے یورپی ملک نے پاکستان کیلئے پہلی مرتبہ ارب کروڑ ڈالرز کے پیکج کا اعلان کر دیا https://reut.r/3P4qHlE | بڑے یورپی ملک نے پاکستان کیلئے پہلی مرتبہ ارب کروڑ ڈالرز کے پیکج کا اعلان کر دیا |
| #PAKvsSL پاک سری لنکا میچ اچانک روک دیا گیا | پاک سری لنکا میچ اچانک روک دیا گیا |
| میں خود سعودیہ رہا ہوں آج سے 15 سال پہلے وہ ہیں اچھا ہی نہیں سمجھتے تھے | میں خود سعودیہ رہا ہوں آج سے سال پہلے وہ ہیں اچھا ہی نہیں سمجھتے تھے |

**TABLE 4.** Guidelines to annotate the facebook posts.

| S# | Urdu | Translation | Level 1 | Level 2 |
|---|---|---|---|---|
| 01 | بھونک بھونک کے مر جاو گے مولوی ڈیزل | Cleric Diesel will die by barking. | Hate Speech | Religious |
| 02 | عمران خان زندہ باد | Long live Imran Khan | Not-hate | NA |
| 03 | بے شرم تم نے ملک کی عزت داؤ پر لگا دیا | Shamelessly, you put the honor of the country at stake | Hate Speech | Political |
| 04 | عورت کے نام پر دھبا ہے یہ کھسرا | She is a stain on the woman's name. A type of transgender | Hate Speech | Gender-based |

2. Emojis/emoticons are replaced by their respective text manually.
3. Abbreviations are replaced by their corresponding text.

After pre-processing, the corpus is ready for feature extraction, thus we explored several semantic and transformer models for extracting significant features.

### B. FEATURES FOR TEXT REPRESENTATION

In this section, we describe the details of two state-of-the-art Urdu transformer models with fine-tuning. These models are used in the proposed framework for hate speech and target community identification from Facebook posts. The motivation here is to utilize the strength of TL for detecting hate speech and target the community with benchmark performance. TL is a paradigm to utilize the knowledge gained from one task/dataset to another similar/related task [36]. Basically, it targets generalized improvement in another related setting. In contrast, traditional approaches design a new model for each task. TF techniques start with pre-trained models/networks and apply already learned knowledge from the source task to the target task. It supports diverse applications from solving data science tasks to training DL models. The advantages of TL are 1) reduces computational costs, 2) accommodates small dataset size, 3) supports generalizability, etc. Mainly, there are three types of transfer learning: Inductive, Unsupervised, and Transductive transfer. We are interested in using transfer learning (Inductive) with fine-tuning approach.

#### 1) URDU-RoBERTa

In 2019, researchers developed an extension of the BERT transformer, that is RoBERTa model [37]. The difference between these two is a few modifications to embedding tweaks and hyperparameters. Furthermore, RoBERTa dropped the objective of the next sentence pre-training and added training with larger learning rates. A variety of RoBERTa models are available such as XLM-RoBERTa, Urdu-RoBERTa, etc.

The XLM model is trained in 100 languages and the Urdu-RoBERTa has already demonstrated significant performance for various NLP tasks such as threatening content identification [6]. We are interested in exploring the potential

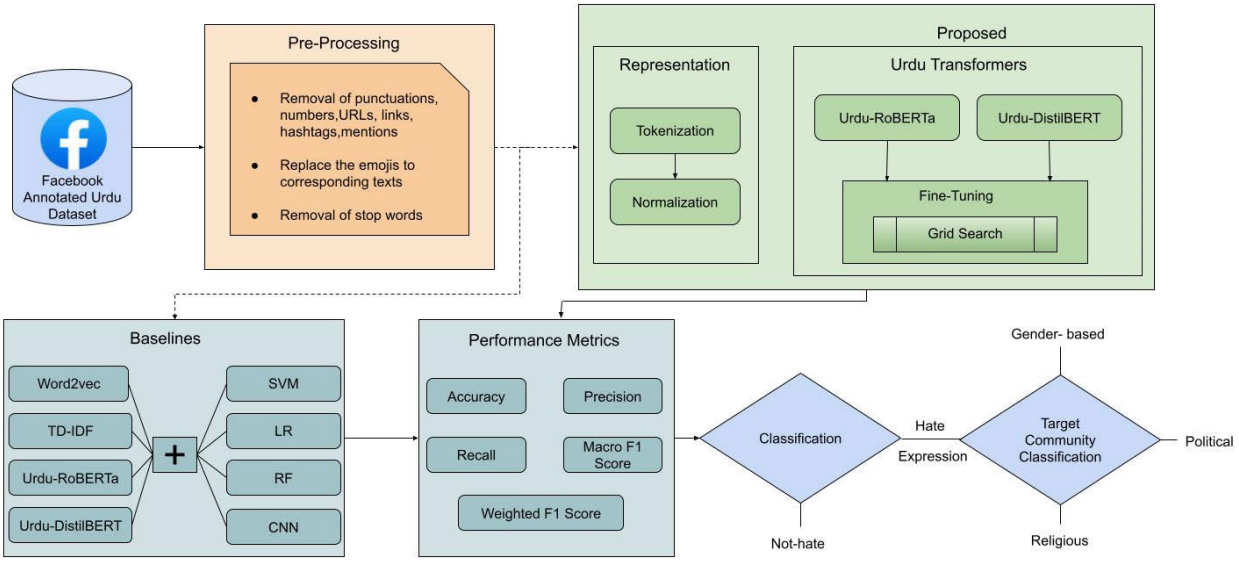



**FIGURE 2.** High-level HSTC architecture.

of the Urdu-RoBERTa transformer with fine-tuning for both levels of classification.

#### 2) URDU-DistilBERT

It is the distilled version of the popular Urdu-BERT base model. It is comprised of 768 dimensions, 6 layers, and 12 heads and is developed for next-sentence prediction and masked language modeling tasks. It is successfully utilized for NLP and downstream mining problems. However, we should be aware of its limitations and biases. In this study, we explored the strengths of Urdu-DistilBERT for hate speech and target community detection on Facebook platforms. The pre-trained language model is fine-tuned on eight hyper-parameters using a grid search technique.

### C. TOKENIZATION AND REPRESENTATION

Here, we provide the details of the tokenizer for relevant transformer models and the steps of data transformation into the required format. In the first step, the tokenization of Facebook posts is done in an appropriate uniform format so that it will be compatible with input to transformer models. The 'CLS' token is concatenated at the start of every post and the 'SEP' token at the end to make it clear from where each sentence is started. This process ends up in a uniform single vector against each Facebook post. Every language model has its tokenizer that can be used to tokenize the data. We used two tokenizers: one for Urdu-RoBERTa (https://huggingface.co/urduhack/roberta-urdu-small), and another for Urdu-DistilBERT (https://huggingface.co /Geotrend/distilbert-base-ur-cased). This tokenization process converted the Facebook post into various tokens. These tokens are then mapped to the corresponding indexes.

**TABLE 5.** Detail of hyper-parameters for fine-tunning.

| Hyper-parameters | Grid Search Methodology |
|---|---|
| Sequence length | 100, 128 |
| Batch size | 16, 32, 64 |
| Learning rate | 4e-5, 3e-5, 2e-5, 1e-5 |
| Weight decay | 0.01-0.1 |
| Warmup ratio | 0.06-0.1 |
| Hidden dropout | 0.05, 0.1 |
| Attention dropout | 0.05, 0.1 |
| Epochs | 1-10 |

### D. FINE-TUNNING AND CLASSIFICATION

The next step is fine-tuning both transformers using eight hyper-parameters for hierarchical classification of hateful content and target community identification in Nastaliq Urdu. The literature usually used two methods for fine-tuning, i.e. manual search and grid search for finding the best values for hyper-parameters. We tried the grid search methodology and got the best performance for both transformers. The names of hyper-parameters and their range of values are described in **Table 5**. The sequence lengths of 100 and 128 are considered with batch sizes of 16, 32, and 64 to analyze their impact on two levels of classification. Similarly, a range of learning rates are tried from 4e-5 to 1e-5 to validate their impact on the accuracy of the classification models. In addition, 10 epochs are tried for fine-tuning and other values of hyper-parameters are added in **Table 5**. The stratified sampling algorithm is used to split the HSTC corpus into three parts. First, a split of 80-20 is applied in which 20% is used for testing, and the remaining 80% is further split into 90-10. This 90% is actually used for training and 10% is used for validation of the trained model.

The descriptions of both transformers are already added in sections IV-B1 and IV-B2. Both transformers have 12 attention heads, 768 hidden-size, and 12 hidden-layers. For setup, each transformer is first trained and then validated. The training and validation loss, validation accuracy, and validation f1-score are calculated. Then that transformer is tested on the testing dataset and relevant measures are calculated. The best results are picked from 10 epochs.

### E. CATASTROPHIC FORGETTING AND OVERFITTING

Catastrophic forgetting is defined as the problem encountered by every transformer model while fine-tuning hyper-parameters, results in forgetting already learned knowledge [6], [28], [38]. This unfortunate situation is common when we perform TL. In this study, we tackled this issue by choosing an appropriate value of learning rate. By exploring a sizeable range, we concluded that small learning rates reveal desired performance and result in best convergence whereas large learning rates result in failures and poor convergence. We got the best performance on the 1e-5 learning rate.

The over-fitting and under-fitting obstacles are also common in the training of deep learning models. An over-fitting is caused by choosing so many epochs whereas an under-fitting is encountered by choosing very few epochs. We are interested in configuring an appropriate number of epochs to handle these issues. So, a range of values are tried and our analysis concluded that ten epochs are appropriate in getting optimal performance.

### F. EVALUATION MEASURES

Four standard evaluation measures are chosen to evaluate the performance of the proposed framework and state-of-the-art baselines. The measures are accuracy, precision, recall, macro, and weighted f1-scores. The HPC local cloud and Google Colab resources are utilized for performing various experiments. Python language is used for development purposes.

### G. BASELINES

To compare the strengths uncovered by the proposed framework for both levels of classification tasks, sixteen baseline models are chosen for the experimental setup. Four different kinds of feature generation models are combined with traditional ML models. The models are TF-IDF [39], word2vec [40], Urdu-RoBERTa, and Urdu-DistilBERT. Why we chose the last two because they are fine-tuned in the proposed framework and we want to test their performance in combination with traditional ML models. In this way, we can better compare them with fine-tuned versions. Furthermore, CNN, RF [41], SVM [42], and LR [43] are chosen as classifiers. These classifiers have proved their effectiveness in related NLP tasks [44], [45], [46], [47]. For the word2vec model, the skip-gram method and 100 dimensions are configured. The description of comparable models is given below:

1. Urdu-RoBERTa + CNN
2. Urdu-RoBERTa + RF
3. Urdu-RoBERTa + SVM
4. Urdu-RoBERTa + LR
5. Urdu-DistilBERT + CNN
6. Urdu-DistilBERT + RF
7. Urdu-DistilBERT + SVM
8. Urdu-DistilBERT + LR
9. Word2vec [100] + CNN
10. Word2vec [100] + RF
11. Word2vec [100] + SVM
12. Word2vec [100] + LR
13. TF-IDF + CNN
14. TF-IDF + RF
15. TF-IDF + SVM
16. TF-IDF + LR

## V. EXPERIMENTS AND RESULTS

This section presents the experiments performed to develop a classification framework for hate speech and target community identification for Nastaliq Urdu Facebook posts. In addition, details of experiments for baseline are added.

### A. HATE SPEECH CLASSIFICATION (LEVEL 1)

As our HSTC corpus is comprised of two levels of classification, so we first perform experiments for binary classification. The objective here is to classify the Nastaliq Urdu Facebook posts into hate or not-hate. In the next section, we will present experiments to classify the target community on the hateful posts. The communities are political, religious, and gender-based.

In this section, we conducted experiments to fine-tune two transformer models (RoBERTa and DistilBERT) for binary classification. After that, their performances are compared with sixteen baseline models to analyze the outperformance of the proposed framework.

#### 1) FINE-TUNNING

As described earlier, we explored the potential of two state-of-the-art transformer models for the first level of the classification (binary) task. The Urdu-RoBERTa and Urdu-DistilBERT are fine-tuned using grid search methodology and eight hyper-parameters are under observation. The detail of hyper-parameters is already described in **Table 5**. The grid search methodology is chosen to find out the appropriate values of hyper-parameters. The transformer models are trained, validated, and then tested. The splitting detail of the HSTC corpus is presented in section **4.2.4.** By exploring various values of hyper-parameters, we reported the best results with hyper-parameter values. For the fine-tuning phase, the steps are described in section **4.2.4**. The sequence length of 100 and 128 are explored and the best performance is observed with a sequence length of 100. The values of hyper-parameters and the resulting confusion matrix are also added to results as shown in **Table 6**. The fine-tunning phase of Urdu-RoBERTa produced the highest performance of 85.50% accuracy and

**TABLE 6.** Results of fine-tuned urdu-distilbert and urdu-roberta models on the test dataset (binary classification).

| **Urdu-RoBERTa** (Epochs = 10, epsilon = 1e-8, Weight decay = 0.01) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **S. L** | **B. S** | **L. R** | **H. D** | **W.R** | **TP** | **TN** | **FP** | **FN** | **ACC** | **F1-score** | | | |
| | | | | | | | | | | **Hate** | **N-hate** | **MAC** | **WTD** |
| 100 | 32 | 1e-5 | 0.01 | 0.01 | 927 | 782 | 159 | 152 | 84.60 | 85.64 | 83.41 | 84.52 | 84.60 |
| 100 | 32 | 2e-5 | 0.005 | 0.06 | 899 | 811 | 130 | 180 | 84.65 | 85.29 | 83.95 | 84.62 | 84.67 |
| 128 | 32 | 1e-5 | 0.005 | 0.06 | 939 | 777 | 164 | 140 | 84.95 | 86.07 | 83.64 | 84.85 | 84.94 |
| 100 | 16 | 1e-5 | 0.05 | 0.06 | 945 | 780 | 161 | 134 | 85.40 | 86.50 | 84.10 | 85.29 | 85.38 |
| 100 | 32 | 1e-5 | 0.05 | 0.06 | 944 | 783 | 158 | 135 | 85.50 | 86.57 | 84.24 | 85.40 | 85.48 |
| | | | | | | | | | | | | | |
| **Urdu-DistilBERT** (Epochs = 10, epsilon = 1e-8, Weight decay = 0.01) | | | | | | | | | | | | | |
| 100 | 32 | 1e-5 | 0.01 | 0.01 | 957 | 772 | 169 | 122 | 85.59 | 86.80 | 84.14 | 85.47 | 85.56 |
| 100 | 16 | 3e-5 | 0.05 | 0.06 | 931 | 801 | 140 | 148 | 85.74 | 86.60 | 84.76 | 85.68 | 85.75 |
| 100 | 64 | 3e-5 | 0.05 | 0.06 | 934 | 802 | 139 | 145 | 85.94 | 86.80 | 84.96 | 85.88 | 85.94 |
| 100 | 32 | 2e-5 | 0.05 | 0.06 | 959 | 778 | 163 | 120 | 85.99 | 87.14 | 84.61 | 85.87 | 85.96 |
| 100 | 32 | 1e-5 | 0.05 | 0.06 | 949 | 800 | 141 | 130 | 86.58 | 87.51 | 85.52 | 86.52 | 86.58 |
| B.S: Batch Size; L.R: Learning Rate; H.D: Hidden Decay; W.R: Warmup Ratio; S.L: Sequence Length; ACC: Accuracy; N-hate: Not Hare; MAC: Macro Averaged; WTD: Weighted Averaged; | | | | | | | | | | | | | |

85.40% macro f1-score with a batch size of 32, hidden dropout of 0.05, warmup-ratio of 0.06, and learning rate of 1e-5. However, the lowest performance is reported for a hidden dropout of 0.01 and a warmup ratio of 0.01.

The performance of fine-tuning Urdu-DistilBERT is reported in the lower part of **Table 6.** The outperformance is observed with a learning rate of 1e-5, batch size of 32, warmup ratio of 0.06, and hidden dropout of 0.05. It is now clearly demonstrated that a lower learning rate results in best convergence and efficiently handles the issue of catastrophic forgetting. In contrast, when a higher learning rate is applied, we get a degradation in performance that is the result of poor convergence. The issue of overfitting is handled by using 10 epochs for training, validation, and testing phases of fine-tuning transformers. It is observed that the fine-tuning phase with epochs greater than 10 got overfitting the classification results, therefore, we reported the best results conducted up to 10 epochs. On top of everything, the highest performance is demonstrated by Urdu-DistilBERT and achieved an accuracy of 86.58% and a macro f1-score of 86.52%. Thus Urdu-DistilBERT outperformed the Urdu-RoBERTa while fine-tuning both models. This completes the phase of fine-tuning both transformer models using grid search methodology.

### 2) COMPARISON WITH STATE-OF-THE-ART

This section describes the comparison and analysis of proposed fine-tuned Urdu-RoBERTa and Urdu-DistilBERT with sixteen SOTA comparable models for binary classification. Five standard evaluation measures are used to test the effectiveness of all classification models and results are provided in **Table 7** and **Fig. 3**. Among the SOTA comparable models, TF-IDF + LR demonstrated comparatively better performance compared to others and showed 84.55% accuracy and 84.54% f1-score. On the other hand, the worst performance is shown by the TF-IDF + CNN model (accuracy of 55.27% and f1-score of 50.70%). Considering the proposed framework, fine-tuning Urdu-RoBERTa achieved an accuracy of 85.50% and an f1-score of 85.40%, thus performing better than all SOTA models.

In addition, the performance of fine-tuning Urdu-RoBERTa is better than all SOTA in precision and recall measures. However, the best performance is demonstrated by fine-tuning Urdu-DistilBERT. It outperformed all SOTA models including fine-tuned Urdu-RoBERTa by achieving benchmark values of accuracy (86.58%), and macro f1-score (86.52%) as shown in **Table 7**. The values of precision and recall obtained by fine-tuned Urdu-DistilBERT are also highest. This verified the significance of our proposed framework for the binary classification of Facebook posts into hate and no-hate classes. The proposed framework improved the accuracy by 2.03%, and macro f1-score by 1.98% compared to sixteen SOTA models.

Next, the comparison of the proposed framework (fine-tuned Urdu-RoBERTa and Urdu-DistilBERT) with SOTA models is conducted on the basis of class-wise, weighted and macro-average measures. The f1-score is used to evaluate the performances and results are shown in **Fig. 3**. For hate speech classification, it is evident that fine-tuned Urdu-DistilBERT model achieved the highest f1-score (87.51%) compared to all SOTA including fine-tuned Urdu-RoBERTa models. This shows the strength of fine-tuned DistilBERT for identifying hate speech class instances on the Facebook platform. The TF-IDF + SVM model presented 85.09% performance which is better than other SOTA models. The proposed model demonstrated an improvement of 2.42% in identifying hate speech instances. Similarly, fine-tuned Urdu-DistilBERT presented benchmark performance in identifying not-hate class instances and achieved an 85.52% threshold. An improvement of 1.41% is observed for not-hate class identification.

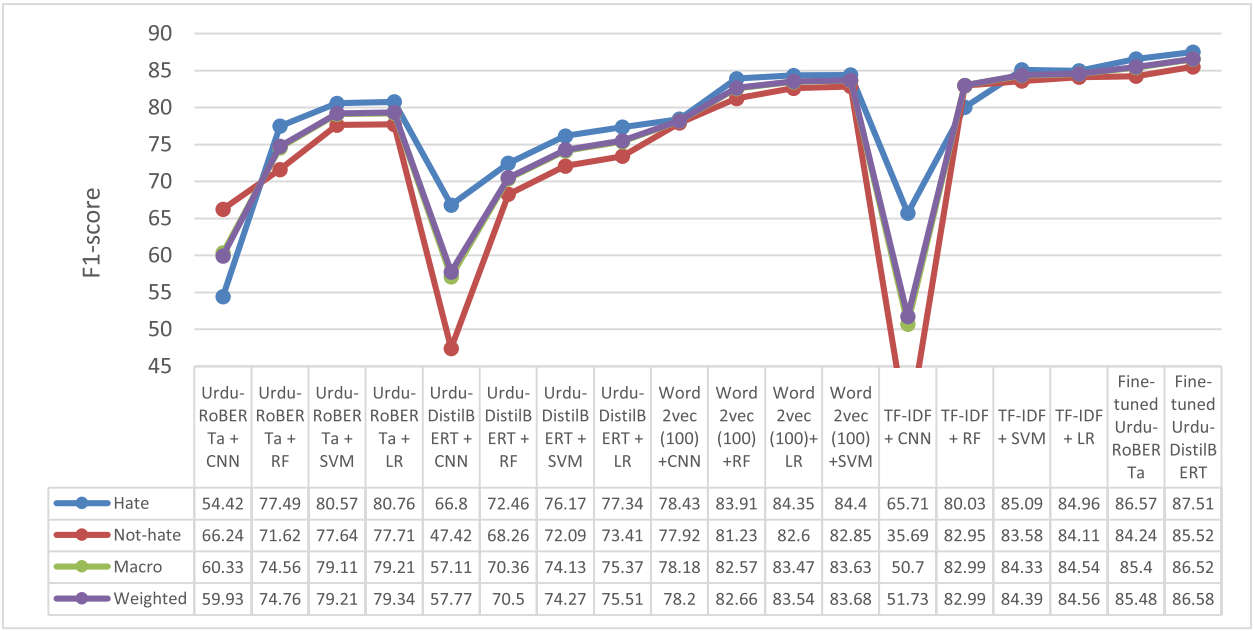



**FIGURE 3.** Comparison of proposed and SOTA models based on class-wise and average measure (binary classification).

**TABLE 7.** Comparison of proposed fine-tuned language models with sixteen sota models (binary classification).

| Type | Models | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Baseline | Urdu-RoBERTa + CNN | 61.21 | 64.40 | 62.52 | 60.33 |
| | Urdu- RoBERTa + RF | 74.89 | 74.99 | 74.99 | 74.56 |
| | Urdu-RoBERTa + SVM | 79.21 | 79.11 | 79.10 | 79.11 |
| | Urdu-RoBERTa + LR | 79.35 | 79.26 | 79.22 | 79.24 |
| | Urdu-DistilBERT + CNN | 59.30 | 59.36 | 58.02 | 57.11 |
| | Urdu-DistilBERT + RF | 70.51 | 70.37 | 70.36 | 70.36 |
| | Urdu-DistilBERT + SVM | 74.29 | 74.18 | 74.10 | 74.13 |
| | Urdu-DistilBERT + LR | 75.53 | 75.43 | 75.34 | 75.37 |
| | Word2vec (100) +CNN | 78.18 | 78.40 | 78.47 | 78.18 |
| | Word2vec (100) +RF | 82.67 | 82.62 | 82.53 | 82.57 |
| | Word2vec (100)+LR | 83.52 | 83.44 | 83.54 | 83.47 |
| | Word2vec (100) +SVM | 83.66 | 83.60 | 83.73 | 83.63 |
| | TF-IDF + CNN | 55.27 | 54.82 | 53.43 | 50.70 |
| | TF-IDF + RF | 82.99 | 83.34 | 83.37 | 82.99 |
| | TF-IDF + SVM | 84.37 | 84.30 | 84.44 | 84.33 |
| | TF-IDF + LR | 84.55 | 84.60 | 84.76 | 84.54 |
| Proposed | Fine-tuned Urdu-RoBERTa | 85.50 | 85.48 | 85.35 | 85.40 |
| | Fine-tuned Urdu-DistilBERT | **86.58** | **86.54** | **86.48** | **86.52** |

Moreover, calculating the weighted average, TF-IDF + LR showed a performance of 84.56% which is better than other SOTA models. Again, fine-tuned DistilBERT outperformed and demonstrated 86.58% performance. This resulted in an improvement of 2.02%. These experiments are sufficient to establish the effectiveness of the proposed framework for binary classification tasks. Thus, our proposed solution based on fine-tuned Urdu-DistilBERT proved its significance.

### *B. TARGET COMMUNITY DETECTION (LEVEL 2)*

This section presents the experiments performed to identify the target community in hate speech instances. It is a multi-class classification problem and we have three communities, i.e. political, religious, and gender-based. For this purpose, two transformer models are fine-tuned using grid search methodology, and their performances are compared with sixteen SOTA models.

#### 1) FINE-TUNNING

As described earlier, the fine-tuning of two transformer models is performed for target community identification considering only hate speech instances. The grid search methodology is used to find optimal values of eight hyper-parameters. Ten epochs are applied to handle the issue of over-fitting and both transformers are trained, validated, and tested on the corresponding parts of the dataset. The testing results of both transformers are added in **Table 8**. We explored various values of each hyper-parameter (described in **Table 5**) but the best results are reported against each combination of hyper-parameter values. For Urdu-RoBERTa, fine-tuning is performed by exploring sequence lengths of 100 and 128, and the best performance is obtained with a sequence length of 100. We have reported the three best performances with specific hyper-parameter values in the upper part of **Table 8**. The highest performance (accuracy of 81.30% and macro f1-score of 81.42%) is achieved with the following hyper-parameter values (batch size: 32, learning rate: 1e-5, hidden dropout: 0.05, and warmup ratio: 0.01).

The fine-tuning of Urdu-DistilBERT is performed by adopting the same process and results are added in the lower part of **Table 8**. Two sequence lengths were tried and we obtained the best performance with a sequence length of 100. After applying several combinations of hyper-parameters, three demonstrated better performances (added in the Table). The batch size of 64 did not produce better performance, however 32 batch size demonstrated benchmark performance. The highest performance is achieved with a learning rate of 1e-5, a hidden dropout of 0.05, and a warmup ratio of 0.01. This completes the fine-tuning process of two transformer models for the task of community detection.

#### 2) COMPARISON WITH STATE-OF-THE-ART

In this section, the performance of fine-tuned Urdu-RoBERTa and Urdu-DistilBERT is compared with sixteen SOTA models to identify the target community on hateful Facebook posts. Four standard measures are used to evaluate the performances and results are added in **Table 9**. Considering sixteen SOTA models, TF-IDF + SVM achieved the highest accuracy (80.58%), and highest macro f1-score (80.51%) whereas DistilBERT + RF showed the lowest performance with the following accuracy (63.31%) and macro f1-score (62.79%). However, the fine-tuned Urdu-RoBERTa model presented better results by achieving an accuracy of 81.30% and f1-score of 81.42% and outperformed all SOTA models. In addition, the precision and recall measures are also the highest compared to the sixteen SOTA models. Likewise, the fine-tuned Urdu-DistilBERT presented benchmark performance and outperformed all models including fine-tuned Urdu-RoBERTa. An accuracy of 84.17% and f1-score of 83.91% is achieved. This resulted in an improvement of 3.59% in accuracy and 3.4% in f1-score. Thus, our proposed framework has proved to be a benchmark model for target community identification for Nastaliq Urdu Facebook posts.

Lastly, the performance of the proposed framework and sixteen benchmarks are compared for class-wise community detection and weighted f1-score. The results are presented in **Fig. 4** and the best twelve out of sixteen SOTA models are added because of their performances. For political community detection, fine-tuned Urdu-DistilBERT demonstrated the highest performance and achieved a 79.15% f1-score. It outperformed all SOTA models including fine-tuned Urdu-RoBERTa and showed an improvement of 4.69%. For gender community identification, again fine-tuned DistilBERT outperformed and achieved 84.21% f1-score. This time, it improved the performance by 1.07%. For religious community detection, the highest performance is achieved by fine-tuning DistilBERT (83.93%). Again, it outperformed all SOTA models including fine-tuned Urdu-RoBERTa, and got an improvement of 4.01% compared to the best SOTA model. Thus, the proposed framework proved its significance for target community detection. The comprehensive set of experiments now established the effectiveness of the proposed framework for binary and multi-class classification tasks of hate speech identification in Nastaliq Urdu on Facebook posts.

## VI. DISCUSSION AND LIMITATIONS

Nowadays, social media platforms have become one of the main sources of information for people worldwide. These platforms support information dissemination and allow individuals to share their opinions on any event/incident. The opinions of people can be classified into positive, negative, or neutral. The negative comments often lead to harm and threats resulting in unrest in society, hate speech is one of them. Hate speech on the basis of religious/political/social/ethnic conflicts, and ethnic violence is an ongoing issue worldwide, especially in Pakistan. Since Urdu is the national language of Pakistan, people use it on social media platforms to spread hateful content to others. Several studies have addressed the issue of hate speech detection in English, however, no study has been conducted to detect hateful text and target communities for Nastaliq Urdu on Facebook posts.

This article provides a complete process of detecting coarse and fine-grained (binary and multi-class) hate speech in Nastaliq Urdu on Facebook posts. The mechanism is based on benchmark performance metrics that aid in preventing hate crimes and offer a reliable system for analyzing social media comments. The use of the proposed framework helped researchers and users to spot the main characteristics that shape the assignment of a comment to a hate or neutral category. This milestone is achieved by exploiting the strength of transfer learning to design an automated solution. Specifically, two Urdu transformers (Urdu-RoBERTa, and Urdu-DistilBERT) are explored with fine-tuning. The grid search technique made it possible to find optimal values of hyper-parameters. The Urdu-DistilBERT demonstrated benchmark performance on all applied metrics for coarse-grained and fine-grained hate speech identification

**TABLE 8.** Fine-tunning urdu-distilbert and urdu-roberta using grid search (multi-class classification).

| **Urdu-RoBERTa** (Epochs = 10, epsilon = 1e-8, Weight decay = 0.01) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **S. L** | **B. S** | **L. R** | **H. D** | **W.R** | **Accuracy** | **F1-score** | | | | |
| | | | | | | **Political** | **Gender** | **Religious** | **MAC** | **WTD** |
| 100 | 64 | 1e-6 | 0.05 | 0.01 | 79.13 | 74.46 | 80.41 | 82.76 | 79.21 | 79.23 |
| 100 | 32 | 1e-6 | 0.05 | 0.01 | 80.57 | 76.08 | 78.35 | 87.64 | 80.70 | 80.74 |
| 100 | 32 | 1e-5 | 0.05 | 0.01 | 81.30 | 78.43 | 81.40 | 81.45 | 81.42 | 81.44 |
| | | | | | | | | | | |
| **Urdu-DistilBERT** (Epochs = 10, epsilon = 1e-8, Weight decay = 0.01) | | | | | | | | | | |
| 100 | 32 | 1e-4 | 0.01 | 0.01 | 80.57 | 72.42 | 83.67 | 83.0 | 80.35 | 80.37 |
| 100 | 64 | 1e-5 | 0.05 | 0.06 | 83.45 | 79.15 | 81.32 | 88.46 | 83.10 | 83.13 |
| 100 | 32 | 1e-5 | 0.05 | 0.01 | 84.17 | 79.15 | 84.21 | 88.0 | 83.91 | 83.93 |

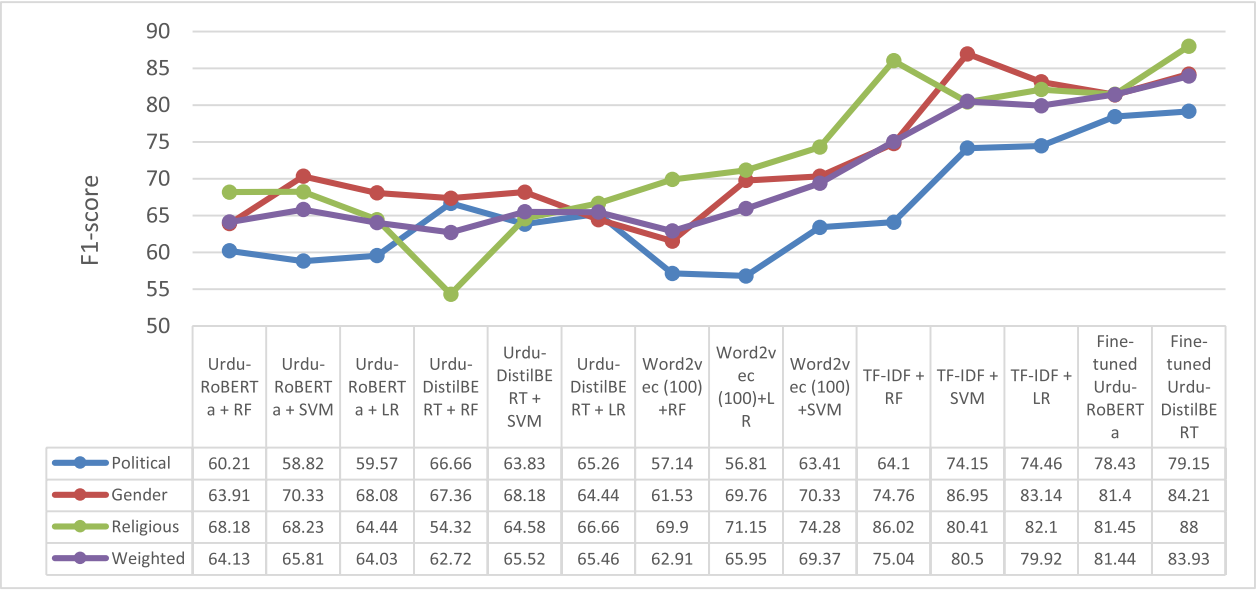
**TABLE 9.** Comparison of fine-tuned language models with sixteen sota models (multi-class classification).

| **Type** | **Models** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|---|---|
| Baseline | Urdu-RoBERTa + CNN | 64.92 | 64.56 | 64.29 | 65.37 |
| | Urdu- RoBERTa + RF | 64.03 | 64.51 | 64.03 | 64.10 |
| | Urdu-RoBERTa + SVM | 65.47 | 65.73 | 65.46 | 65.53 |
| | Urdu-RoBERTa + LR | 64.03 | 64.15 | 64.05 | 64.03 |
| | Urdu-DistilBERT + CNN | 64.45 | 64.80 | 64.45 | 64.67 |
| | Urdu-DistilBERT + RF | 63.31 | 63.58 | 63.43 | 62.79 |
| | Urdu-DistilBERT + SVM | 65.47 | 65.73 | 65.46 | 65.53 |
| | Urdu-DistilBERT + LR | 65.47 | 65.52 | 65.46 | 65.46 |
| | Word2vec (100) + CNN | 65.73 | 65.96 | 65.72 | 66.06 |
| | Word2vec (100) + RF | 63.31 | 63.22 | 63.21 | 62.86 |
| | Word2vec (100) + LR | 66.19 | 66.48 | 66.10 | 65.91 |
| | Word2vec (100) + SVM | 69.78 | 70.19 | 69.69 | 69.34 |
| | TF-IDF + CNN | 79.97 | 80.12 | 79.11 | 80.07 |
| | TF-IDF + RF | 75.54 | 76.89 | 75.47 | 74.96 |
| | TF-IDF + SVM | 80.58 | 80.57 | 80.56 | 80.51 |
| | TF-IDF + LR | 79.86 | 80.07 | 79.83 | 79.91 |
| Proposed | Fine-tuned Urdu-RoBERTa | 81.30 | 82.43 | 81.29 | 81.42 |
| | Fine-tuned Urdu-DistilBERT | **84.17** | **85.36** | **84.10** | **83.91** |

at the comment level. Furthermore, the performance of the proposed framework is superior to sixteen SOTA models. Thus the proposed solution provides a reliable ways of detecting hateful content in Nastaliq Urdu on social media at coarse-grained and fine-grained levels.

This research has several implications in the related domains. The prior works in hate speech detection have produced distinctive solutions and worthy but diverse findings, sometimes drawing inconsistent conclusions. The proposed system used a transfer learning paradigm to design an automated model by exploiting the potentials of Urdu-DistilBERT and Urdu-RoBERTa with fine-tuning. The proposed model outperformed the sixteen SOTA baselines. The findings of the current study provide insights for law enforcement organizations in the early detection of hateful content in Nastaliq Urdu and then filter out unwanted material from social media to promote peace and harmony in society. Moreover, the findings of the study help to save the vulnerable community by offering timely identification of the community exposed to be a target. These measures are very helpful for social media and are in high demand to eliminate unrest. Lastly, the findings of the proposed study provide grounds for a multi-model system to detect hateful content and then detoxify the text to promote positivity.

Highlighting the advantages of the proposed framework, we summarize following: First, the proposed methodology used the transformers to generate automated features instead of using hand-crafted features. This reduces the manual efforts required to calculate the hand-crafted features. Second, transformers are better than hand-crafted features to capture the context of hate speech in Nastaliq Urdu and to handle the complexity of the language for identification of hate speech. Prior studies mostly used frequency and semantic features for this task. Third, the proposed framework improved the accuracy of identifying hate speech in Nastaliq Urdu compared to SOTA baselines. Moreover, offers




| | Urdu-RoBERTa + RF | Urdu-RoBERTa + SVM | Urdu-RoBERTa + LR | Urdu-DistilBERT + RF | Urdu-DistilBERT + SVM | Urdu-DistilBERT + LR | Word2vec (100) +RF | Word2vec (100)+LR | Word2vec (100) +SVM | TF-IDF + RF | TF-IDF + SVM | TF-IDF + LR | Fine-tuned Urdu-RoBERTa | Fine-tuned Urdu-DistilBERT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Political | 60.21 | 58.82 | 59.57 | 66.66 | 63.83 | 65.26 | 57.14 | 56.81 | 63.41 | 64.1 | 74.15 | 74.46 | 78.43 | 79.15 |
| Gender | 63.91 | 70.33 | 68.08 | 67.36 | 68.18 | 64.44 | 61.53 | 69.76 | 70.33 | 74.76 | 86.95 | 83.14 | 81.4 | 84.21 |
| Religious | 68.18 | 68.23 | 64.44 | 54.32 | 64.58 | 66.66 | 69.9 | 71.15 | 74.28 | 86.02 | 80.41 | 82.1 | 81.45 | 88 |
| Weighted | 64.13 | 65.81 | 64.03 | 62.72 | 65.52 | 65.46 | 62.91 | 65.95 | 69.37 | 75.04 | 80.5 | 79.92 | 81.44 | 83.93 |

**FIGURE 4.** Comparison of proposed and SOTA models [class-wise and average measure for multi-class classification].

a mechanism for target community identification on the basis of religious. political, and gender based conflicts.

This study has some limitations. First, the dataset was collected from the Facebook platform, which has different dynamics compared to Twitter, YouTube, and other platforms. Furthermore, the comments represent the opinions of the Pakistani Community, further limiting the dataset scope, although Urdu is being spoken in other parts of Asia, Europe, and some regions of Canada. So the results obtained cannot be generalized beyond the intended scope. Second, the size of the dataset is not enough to reach some solid conclusions in general, thus a big corpus is needed. Third, the proposed framework did not support the interpretability of the classification process at coarse and fine-grained hate speech identification, thus leaving the explainability of black-box logic unattended. Future studies can address this issue to design a robust interpretable detection model for hate speech in Nastaliq Urdu. This extension will enable the researchers to address new diverse scenarios by exploring state-of-the-art visualization techniques.

## VII. CONCLUSION AND FUTURE SCOPE

This study investigated the task of identifying hateful content and vulnerable target community detection in Nastaliq Urdu Facebook posts. To the best of our knowledge, it is the first attempt in this area to offer a two level classification system for hateful content (coarse-grained) and target community (fine-grained) detection in Nastaliq Urdu. In addition, no labeled corpus is available for this task in Nastaliq Urdu.

This study contributes to the literature in three ways: First, the construction of labeled Nastaliq Urdu corpus for HSTC detection on Facebook posts. Second, it offers a robust two-level classification model for hateful content. Third, the utilization of two state-of-the-art Urdu transformers with fine-tuning instead of hand-craft features. Several pre-processing steps are applied to the HSTC corpus for preparation. After that, Urdu-RoBERTa and Urdu-DistilBERT transformers are fine-tuned using grid search technique for detecting hate speech and then target communities (religious, political, and gender-based) on hateful text. The experiments revealed that grid search is an effective technique for finding optimal hyper-parameters values. Sixteen baselines are chosen for SOTA comparisons. The results of extensive experiments revealed that fine-tuned DistilBERT is the most effective model for binary classification and fine-grained multi-class classification of hate speech in Nastaliq Urdu. It outperformed all SOTA models including fine-tuned Urdu-RoBERTa with substantial improvement in accuracy, weighted and macro f1-scores. The Urdu-DistilBERT achieved 86.52% f1-score for binary (coarse-grained) classification and 83.91% f1-score for target community (fine-grained) identification.

Several future directions can be considered to the extension of this study. One direction is to design an interpretable identification system for hate speech in Nastaliq Urdu that will interpret the black-box logic of classification inference. Another direction is to extend the current framework for other under-resource languages such as Hindi, Bengali, Greek, etc. Another direction is to transform the supervised approach into semi-supervised or un-supervised approach for hate speech and target community identification in Nastaliq Urdu because designing a labeled corpus requires manual efforts. Another direction is to utilize evolutionary algorithms with deep learning models to design a more robust identification system.

## ACKNOWLEDGMENT

This study was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number

(PNURSP2024R104), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. This article is the output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University). Moreover, this research was supported in part by computational resources of HPC facilities at HSE University.

## REFERENCES

[1] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: A systematic review," *Lang. Resour. Eval.*, vol. 55, no. 2, pp. 477–523, Jun. 2021.

[2] R. Delgado and J. Stefancic, "Images of the outsider in American law and culture: Can free expression remedy systemic social ills," *Cornell L. Rev.*, vol. 77, p. 1258, Feb. 1991.

[3] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–30, Jul. 2019.

[4] YouTube. (2019). *YouTube Hate Policy*. [Online]. Available: https://support.google.com/youtube/answer/2801939?hl=en

[5] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop Natural Lang. Process. Social Media*, 2017, pp. 1–10.

[6] M. S. I. Malik, "Threatening expression and target identification in under-resource languages using NLP techniques," in *Proc. Int. Conf. Anal. Images, Social Netw. Texts*. Cham, Switzerland: Springer, 2023, pp. 3–17.

[7] P. Sheth, R. Moraffah, T. S. Kumarage, A. Chadha, and H. Liu, "Causality guided disentanglement for cross-platform hate speech detection," in *Proc. 17th ACM Int. Conf. Web Search Data Mining*, Mar. 2024, pp. 626–635.

[8] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, Aug. 2023, Art. no. 126232.

[9] N. Makouar, L. Devine, and S. Parker, "Legislating to control online hate speech: A corpus-assisted semantic analysis of French parliamentary debates," *Int. J. Semiotics Law-Revue Internationale de Sémiotique Juridique*, vol. 36, no. 6, pp. 2323–2353, Dec. 2023.

[10] R. M. Al-Ibrahim, M. Z. Ali, and H. M. Najadat, "Detection of hateful social media content for Arabic language," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 9, pp. 1–26, Sep. 2023.

[11] M. Bilal, A. Khan, S. Jan, S. Musa, and S. Ali, "Roman Urdu hate speech detection using transformer-based model for cyber security applications," *Sensors*, vol. 23, no. 8, p. 3909, Apr. 2023.

[12] M. U. Arshad, R. Ali, M. O. Beg, and W. Shahzad, "UHated: Hate speech detection in Urdu language using transfer learning," *Lang. Resour. Eval.*, vol. 57, no. 2, pp. 713–732, Jun. 2023.

[13] M. Z. Ali, Ehsan-Ul-Haq, S. Rauf, K. Javed, and S. Hussain, "Improving hate speech detection of Urdu tweets using sentiment analysis," *IEEE Access*, vol. 9, pp. 84296–84305, 2021.

[14] M. M. Khan, K. Shahzad, and M. K. Malik, "Hate speech detection in Roman Urdu," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 1, pp. 1–19, Jan. 2021.

[15] M. H. Akram, K. Shahzad, and M. Bashir, "ISE-Hate: A benchmark corpus for inter-faith, sectarian, and ethnic hatred detection on social media in Urdu," *Inf. Process. Manage.*, vol. 60, no. 3, May 2023, Art. no. 103270.

[16] R. Saeed, H. Afzal, S. A. Rauf, and N. Iltaf, "Detection of offensive language and ITS severity for low resource language," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 6, pp. 1–27, Jun. 2023.

[17] 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.

[18] S. Kausar, B. Tahir, and M. A. Mehmood, "ProSOUL: A framework to identify propaganda from online Urdu content," *IEEE Access*, vol. 8, pp. 186039–186054, 2020.

[19] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, and M. T. Sadiq, "Automatic detection of offensive language for Urdu and Roman Urdu," *IEEE Access*, vol. 8, pp. 91213–91226, 2020.

[20] 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, Nov. 2020.

[21] M. Sohail, A. Imran, H. Ur Rehman, and M. Salman, "Anti-social behavior detection in Urdu language posts of social media," in *Proc. 3rd Int. Conf. Comput., Math. Eng. Technol. (iCoMET)*, Jan. 2020, pp. 1–7.

[22] H. Rizwan, M. H. Shakeel, and A. Karim, "Hate-speech and offensive language detection in Roman Urdu," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 2512–2522.

[23] F. Adeeba, M. I. Yousuf, I. Anwer, S. U. Tariq, A. Ashfaq, and M. Naqeeb, "Addressing cyberbullying in Urdu tweets: A comprehensive dataset and detection system," *PeerJ Comput. Sci.*, vol. 10, p. e1963, Apr. 2024.

[24] M. T. Jahangir, M. Ahmad, and H. Rehman, "Efficient intelligent system for cyberbullying detection in English and Roman Urdu social media posts," *J. Comput. Biomed. Inform.*, Apr. 2024.

[25] 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. 128302–128313, 2021.

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

[27] 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," *Appl. Sci.*, vol. 12, no. 20, p. 10342, Oct. 2022.

[28] M. Rehan, M. S. I. Malik, and M. M. Jamjoom, "Fine-tuning transformer models using transfer learning for multilingual threatening text identification," *IEEE Access*, vol. 11, pp. 106503–106515, 2023.

[29] A. Ullah, K. U. Khan, A. Khan, S. T. Bakhsh, A. U. Rahman, S. Akbar, and B. Saqia, "Threatening language detection from Urdu data with deep sequential model," *PLoS ONE*, vol. 19, no. 6, Jun. 2024, Art. no. e0290915.

[30] H. H. Saeed, M. H. Ashraf, F. Kamiran, A. Karim, and T. Calders, "Roman Urdu toxic comment classification," *Lang. Resour. Eval.*, vol. 55, no. 4, pp. 971–996, Dec. 2021.

[31] S. Hussain, M. S. I. Malik, and N. Masood, "Identification of offensive language in Urdu using semantic and embedding models," *PeerJ Comput. Sci.*, vol. 8, p. e1169, Dec. 2022.

[32] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. AbdelMajeed, and T. Zia, "Abusive language detection from social media comments using conventional machine learning and deep learning approaches," *Multimedia Syst.*, vol. 28, no. 6, pp. 1925–1940, Dec. 2022.

[33] M. S. Khan, M. S. I. Malik, and A. Nadeem, "Detection of violence incitation expressions in Urdu tweets using convolutional neural network," *Expert Syst. Appl.*, vol. 245, Jul. 2024, Art. no. 123174.

[34] A. Khan, A. Ahmed, S. Jan, M. Bilal, and M. F. Zuhairi, "Abusive language detection in Urdu text: Leveraging deep learning and attention mechanism," *IEEE Access*, vol. 12, pp. 37418–37431, 2024.

[35] F. Mehmood, H. Ghafoor, M. N. Asim, M. U. Ghani, W. Mahmood, and A. Dengel, "Passion-Net: A robust precise and explainable predictor for hate speech detection in Roman Urdu text," *Neural Comput. Appl.*, vol. 36, no. 6, pp. 3077–3100, Feb. 2024.

[36] M. S. I. Malik, M. Z. Younas, M. M. Jamjoom, and D. I. Ignatov, "Categorization of tweets for damages: Infrastructure and human damage assessment using fine-tuned BERT model," *PeerJ Comput. Sci.*, vol. 10, p. e1859, Feb. 2024.

[37] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[38] 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," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 8, Sep. 2023, Art. no. 101736.

[39] 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 Syst. Appl.*, vol. 227, Oct. 2023, Art. no. 120236.

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

[41] M. S. I. Malik, A. Nawaz, M. M. Jamjoom, and D. I. Ignatov, "Effectiveness of ELMo embeddings, and semantic models in predicting review helpfulness," *Intell. Data Anal.*, vol. 28, no. 4, pp. 1045–1065, Jul. 2024.

[42] M. S. I. Malik, F. Rehman, and D. I. Ignatov, "Ensemble learning with linguistic, summary language and psychological features for location prediction," *Int. J. Inf. Technol.*, vol. 16, no. 1, pp. 193–205, Jan. 2024.

[43] Y. Abbas and M. S. I. Malik, "Defective products identification framework using online reviews," *Electron. Commerce Res.*, vol. 23, no. 2, pp. 899–920, Jun. 2023.
[44] M. S. I. Malik and A. Nawaz, "SEHP: Stacking-based ensemble learning on novel features for review helpfulness prediction," *Knowl. Inf. Syst.*, vol. 66, no. 1, pp. 653–679, Jan. 2024.
[45] G. Ali and M. S. I. Malik, "Rumour identification on Twitter as a function of novel textual and language-context features," *Multimedia Tools Appl.*, vol. 82, no. 5, pp. 7017–7038, Feb. 2023.
[46] A. Nawaz and M. Malik, "Rising stars prediction in reviewer network," *Electron. Commerce Res.*, vol. 22, no. 1, pp. 53–75, Mar. 2022.
[47] A. Mehboob and M. S. I. Malik, "Smart fraud detection framework for job recruitments," *Arabian J. Sci. Eng.*, vol. 46, no. 4, pp. 3067–3078, Apr. 2021.

**MUHAMMAD SHAHID IQBAL MALIK** received the master's degree in computer engineering and the Ph.D. degree in data mining from the International Islamic University, Islamabad, Pakistan, in 2011 and 2018, respectively. Recently, he completed his Postdoctoral Fellowship with the Laboratory for Models and Methods of Computational Pragmatics, National Research University Higher School of Economics, Moscow, Russia. He is currently an Associate Professor with the Department of Computer Science, University of Wah, Pakistan. Furthermore, he served 12 years in the HVAC industry, Islamabad, and developed several embedded system solutions for air-conditioning systems. He has authored more than 26 research papers published in reputed SCI and Scopus-indexed journals, and conferences. His research interests include data mining, social media mining, natural language processing, text mining, and social computing.

**AFTAB NAWAZ** received the M.S. degree in computer science from Comsats University Islamabad, Attock Campus, Punjab Pakistan. His research interests include sentiment analysis, data mining, and NLP.

**MONA MAMDOUH JAMJOOM** received the Ph.D. degree in computer science from King Saud University. She is currently an Associate Professor with the Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia. Her current research interests include artificial intelligence, machine learning, deep learning, medical imaging, and data science. She has published several research articles in her field.