IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 August 2023, accepted 24 September 2023, date of publication 27 September 2023, date of current version 4 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3320062

RESEARCH ARTICLE

Fine-Tuning Transformer Models Using Transfer Learning for Multilingual Threatening Text Identification

MUHAMMAD REHAN[1], MUHAMMAD SHAHID IQBAL MALIK[2], AND MONA MAMDOUH JAMJOOM[3]

[1]Department of Computer Science, Capital University of Science and Technology, Islamabad 44000, Pakistan
[2]Department of Computer Science, National Research University Higher School of Economics, 109028 Moscow, Russia
[3]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 in part by Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, under Project PNURSP2023R104; and in part by the Computational Resources of High Performance Computing (HPC) Facilities at National Research University Higher School of Economics (HSE University).

ABSTRACT Threatening content detection on social media has recently gained attention. There is very limited work regarding threatening content detection in low-resource languages, especially in Urdu. Furthermore, previous work explored only mono-lingual approaches, and multi-lingual threatening content detection was not studied. This research addressed the task of Multi-lingual Threatening Content Detection (MTCD) in Urdu and English languages by exploiting transfer learning methodology with fine-tuning techniques. To address the multi-lingual task, we investigated two methodologies: 1) Joint multi-lingual, and 2) Joint-translated method. The former approach employs the concept of building a universal classifier for different languages whereas the latter approach applies the translation process to transform the text into one language and then perform classification. We explore the Multilingual Representations for Indian Languages (MuRIL) and Robustly Optimized BERT Pre-Training Approach (RoBERTa) with fine-tuning that already demonstrated state-of-the-art in capturing the contextual and semantic characteristics within the text. For hyper-parameters, manual search and grid search strategies are utilized to find the optimum values. Various experiments are performed on bi-lingual English and Urdu datasets and findings revealed that the proposed methodology outperformed the baselines and showed benchmark performance. The RoBERTa model achieved the highest performance by demonstrating 92% accuracy and 90% macro F1-score with the joint multi-lingual approach.

INDEX TERMS Multi-lingual, Urdu, XLM-RoBERTa, threatening text, fine-tunning, MuRIL.

I. INTRODUCTION

Internet technology allows us to communicate easily and enables us to connect with friends and family from all over the world. In the last decades, social media facilitates us to share our views and opinions and these contents can reach billions of users in mere seconds, leading to not only positive content but also negative exchange of ideas and abusive content [1]. An abusive language can be defined as any type of offensive, vulgarity, profanity, insult, or threatening material to target someone [2]. It is not feasible anymore to identify abusive content manually on social media platforms like Twitter and Facebook etc. Therefore, an automated process is needed to process the abusive content using state-of-the-art Natural Language Processing (NLP) approaches. Threatening content detection from social media has recently got attention but the research in this context is limited. It was explored in the mono-lingual context by building a separate identification system for each language like English [3], and Urdu [4], [5], [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

Various languages exist worldwide and are being spoken which leads to diversity in several multi-lingual tasks related to NLP, such as threatening text, and hate speech detection in social media etc. The task of "categorization of a set of posts written in different languages (Urdu, English, Roman-Urdu, etc.) into predefined groups across languages" is referred to as Multi-lingual Text Classification (MTC). In contrast, "content written in one language but categorized by a classification model learned in another language" is defined as cross-language content classification. Threatening content identification was not handled earlier as a cross-language or MTC methodology, prior approaches addressed it as the mono-lingual paradigm. We are interested to handle it as an MTC task, so MTCD is a binary classification framework, that needs to be trained and tested on the multi-lingual threatening content dataset in English and Urdu languages.

In the literature, three types of techniques were explored to handle the task of MTC. The first approach enforces the design of a separate classification framework for each language [7], [8]. The second approach employs the step of translation for all languages into one universal language before applying a single classification system [9], [10]. The third approach emphasizes the development of one classification framework for different languages. In literature, the MTC was explored in a limited context despite its importance. In this study, we dig into the second and third types of approaches to address the design of MTCD in Urdu and English languages. For this purpose, joint-translation and joint multi-lingual methodologies are investigated [11] and RoBERTa and MuRIL transformer models are considered to employ the MTC paradigm. The RoBERTa model has exhibited state-of-the-art performance for cross-lingual and multi-lingual NLP tasks [12]. Likewise, the MuRIL model already demonstrated benchmark performance in mono-lingual and multi-lingual classification tasks for Indian languages [13].

In this study, a robust solution for multi-lingual threatening content identification is proposed by exploiting the MTC paradigm with the strength of the transfer learning technique. The RoBERTa and MuRIL transformer models are used with fine-tuning hyper-parameters. Joint multi-lingual (XLM-RoBERTa, MuRIL) and joint translation-based MTC approaches are explored. Furthermore, the joint-translated approach is further divided into two parts; 1) Urdu-RoBERTa, and Urdu-MuRIL, and 2) English-RoBERTa, and English-MuRIL. The main contributions of this study addressing the MTCD are presented below:

1. It is the first attempt to propose a multi-lingual threatening content identification framework for English and Urdu languages.
2. We explored joint multi-lingual and joint-translated methodologies for MTC and an algorithm is also designed to understand the methodology and re-produce the results easily.
3. Contextual embeddings offered by state-of-the-art RoBERTa and MuRIL transformer models with fine-tuning are employed for threatening content identification in English and Urdu languages.
4. The experiments showed the effectiveness of the proposed framework by achieving benchmark performance and outperformed the baselines.
5. The proposed framework demonstrated the highest performance with the joint multi-lingual approach by obtaining 92% accuracy, and 90% macro F1-score.
6. The joint multi-lingual approach improved performance by 3.3% in threatening, 10.31% in not-threatening, 6.81% in macro, and 5.3% in weighted F1-score.

The remainder of the paper is organized as follows: Related work is described in section II containing the summary of threatening content identification works and multi-lingual approaches. Section III presents the proposed methodology in detail, followed by section IV, which describes the detail of the experimental setup. Section V presents the results and their analysis. Discussion and limitations are presented in section VI. At last, the conclusion is presented in section VII.

II. RELATED WORK

This section briefly reviews the works related to threatening text detection from social media as a mono-lingual approach. In addition, some state-of-the-art multi-lingual approaches addressing the low and high-resources languages are reviewed.

A. THREATENING TEXT DETECTION

Due to the increasing volume of abusive content on social media, it is difficult to discriminate threatening material from abusive content manually. The first detection model for threatening content was proposed by [14] to detect threats in the Dutch language using n-gram representation. Then, the authors proposed another threatening text detection model [15] in the Dutch language using a shallow parsing mechanism. Later in 2016, a study [16] proposed a threatening content detection method for YouTube comments. They compared lexical, syntactic, and semantic features and concluded that lexical features outperformed them. Then, another study [17] proposed a detection model for threats on Twitter. Authors used Glove embeddings with Convolutional Neural Network (CNN) and their model demonstrated effective performance. Later, [18] explored bag-of-word features with the logistic regression model for threatening text identification from tweets. Their model achieved 98% F1-score.

Likewise, an annotated big corpus containing violent threat material is released by [19]. The dataset contains 10,000 YouTube comments. A novel approach was proposed [3] by addressing threat detection and then target identification on Twitter for the English language. They explored bag-of-words and FastText features. Moreover, Glove embeddings, CNN, and Long-Short Term Memory (LSTM) models are used. Their framework achieved an 85% F1-score. The first identification model for threatening language and its target in Urdu was proposed by [6]. Authors explored FastText, char, and word n-gram models with Machine Learning (ML) and

Deep Learning (DL) models. The experiments revealed that word n-gram and FastText are the best feature models. Then, another study [20] addressed the task of threatening content detection in the Urdu language by utilizing the transformer model. However, their dataset is not balanced.

Likewise, a framework for abusive and threatening context detection in Urdu is proposed by [21]. The authors explored several benchmark features like Bidirectional Encoder Representation Transformer (BERT), mBERT with XGboost, and other ML models. Their framework achieved an 88% F1-score. Recently, a study [5] developed a detection model for Urdu-threatening language by exploring n-grams, TF-IDF, and BOW features with the stacking ensemble model. Their model obtained 73.99% F1-score. Then, another study handled the task of abusive and threatening content identification [22]. They explored word2vec and TF-IDF features with several ML models, but their model did not achieve significant performance. The semantic network-based pipeline [23] is designed for threatening text and target identification in tweets. Their proposed model outperformed the ML models and obtained 76% accuracy. More recently, a robust approach is introduced to identify threatening text and target identification [4] from Urdu tweets. The BERT model is fine-tuned using important hyperparameters and it outperformed the benchmarks. The summary of all the works for threatening content detection is presented in **Table 1**.

B. MTC APPROACHES

In the last decade, multi-lingual content classification got attention from the research community but the research in this area is restricted. Not so many but some works proposed MTC solutions to develop classification systems for low-resource and high-resource languages. The first work was presented in 2006 [24], in which English and Chinese content was classified using the latent semantic indexing approach but they addressed it as mono-lingual. Later, a study [9] explored two WordNet methods to handle the MTC. One method did not consider translation and used directly WordNet related to each language, whereas another approach adopted a translation phase to access WordNet. Then, [25] developed an MTC system for Spanish, Italian, and English languages by exploring n-gram features. They used the Naïve Bayes algorithm for classification. Later, the MTC system for English and Hindi languages is developed [26] by exploring several ML models including genetic algorithm and self-organizing map and they achieved benchmark accuracy by using feature selection algorithms.

Recently, contextual embeddings based on transformer models and deep neural networks are introduced for English [27] and Arabic languages [28]. Furthermore, some multi-lingual transformer models (pre-trained in many languages) like mBERT [29], XLM-RoBERTa [30], and MuRIL [31] are developed. The study [11] proposed an MTC pipeline for offensive text identification in English and Arabic languages. They explored joint multi-lingual and joint-translation approaches and achieved the best performance with the translation-based method (Arabic-BERT). Later, [32] addressed the offensive language detection problem using the MTC framework for English and Bengali languages. The authors proposed a Deep-BERT model and obtained effective performance.

Summarizing the literature, to the best of our knowledge, we did not find any work on multi-lingual threatening content identification. Furthermore, threatening language detection was only addressed in Dutch, English, and Urdu languages. The following limitations related to threatening content detection are identified in the literature:

- **Mono-lingual Methodology:** To the best of our knowledge, prior works used only mono-lingual techniques for designing identification (threatening text) systems for each language.
- **Feature Engineering Methods:** The prior works mainly explored lexical, syntactic, and semantic features for threatening content identification.
- **MTC Methodology:** To the best of our knowledge, we did not find any work on multi-lingual threatening content identification.

III. PROPOSED METHODOLOGY

This section describes the proposed methodology in detail. We address multi-lingual threatening text detection as a binary classification task. The steps employed in the proposed pipeline are presented in **Fig. 1**. Each part of the pipeline is described step-by-step in this section.

A. TRANSLATION PHASE

For the joint-translation approach, we need to translate our multi-lingual Twitter dataset into a single language. This process includes two steps: first we transform the multi-lingual dataset into English; second, we transform it into Urdu language. The detail of the steps is provided below:

- ❖ **Universal Urdu Corpus:** We already have Urdu corpora [4] in the multi-lingual dataset. The other corpora (English) [3] is translated into Urdu using the Google translator API. The translated data is edited manually to resolve the issues and inconsistencies. After that, both corpora are combined to get a single Urdu corpus.
- ❖ **Universal English Corpus:** The Urdu corpus is translated into English using the services of Google translator. After the manual editing of translated data, we combined it with the English part of the multi-lingual dataset to finalize a single English corpus.

B. PRE-PROCESSING PHASE

Pre-processing steps are an important phase for automated text classification tasks. After pre-processing, it is convenient to extract precise information from the dataset. We employed the following steps to pre-process the multi-lingual dataset.

- Punctuations, mentions, hashtags, numbers, HTML tags, and URLs are removed.

TABLE 1. Summary of Mono-lingual approaches for threatening content detection.

[Ref]	Language	Target		Feature Models	ML & DL Models
		MONO	MULTI		
[14]	Dutch	√		N-grams	Custom classifier
[15]	Dutch	√		Shallow parsing mechanism	---
[16]	English	√		Lexical, syntactic and semantic	MEC, SVM, RF
[17]	English	√		Glove	CNN
[18]	English	√		Bag of words	LR
[19]	English	√		Released a dataset	-----------------
[3]	English	√		Bag of words, FastText, Glove	CNN, LSTM
[6]	Urdu	√		Word & char n-gram, FastText embeddings	SVM, CNN, LSTM, RF, LR, MLP
[20]	Urdu	√		Transformer Model	----------
[21]	Urdu	√		BERT transformer	mBERT, XGboost,
[5]	Urdu	√		TF-IDF, n-grams	Stacking ensemble model
[22]	Urdu	√		Word2vec, TF-IDF	CNN, SVM, AdaBoost
[23]	English	√		Named entity recognition	Custom-made
[4]	Urdu	√		Fine-tuning BERT	BERT
Proposed	**English, Urdu**		√	**Fine-tuning RoBERTa, MuRIL models**	**Customized Classifiers**

MONO: Monolingual; MULTI: Multi-lingual; MEC: Maximum Entropy Classifier; SVM: Support Vector Machine; RF: Random Forest; LR: Logistic Regression; CNN: Convolutional Neural Network; LSTM: Long Short Term Memory Model; MLP: Multi-layer Perceptron;

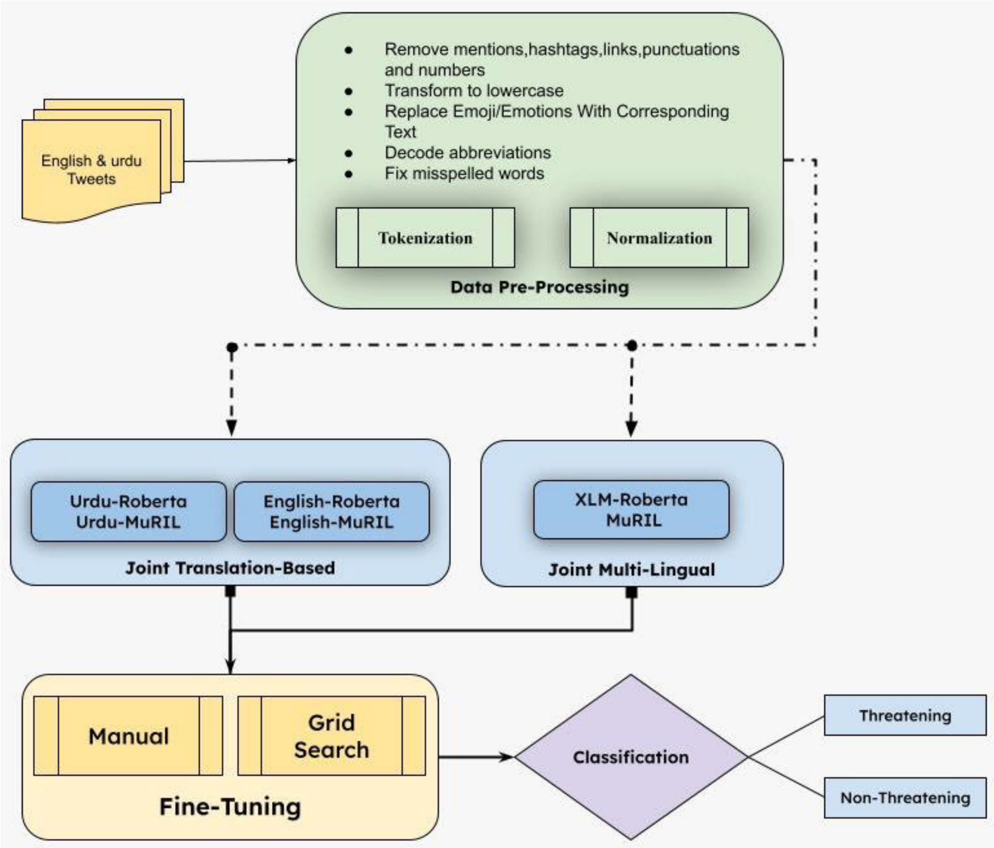

FIGURE 1. Proposed pipeline for multi-lingual threatening content detection.

- Emoji/Emoticons are replaced with relevant text, since these hold important information.
- English text to lowercase conversion.
- Address the issue of miss-spelled words (English text).
- Decode the English abbreviations (pls, thx, etc)

After the phase of pre-processing, we employ two approaches of MTC; a joint-translated approach and a joint multi-lingual approach. In joint multi-lingual approach, we combine the tweets of English and Urdu languages whereas in joint-translated, we use two corpora one by one (Universal Urdu and Universal English) described in section III-A. The algorithmic pseudo-code of our pipeline is described in **Fig. 2**.

C. TRANSFORMER MODELS

In this study, we explore two transformer models with fine-tuning to design an effective multi-lingual threatening content detection in English and Urdu languages. The detail of two state-of-the-art transformer models is presented below:

1) RoBERTa

The RoBERTa transformer model was introduced by [33] in 2019. The difference between BERT and RoBERTa is that the RoBERTa eliminates the objective of next sentence pre-training and re-visits the significant hyperparameters and applies them with larger ranges, but it is mainly based on the architecture of the BERT model. XLM-RoBERTa is trained on 100 languages. Here, we are interested to explore the RoBERTa with fine-tuning of significant hyper-parameters to design a robust multi-lingual threatening text detection model. For this purpose, we investigate XLM-RoBERTa for the joint multi-lingual task and RoBERTa and Urdu-RoBERTa for the joint-translation approach.

2) MuRIL

The MuRIL transformer is the most recent multi-lingual model on Google. It is pre-trained on 17 Indian languages to promote some downstream NLP operations like spelling differences and transliteration to enhance linguistic interoperability. According to the cross-lingual XTREME test, MuRIL presented better than the BERT model [34]. In this study, we are using three models of MuRIL: 1) For multi-lingual text, 2) For English text, and 3) For Urdu text.

D. TOKENIZATION & REPRESENTATION

To make the input data compatible with the two transformer models (RoBERTa and MuRIL), a few steps are needed. First of all, we have to perform tokenization of English and Urdu posts to transform them into a unified format. For this purpose, the [CLS] token is added at the start of every post and the [SEP] token is added at the end of every post. This process reveals that from where each sentence starts and ends and the resultant entity is a single vector for the whole post. It results in a universal vector to input for the MuRIL and RoBERTa classifiers.

TABLE 2. Search space of hyper-parameters for RoBERTa and MuRIL transformers.

Hyperparameters	Grid Search
Sequence length	64, 128
Batch size	16, 32, 64
Learning rate	3e-5, 2e-5, 1e-5, 0.99e-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

Regarding the RoBERTa transformer, we use RoBERTa-Tokenizer [33] for English text, RoBERTa-UrduTokenizer (https://huggingface.co/urduhack/roberta-urdu-small) for Urdu language and XLMRoBERTaTokenizer [35] for multi-lingual text. For the MuRIL transformer model, we again use three tokenizers [31] for Urdu, English, and multi-lingual text. After that, each token (tweet) is mapped to an index corresponding to the transformer model vocabulary (RoBERTa or MuRIL).

E. FINE-TUNNING

The next task is the fine-tuning of both transformer models (RoBERTa and MuRIL). We applied two methods to explore suitable values of hyper-parameters for fine-tuning, i.e. manual search and grid search. The number of hyper-parameters is eight, the search space for the eight parameters is presented in **Table 2**. The maximum number of characters supported by a tweet is 280, therefore 128 could be the maximum value of sequence length. Thus two sequence lengths are investigated to analyze their impact on binary classification, i.e. 64, and 128. Three batch sizes (16, 32, 64) are evaluated one by one to investigate the impact of each batch size. Likewise, three learning rates are explored to see their impact on the training, validation, and test part of the multi-lingual dataset. The other parameters and their corresponding ranges are presented in **Table 2**. These parameters are used to fine-tune RoBERTa and MuRIL transformer models.

For RoBERTa, we are using its un-cased pre-trained base model for Urdu, English, and multi-lingual dataset. The hidden size is 768, attention_heads are 12 and hidden_layers are also 12. For the MuRIL transformer, we are using its cased pre-trained base model, trained on 17 Indian languages with the MLM layer intact. We are interested to explore the strengths of RoBERTa and MuRIL models by fine-tuning eight hyper-parameters.

For dataset splitting, we employed stratified data splitting method and split our datasets into 80-20, in which 20% is used for testing the validated model and the remaining 80% is further split into 90-10, where 90% is actually used for training and 10% is used for validation. The optimizer function is utilized for updating the parameters for each epoch and the output of each training and validating cycle is measured using training loss, validation loss, accuracy, and macro

```
Algorithm 1: Multi-lingual Threatening Content Detection (English-dataset, Urdu-dataset)
procedure multi-lingual-threatening-content-detection (English-dataset, Urdu-dataset)
    S1 ← Pre-Processing (English-dataset);             // pre-processing of English dataset
    S2 ← Pre-Processing (Urdu-dataset);                // pre-processing of Urdu dataset
    ENG ← Google-Translate (S2);                       // translate Urdu dataset into English using google API
    Eng-S ← Merge (S1, ENG);                           // combine both English and translated English datasets
    URU ← Google-Translate (S1);                       // translate English dataset into Urdu using google API
    URU-S ← Merge (S2, URU);                           // combine both Urdu and translated Urdu datasets
    Com-S ← Merge (Eng-S, URU-S);                      // combine both English and Urdu datasets
    Eng-S1 ← Eng-RoBERTa-Tokenizer (Eng-S);            // Apply English RoBERTa Tokenizer
    Urdu-S1 ← Urdu-RoBERTa-Tokenizer (URS-S);          // Apply Urdu RoBERTa Tokenizer
    Com-S1 ← XLM-RoBERTa-Tokenizer (Com-S);            // Apply XLM-RoBERTa Tokenizer
    Classification (Eng-S1, 1);                        // classification using Joint-translated English-RoBERTa model
    Classification (Urdu-S1, 2);                       // classification using Joint-translated Urdu-RoBERTa model
    Classification (Com-S1, 3);                        // classification using Joint multi-lingual XLM-RoBERTa model
end procedure

procedure Pre-Processing (D)
        D1 ← Cleaning (D);                             // Remove hashtags, HTML tags, mentions, punctuations, URLs, and numbers
        D2 ← Lower-Case (D1);                          // Lower-case conversion for case-sensitive language
        D3 ← Replace-Emoji (D2);                       // Replace emoji/emoticons with the corresponding text
        Return D3;
end procedure

procedure Classification (dataset D, mode)             // classification using fine-tuning of relevant RoBERTa
    if (mode = 1)
            Model ← fine-tunning (D, English-RoBERTa, 80-20)    // fine-tuning with English RoBERTa
      else if ((mode = 2)
            Model ← fine-tunning (D, Urdu-RoBERTa, 80-20)       // fine-tuning with Urdu RoBERTa
      else
            Model ← fine-tunning (D, XLM-RoBERTa, 80-20)        // fine-tuning with XLM-RoBERTa
     confusion-matrix ← generate-results (Model);
     accuracy ← compute-accuracy (confusion-matrix);
     precision ← compute-precision (confusion-matrix);
     recall ← compute-recall (confusion-matrix);
     F1-score ← compute-F1(confusion-matrix);
end procedure
```

FIGURE 2. Pseudo-code of MTCD methodology.

F1-score. Google Colab and High-Performance Computing (HPC) local cloud is used to conduct the experiments of fine-tuning two transformers. The transformer models are initialized in their pre-trained settings and then annotated datasets are used to fine-tune important parameters.

F. CLASSIFICATION

For the classification task, we appended a single layer containing the Softmax function on the top of the RoBERTa and MuRIL transformer models. This layer is used to classify the tweets as threatening or non-threatening. The fine-tuned tweet is forwarded to the softmax function and the transformer model is trained to optimize the cross-entropy loss.

G. CATASTROPHIC FORGETTING

The literature reveals that the transformer models encounter the problem of forgetting already learned knowledge while attempting to learn new knowledge by fine-tuning the hyperparameters. This concept is termed as "Catastrophic forgetting in transfer learning" [4]. We exhaustively tried several learning rates to analyze the risk of catastrophic forgetting while fine-tuning the MuRIL and RoBERTa models. After several trials, we found that higher learning rates make poor convergence and result in failure most of the time. Thus, we obtained the best results with learning rates $\leq$ 1e-5.

H. OVERFITTING

For deep learning models, it is a common issue to choose the appropriate number of epochs because choosing very few epochs results in under-fitting, and choosing too many results in over-fitting. We investigated the impact of the number of epochs (5 or 10) on the validation and test parts of the multi-lingual threatening content dataset using the loss function. The performance of RoBERTa and MuRIL transformer models are monitored and we reached on a conclusion that 10 epochs are suitable for getting the desired results by evaluating the trained model on the validation and test parts of the dataset.

TABLE 3. Detail of multi-lingual threatening corpus (English and Urdu).

Dataset	Threatening	Non-Threatening	Total	Train	Validate	Test
English	1185	128	1313	945	105	263
Urdu	1200	1200	2400	1728	192	480

IV. EXPERIMENTAL SETUP

The parameters for fine-tuning the MuRIL and RoBERTa models are described in the previous section. Here we describe the detail of datasets for multi-lingual threatening content detection tasks, the baselines, and the evaluation metrics to evaluate the performance of the classifiers.

A. DATASET DESCRIPTION (MULTI-LINGUAL CORPUS)

As described earlier, we address the task of binary classification for threatening content detection in English and Urdu posts. We used the threatening content Urdu dataset [4], which was collected from Twitter containing 1200 threatening and 1200 non-threatening instances. The dataset was collected from Pakistani Twitter accounts ranging from August 2020 to August 2022. The other dataset is in English and consists of YouTube comments posted on videos related to religion and politics [3]. This dataset has several inconsistencies, therefore we manually resolved those issues. The cleaning process results in 1313 instances, among which 128 are non-threatening and the remaining are threatening. Further detail of both datasets including train, validate, and test instances are presented in **Table 3**.

B. BASELINE

To compare our proposed methodology with benchmarks, we chose word2vec and RoBERTa as feature models and combine them with Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), CNN, and Bi-LSTM algorithms. The reason for choosing these classifiers is that they demonstrated state-of-the-art performance for binary classification in NLP tasks [36]. The following combinations of models are chosen as the baselines:

- Word2vec + SVM
- Word2vec + LR
- Word2vec + RF
- Word2vec + CNN
- Word2vec + Bi-LSTM
- RoBERTa + SVM
- RoBERTa + LR
- RoBERTa + RF
- RoBERTa + CNN
- RoBERTa + Bi-LSTM

Next, we describe each feature model briefly.

1) Word2vec

Word2vec word embedding model has shown state-of-the-art performance in many classification tasks related to the NLP domain [36], [37], [38]. There are two methods supported by word2vec to generate word embeddings; skip-gram and CBOW. In this study, we used the skip-gram model with 300 dimensions.

2) RoBERTa AS A FEATURE MODEL

We used RoBERTa as a feature model and combined it with above mentioned ML and DL models to design comparable models to compare with fine-tuned RoBERTa and MuRIL transformer models.

C. EVALUATION METRICS

Four state-of-the-art metrics are used to evaluate the performance of the proposed and baseline classifiers. The metrics with their mathematical formulation are presented next:

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN} \tag{1}$$

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP} \tag{2}$$

$$\text{F1} - \text{score} = 2.(\text{precision} * \text{recall})/\text{precision} + \text{recall} \tag{3}$$

$$\text{Accuracy} = \text{TP} + \text{TN}/\text{TP} + \text{TN} + \text{FP} + \text{FN} \tag{4}$$

where

TP = Number of positive instances and predicted as positive.

FP = Number of negative instances but incorrectly predicted as positive.

TN = Number of negative instances and predicted as negative.

FN = Number of positive instances but incorrectly predicted as negative.

V. RESULTS AND ANALYSIS

In this section, three sets of experiments are performed to investigate the effectiveness of the proposed framework for threatening content identification in English and Urdu languages.

A. JOINT MULTI-LINGUAL RESULTS

In this section, we conducted experiments to fine-tune the XLM-RoBERTa and MuRIL transformers on the joint dataset (Urdu and English) to evaluate the joint multi-lingual methodology. We used already described hyper-parameters (**Table 2**) for the fine-tuning process and transformer models are trained, validated, and tested by employing the given mechanism (section III-C2).

The fine-tuning results of both models (XLM-RoBERTa and MuRIL) are presented in **Table 4**. For each transformer model, we presented only the best results while trying several

TABLE 4. Fine-tunning results of XLM-Roberta and MURIL on the test part of multi-lingual dataset.

Models	L. Rate	H.Dropout	W.Decay	Accuracy	F1-score			
					Threat	Not	Macro	Weighted
MuRIL	2e-5	0.05	0.01	90.15	93.25	81.85	87.55	89.99
	1e-5	0.05	0.01	89.38	92.70	80.57	86.63	89.23
	1e-5	**0.1**	**0.1**	**90.35**	**93.39**	**82.14**	**87.76**	**90.17**
XLM-RoBERTa	2e-5	0.05	0.01	89.19	92.65	79.56	86.11	88.91
	0.99e-5	0.05	0.0208	91.51	94.10	84.83	89.46	91.45%
	1e-5	0.1	0.1	88.80	92.41	78.68	85.54	88.48
	1e-5	**0.05**	**0.01**	**91.89**	**94.40**	**85.31**	**89.86**	**91.80**

TABLE 5. Comparison of joint multi-lingual models with ten baselines.

Classifiers	Acc	F1-score				Precision	Recall
		Threat	Not	Macro	Weighted		
Word2vec+SVM	82.05	87.21	69.90	78.56	82.26	82.59	82.05
Word2vec+LR	76.64	84.93	48.07	66.50	74.40	75.08	76.64
Word2vec+RF	85.52	90.09	73.12	81.61	85.24	85.18	85.52
Word2vec+CNN	77.22	85.82	42.16	63.99	73.34	77.13	77.22
Word2vec+Bi-LSTM	74.90	81.64	60.37	71.00	75.56	76.79	74.90
XLM-RoBERTa+SVM	84.56	89.80	68.25	79.02	83.64	84.36	84.56
XLM-RoBERTa+LR	86.68	91.07	73.76	82.42	86.13	86.49	86.68
XLM-RoBERTa+RF	86.29	90.82	73.00	81.91	85.73	86.06	86.29
XLM-RoBERTa+CNN	86.87	91.10	75.00	83.05	86.50	86.59	86.87
XLM-RoBERTa+Bi-LSTM	72.20	79.49	56.89	68.19	73.03	74.62	72.20
Fine-tuned MuRIL	90.35	93.39	82.14	87.76	90.17	90.21	90.35
Fine-tuned XLM-RoBERTa	**91.89**	**94.40**	**85.31**	**89.86**	**91.80**	**91.80**	**89.05**

hyper-parameters and for the F1-score metric, class-wise performance is presented (threatening, not-threatening, macro, and weighted average). The reported results contain learning rate, hidden-dropout, and weight-decay parameters, and the sequence length is 128. It is evident that the best performance by the MuRIL model is obtained on 1e-5 learning rate, 0.1 hidden-dropout, and 0.1 weight-decay resulting in 90.35% accuracy, and 90.17% weighted F1-score. Likewise, the best performance with XLM-RoBERTa is 91.89% accuracy and 91.80% weighted F1-score on 1e-5 learning rate, 0.05 hidden-dropout, and 0.01 weight-decay. Thus, this experiment concluded that the best performance is obtained by fine-tuned XLM-RoBERTa model as compared to fine-tuned MuRIL.

The next experiment compared the performances of the fine-tuned MuRIL and XLM-RoBERTa models with ten baselines and the results are shown in **Table 5**. The results are reported in accuracy, precision, recall, and class-wise performances in F1-score. In baselines, word2vec with the RF model and XLM-RoBERTa with the CNN model presented better performance in comparison to other combinations. In the proposed models, fine-tuned MuRIL presented better than all baselines, and fine-tuned XLM-RoBERTa presented better than all baselines including fine-tuned MuRIL. The XLM-RoBERTa improved accuracy by 5.02%, precision by 5.21%, and recall by 2.185 in comparison with baselines. Furthermore, it improved 3.3% in threatening, 10.31% in not-threatening, 6.81% in macro, and 5.3% in weighted F1-score compared to baselines as shown in **Fig 3**. The highest improvement is observed for detecting not-threatening class instances in the F1-score. The word2vec embeddings combined with ML and DL models did not perform well. Thus, the joint multi-lingual methodology proved itself by demonstrating benchmark performance and also outperformed the baselines in class-wise, macro, and weighted evaluation metrics.

B. JOINT-TRANSLATED RESULTS

In this section, we performed two sets of experiments; one for joint-translated English and the other for joint-translated Urdu model. The objective here is to evaluate the strength of the joint-translated approach for multi-lingual threatening content identification. We already devised two corpora for the joint-translated approach; 1) English, and 2) Urdu. First, we evaluate the effectiveness of the joint-translated English approach and then the joint-translated Urdu approach.

The fine-tunning of English-RoBERTa and MuRIL is performed using the hyper-parameters enlisted in **Table 2** and best results are reported in the **Table 6**. The joint-translated English corpora is used for experimental setup and

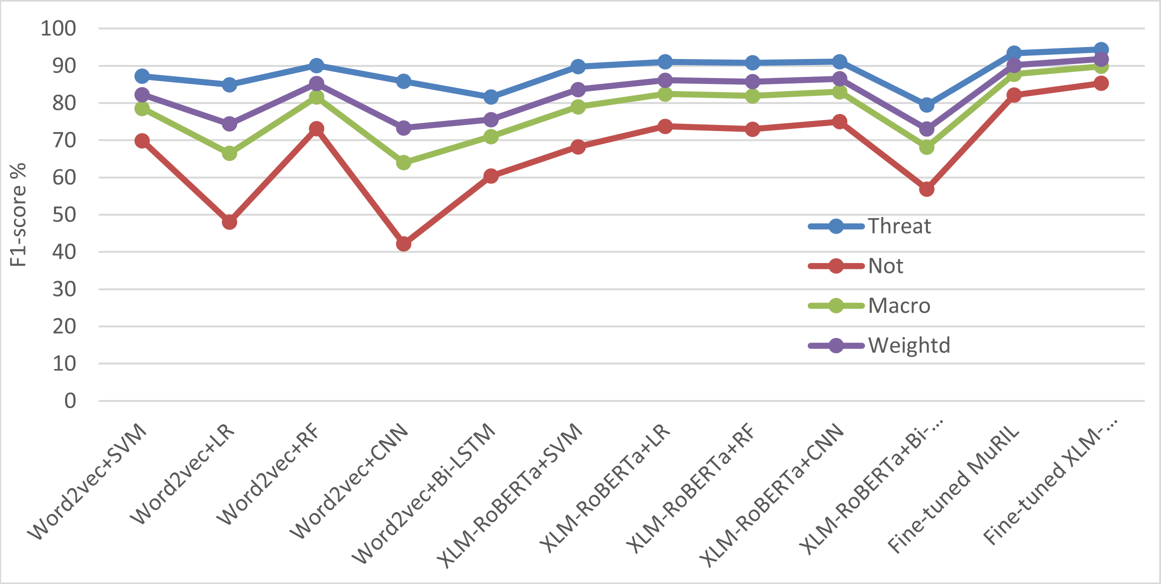

FIGURE 3. Class-wise performance of proposed framework and baselines (F1-score).

TABLE 6. Comparison of joint-translated English models with baselines.

Classifiers	Acc	F1-score				Precision	Recall
		Threat	Not	Macro	Weighted		
Word2vec+SVM	74.16	80.30	62.43	71.37	75.11	77.75	74.16
Word2vec+LR	65.81	71.33	57.64	64.48	67.36	75.57	65.81
Word2vec+RF	80.12	86.95	58.33	72.64	78.64	79.40	80.12
Word2vec+CNN	76.34	85.29	39.59	62.44	72.03	76.37	76.34
Word2vec+Bi-LSTM	68.79	74.80	59.01	66.90	70.22	76.01	68.79
RoBERTa+SVM	82.50	87.13	72.67	79.90	82.94	83.98	82.50
RoBERTa+LR	82.50	87.25	72.15	79.70	82.87	83.62	82.50
RoBERTa+RF	76.74	85.61	39.38	62.49	72.19	77.63	76.74
RoBERTa+CNN	84.10	88.54	74.03	81.28	84.33	84.74	84.10
RoBERTa+Bi-LSTM	69.78	76.76	56.82	66.79	70.97	74.07	69.78
Fine-tuned MuRIL	87.48	91.41	76.92	84.16	87.20	87.23	87.48
Fine-tuned English-RoBERTa	**89.86**	**92.97**	**81.85**	**87.41**	**89.74**	**89.72**	**89.86**

performance of classifiers are reported using four metrics. Furthermore, the performance of ten baselines are also added to compare them with proposed joint-translated english models.

The word2vec and XLM-RoBERTa are combined with three ML and two DL models to create comparable models. Considering baselines, the superior performance is demonstrated by RoBERTa+CNN by achieving 84.33% weighted F1-score. In contrast, the MuRIL fine-tuned model obtained 87.20% weighted F1-score and fine-tuned RoBERTa model demonstrated 89.74% weighted F1-score. Thus, the proposed joint-translated English methodology outperformed the baselines and demonstrated benchmark performance. The performances of baselines and proposed frameworks per class-wise and in macro and weighted F1-score are presented in **Fig.** 4. It is observable that English-RoBERTa improved accuracy by 5.76%, precision by 4.98%, recall by 5.76% in comparison with baseline. Furthermore, the proposed framework improved the threatening class by 4.43%, not-threatening by 7.82%, macro F1-score by 6.13%, and weighted F1-score by 5.41%. We noticed substantial improvement achieved by the proposed joint-translated English model in comparison with baselines, indicating the effectiveness of the proposed methodology. The largest improvement is observed in identifying not-threatening class instances. This proves the strength of the proposed methodology for the joint-translated English approach to address the problem of multi-lingual threatening content identification.

The last set of experiments is performed to investigate the effectiveness of proposed joint-translated Urdu models for multi-lingual threatening content identification task. The proposed models are Urdu-RoBERTa and MuRIL. We performed fine-tuning of Urdu-RoBERTa and MuRIL using the same parameters mentioned earlier. After fine-tuning,

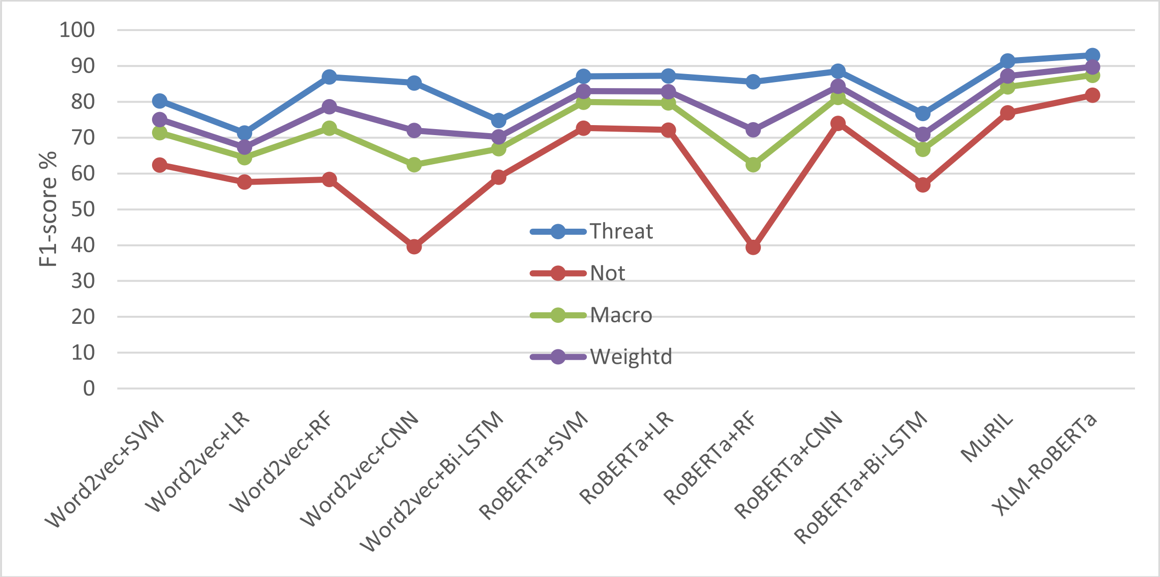

FIGURE 4. Class-wise performance of joint-translated English models and baselines (F1-score).

TABLE 7. Comparison of joint-translated Urdu models with baselines.

Classifiers	Acc	F1-score				Precision	Recall
		Threat	Not	Macro	Weighted		
Word2vec+SVM	75.95	81.65	65.12	73.38	76.85	72.64	76.33
Word2vec+LR	68.94	75.20	58.45	66.82	70.33	67.26	70.78
Word2vec+RF	80.36	86.97	60.16	73.57	79.18	77.71	71.70
Word2vec+CNN	77.15	84.30	58.09	71.19	76.68	72.23	70.46
Word2vec+Bi-LSTM	70.54	76.70	59.95	68.32	71.83	68.46	72.11
Urdu-RoBERTa+SVM	81.16	87.23	64.12	75.68	80.51	77.91	74.30
Urdu-RoBERTa+LR	81.56	86.86	69.13	77.99	81.71	77.59	78.46
Urdu-RoBERTa+RF	78.96	86.56	51.61	69.08	76.40	78.47	67.05
Urdu-RoBERTa+CNN	83.77	89.30	66.39	77.84	82.64	83.60	75.33
Urdu-RoBERTa+Bi-LSTM	63.33	71.80	47.56	59.68	64.76	59.83	61.53
Fine-tuned Urdu-RoBERTa	86.37	90.81	73.64	82.23	85.82	85.56	80.22
Fine-tuned MuRIL	87.17	90.93	78.08	84.51	87.20	84.37	84.65

the best results are reported. In addition, for state-of-the-art comparison, we compared the fine-tuned models with ten baselines, and the results are presented in **Table 7**. Among baselines, the best performance is demonstrated by the Urdu-RoBERTa+CNN model by achieving 83.77% accuracy and 77.84% macro F1-score. On the other end, fine-tuned Urdu-RoBERTa transformer presented 86.37% accuracy and 82.23% macro F1-score. Furthermore, fine-tuned MuRIL model demonstrated the best performance by obtaining 87.17% accuracy and 84.51% macro F1-score. Thus proposed fine-tuned MuRIL model outperformed the baselines including fine-tuned Urdu-RoBERTa model and showed benchmark performance. It is important to note that fine-tuned MuRIL model beat the fine-tuned Urdu-RoBERTa for the joint-translated Urdu approach. The class-wise performance of all classification models in the F1-score is presented in **Fig. 5**. It is visible that the proposed joint-translated Urdu approach obtained substantial improvement in performance, i.e. 3.4% in accuracy, 0.77% in precision, and 9.32% in recall.

Furthermore, we observed 1.63% improvement in threatening class, 11.69% improvement in not-threatening, 6.67% improvement in macro F1-score, and 4.56% improvement in weighted F1-score.

After an extensive set of various experiments, we conclude that the proposed methodology is very helpful in identifying multi-lingual threatening content in English and Urdu. It outperformed the ten baseline models while testing both types (joint multi-lingual and joint-translated) of approaches. A substantial improvement is observed while evaluating the classifiers in macro and weighted F1-scores.

VI. DISCUSSION AND LIMITATIONS

To extract new insights and knowledge from the plethora of data and to automate the business process, content

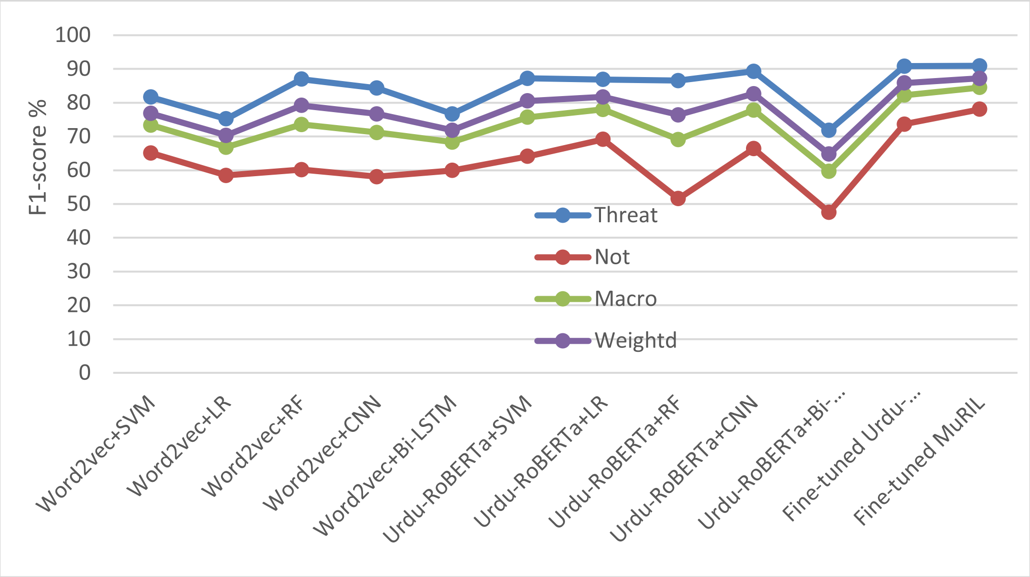

FIGURE 5. Class-wise performance of joint-translated Urdu models and baselines (F1-score).

classification is a significant task. Furthermore, the requirement for an accurate and efficient multi-lingual NLP framework is in high demand due to the large multilingualism on social media. To handle part of these issues, this study conducted a comparative analysis and evaluation of a proposed multi-lingual threatening content detection framework on a bi-lingual dataset. The task is of business interest and high research and the developed framework is evaluated in terms of accuracy, precision, recall, and macro & weighted F1-score. Moreover, the findings of this research help to unveil important characteristics that are useful in the identification of threatening expressions from Twitter and YouTube comments. The current study has brought out worthful insights for social media users and our society to save them and promote peace and harmony. The research on threatening content detection in low-resource languages is very restricted and the main emphasis was on monolingual techniques. According to our knowledge, work on multi-lingual threatening content detection is missing in the literature. We attempted to fill out this gap by designing a multi-lingual pipeline using two techniques (joint multilingual and joint-translated approach) to identify threatening content in English and Urdu languages.

The proposed pipeline is evaluated on a bi-lingual semi-supervised setup containing English and Urdu corpora. The transfer learning methodology in the form of fine-tuning of RoBERTa and MuRIL transformers is utilized. The experiments help us to discover insights for further research and aid in making a practical approach. The findings disclosed that the proposed pipeline obtained state-of-the-art performance by getting 92% accuracy and 90% macro F1-score with the joint multi-lingual approach. Furthermore, it outperformed the baselines and proved itself a benchmark approach for multi-lingual threatening content detection. Therefore, these findings suggest that our approach can be applied to similar MTC in NLP tasks.

The current study has some limitations; First of all, only two languages (English and Urdu) are considered to test the performance of the proposed framework, more languages can be incorporated to deal with the task of MTC, especially low-resource languages like Russian, Chinese, roman Urdu and Hindi, etc. Second, the proposed framework can be tested on a larger corpus to make the framework more generalizable. Third, threatening content identification is addressed here as a binary classification task. It will be more appropriate if we address the task of ''who is being threatened, individual or community''. It will be helpful to locate the targeted community.

VII. CONCLUSION

In this paper, we developed a multi-lingual text classification framework to handle the task of multi-lingual threatening content detection in English and Urdu languages. We took advantage of the transfer learning approach to deal with the complexity and overhead of designing a separate classification system for each language. Joint multi-lingual and joint-translated techniques are explored to design the robust MTCD system. The proposed MTCD system is based on the RoBERTa and MuRIL transformer models, which were fine-tuned on bilingual semi-supervised threatening content detection corpus. The proposed methodology is also transformed into an algorithm for readers and researchers to reproduce the results and understand the methodology easily. Two benchmark transformer models (RoBERTa and MuRIL) are chosen and their effectiveness for the MTCD task is explored extensively by fine-tuning. The proposed pipeline is comprised of four modules (pre-processing, tokenization, fine-tuning, and classification). The experiments on bilingual semi-supervised corpus revealed that the proposed methodology demonstrated superior performance than ten baselines for joint multi-lingual and joint-translated approaches. The best performance is observed against the joint multi-lingual approach, that is 92% accuracy and 90% macro F1-score.

In future work, the proposed pipeline can be extended for other low-resource languages such as Russian, Chinese, Roman Urdu, etc. In addition, the proposed framework can be easily applicable to other binary and multi-class classification tasks in the NLP field. Another possible direction would be to re-visit the methodology by hybridizing the transformer architecture with some robust algorithms to improve performance.

ACKNOWLEDGMENT

This work was supported in part by Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, under project PNURSP2023R104. Moreover, this research was supported in part by the Computational Resources of High Performance Computing (HPC) facilities at National Research University Higher School of Economics. This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] S. Malliga, K. Shanmugavadivel, R. Chinnasamy, N. Subbarayan, A. Ganesan, D. Ravi, V. Palanikumar, and B. R. Chakravarthi, "On fine-tuning adapter-based transformer models for classifying abusive social media Tamil comments," Kongu Eng. College, Perundurai, India, Tech. Rep., 2023.

[2] M. Anand, K. B. Sahay, M. A. Ahmed, D. Sultan, R. R. Chandan, and B. Singh, "Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques," *Theor. Comput. Sci.*, vol. 943, pp. 203–218, Jan. 2023.

[3] N. Ashraf, R. Mustafa, G. Sidorov, and A. Gelbukh, "Individual vs. group violent threats classification in online discussions," in *Proc. Companion Web Conf.*, Apr. 2020, pp. 629–633.

[4] M. S. I. Malik, U. Cheema, and D. I. Ignatov, "Contextual embeddings based on fine-tuned Urdu-BERT for Urdu threatening content and target identification," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 7, Jul. 2023, Art. no. 101606.

[5] 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.

[6] 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.

[7] T. Gonalves and P. Quaresma, "Multilingual text classification through combination of monolingual classifiers," in *Proc. 4th Workshop Legal Ontologies Artif. Intell. Techn.*, vol. 605, 2010, pp. 29–38.

[8] M. R. Amini, C. Goutte, and N. Usunier, "Combining coregularization and consensus-based self-training for multilingual text categorization," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2010, pp. 475–482.

[9] M. A. Bentaallah and M. Malki, "The use of WordNets for multilingual text categorization: A comparative study," in *Proc. ICWIT*, 2012, pp. 121–128.

[10] B. P. Prajapati, S. Garg, and M. H. Panchal, "Automated text categorization with machine learning and its application in multilingual text categorization," in *Proc. Nat. Conf. Advance Comput. (NCAC)*, 2009, pp. 204–209.

[11] F.-Z. El-Alami, S. Ouatik El Alaoui, and N. En Nahnahi, "A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 8, pp. 6048–6056, Sep. 2022.

[12] A. Höfer and M. Mottahedin, "Minanto at SemEval-2023 task 2: Fine-tuning XLM-RoBERTa for named entity recognition on English data," in *Proc. The 17th Int. Workshop Semantic Eval. (SemEval)*, 2023, pp. 1127–1130.

[13] R. Rajalakshmi, S. Selvaraj, F. R. Mattins, P. Vasudevan, and A. M. Kumar, "HOTTEST: Hate and offensive content identification in Tamil using transformers and enhanced stemming," *Comput. Speech Lang.*, vol. 78, Mar. 2023, Art. no. 101464.

[14] N. Oostdijk and H. van Halteren, "N-gram-based recognition of threatening tweets," in *Computational Linguistics and Intelligent Text Processing*. Cham, Switzerland: Springer, 2013, pp. 183–196.

[15] N. Oostdijk and H. van Halteren, "Shallow parsing for recognizing threats in Dutch tweets," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2013, pp. 1034–1041.

[16] A. Wester, L. Øvrelid, E. Velldal, and H. L. Hammer, "Threat detection in online discussions," in *Proc. 7th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2016, pp. 66–71.

[17] A. Wester, L. Øvrelid, E. Velldal, and H. L. Hammer, "Threat detection in online discussions," in *Proc. 7th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2016, pp. 66–71.

[18] C. D. Lim, "Detecting legally actionable threats on Twitter using natural language processing and machine learning," M.S. thesis, Dept. Cogn. Sci. Artif. Intell., Tilburg Univ., Tilburg, The Netherlands, 2018.

[19] H. L. Hammer, M. A. Riegler, L. Øvrelid, and E. Velldal, "THREAT: A large annotated corpus for detection of violent threats," in *Proc. Int. Conf. Content-Based Multimedia Indexing (CBMI)*, Sep. 2019, pp. 1–5.

[20] S. Kalraa, M. Agrawala, and Y. Sharmaa, "Detection of threat records by analyzing the tweets in Urdu language exploring deep learning transformer-based models," in *Proc. CEUR Workshop*, 2021, pp. 1–7.

[21] 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*.

[22] M. Humayoun, "Abusive and threatening language detection in Urdu using supervised machine learning and feature combinations," 2022, *arXiv:2204.03062*.

[23] F. Fkih and G. Al-Turaif, "Threat modelling and detection using semantic network for improving social media safety," *Int. J. Comput. Netw. Inf. Secur.*, vol. 15, no. 1, pp. 39–53, Feb. 2023.

[24] C.-H. Lee, H.-C. Yang, and S.-M. Ma, "A novel multilingual text categorization system using latent semantic indexing," in *Proc. 1st Int. Conf. Innov. Comput., Inf. Control (ICICIC)*, vol. 2, Sep. 2006, pp. 503–506.

[25] P. P. Dhyani and S. Mittal, "Multilingual text classification," *Int. J. Eng. Res.*, vol. 4, no. 3, pp. 99–101, Mar. 2015.

[26] K. Rani, "Satvika: Text categorization on multiple languages based on classification technique," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 3, pp. 1578–1581, 2016.

[27] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019.

[28] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," 2020, *arXiv:2003.00104*.

[29] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, p. 2.

[30] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.

[31] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, and P. Talukdar, "MuRIL: Multilingual representations for Indian languages," 2021, *arXiv:2103.10730*.

[32] M. Anwar Hussen Wadud, M. F. Mridha, J. Shin, K. Nur, and A. Kumar Saha, "Deep-BERT: Transfer learning for classifying multilingual offensive texts on social media," *Comput. Syst. Sci. Eng.*, vol. 44, no. 2, pp. 1775–1791, 2023.

[33] 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*.

[34] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4411–4421.

[35] R. Mehta and V. Varma, "LLM-RM at SemEval-2023 task 2: Multilingual complex NER using XLM-RoBERTa," 2023, *arXiv:2305.03300*.

[36] 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.

[37] 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.

[38] 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.

MUHAMMAD REHAN received the M.S. degree in computer science from the Capital University of Science and Technology, Islamabad, Pakistan. His current research interests include data mining, NLP, and machine learning.

MUHAMMAD SHAHID IQBAL MALIK received the master's degree in computer engineering in 2011, and the Ph.D. degree in data mining from International Islamic University, Islamabad, Pakistan, in 2018. He is currently a Postdoc Fellow with the Lab for Models and Methods of Computational Pragmatics, National Research University Higher School of Economics, Moscow, Russia. Previously, he served more than three years as an Assistant Professor at the Capital University of Science and Technology, and four years as a Lecturer at Comsats University Islamabad, Pakistan. In addition, he served 12 years in HVAC industry, Islamabad and developed several embedded systems solutions for Air-conditioning systems. He authored more than 23 research papers published in leading International Journals and Conferences. His research interests include social media mining, natural language processing, predictive analytics and social computing. He is the reviewers of famous International Journals.

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.