Threatening Expression and Target Identification in Under-Resource Languages Using NLP Techniques

Muhammad Shahid Iqbal Malik(✉)

Department of Computer Science, National Research University Higher School of Economics, 11 Pokrovskiy Boulevard, Moscow 109028, Russian Federation
mumalik@hse.ru

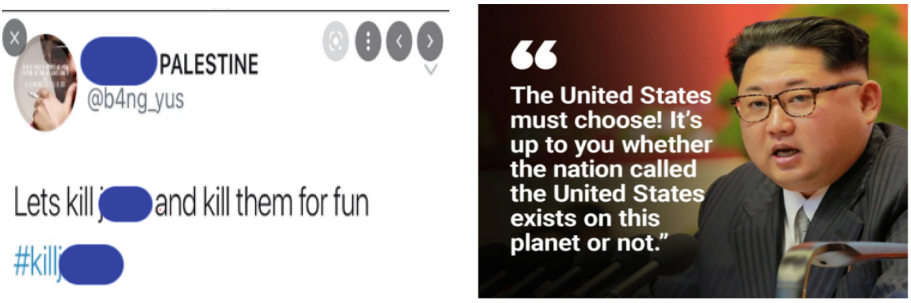
Abstract. In recent decades, hate speech on social media platforms has been on the rise. It is highly desired to control this kind of material because it initiates unrest and harms to the society. Literature describes several forms of the hate speech and it is quite challenging to differentiate between these forms and to design an automated detection system, especially for under-resource languages. In this study, we propose a robust framework for threatening expressions and its target identification in Urdu (Nastaliq style) language. The proposed methodology presents each step in detail like data collection & annotation, cleaning & pre-processing step, and fine-tuning of Robustly Optimized Bidirectional Encoder Representations from Transformer (Urdu-RoBERTa) with grid search technique for hyper-parameters optimization. The study exploits the strength of a pre-trained Urdu-RoBERTa as a transfer learning technique with grid search fine-tuning. The proposed framework is compared with state-of-the art baseline and ten comparable models and it outperformed all for both tasks (threatening expression and target identification). Furthermore, the proposed framework obtained benchmark performance and improved the f1-score with substantial margin.

Keywords: Natural language processing · threatening expression · low-resource · target identification · RoBERTa · hyper-parameters

1 Introduction

A large number of users who use social media platforms is escalating at a very high speed according to the recent statistics, and various social media platforms are commonly used for sharing views and opinions. As social media is open for everyone by providing freedom of speech, it is being used for the spreading of positive views as well as for the propagation of hate speech and negative content. Therefore, it is highly desired to control this kind of material because it initiates unrest and harms to individuals and affect society by arousing violence, terrorist activities, aggression, etc. Two examples of hate speech expressions are presented in Fig. 1. In the left part (a), a tweet was posted in 2014 motivating killings of Jews for fun and in the right part (b), a leader is giving a threat to the United States.

D. I. Ignatov et al. (Eds.): AIST 2023, LNCS 14486, pp. 3–17, 2024.
https://doi.org/10.1007/978-3-031-54534-4_1



(a) (b)

Fig. 1. Two examples of hate speech expressions on social media [1] (the sensitive information is patched by blue boxes by the editors) (Color figure online)

There is a consensus among the researchers on the definition of hate speech that "it is a language used to attack/target an individual or a group on the basis of ethnicity, race, gender, or religion etc." [2]. In literature, several state-of-the-art definitions of hate speech are presented and their sources are mainly from scientific studies and popular social media platforms [3–5]. Some examples are:

- "Hate speech attack others dependent on racism, ethnicity, public start, sexual bearing, sex, character, age, handicap, or genuine illness" [6]
- "Hate Speech is a purposeful attack on a specific social occasion of people motivated by the pieces of the group's character" [7]
- "Hate Speech is a toxic speech attack on a person's individuality and likely to result in violence when targeted against groups based on specific grounds like religion, race, place of birth, language, residence, caste, community, etc." [1]

These benchmark definitions describe the hate speech in several perspectives and researchers address hate speech according to their understanding, knowledge, and thinking prospective. In addition, literature described several forms of hate speech like toxicity [8], profanity [9], discrimination [10], Cyberbullying [11] etc., and these forms of hate speech are presented in the Fig. 2. The definitions of some of the forms and their differences compared to hate speech are presented below:

- Profanity vs Hate Speech: "Hostile or indecent words or expressions but hate speech can use profane words but not always"
- Toxicity vs Hate Speech: "Conveying content that is disrespectful, abusive, unpleasant, and harmful but Not all toxic comments contain hate speech"
- Discrimination vs Hate Speech: "Interaction via a distinction and afterward utilized as the premise of unreasonable treatment but Hate speech is a virulent form of discrimination"

Threatening text or violent threat is one of the form of hate speech. Few studies handled the task of threatening expression detection in high-resource languages like English [12–14] etc., but under-resource languages have very limited such approaches. Furthermore, target identification from threatening expression is almost ignored for

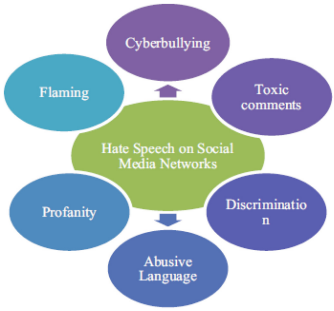


Fig. 2. Various forms of hate speech

under-resource languages. A large variety of languages are spoken worldwide. The landscape of world's popular spoken languages is presented in the Fig. 3. We can analyze the proportion of population speaking under-resource languages like Arabic, Urdu, Hindi, Chinese, Bengali and Russian in the Asian subcontinent. Urdu is the national language of Pakistan and it is being spoken by approximately 300 million people in Canada, USA, UK and India. Furthermore, it is an indigenous language of approximately 170 million speakers in the Asian region. Urdu language is identified as an under-resource language because several content processing toolkits and other resources are not available [15].

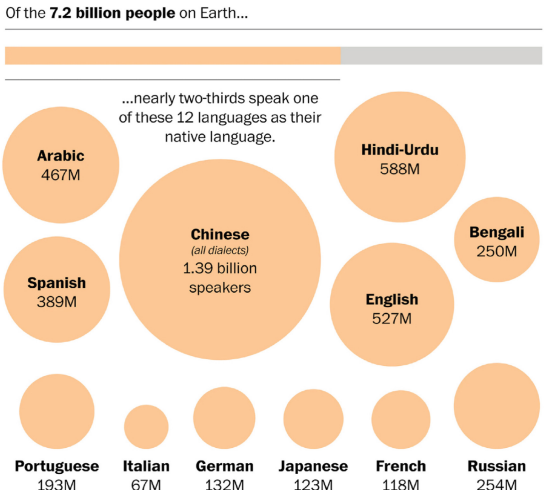


Fig. 3. The landscape of world languages in seven maps and charts (according to Washington post, 2022)

There are two writing styles in Urdu language, one is Nastaliq and other is Roman. In this study, we address Urdu language in Nastaliq writing style and introduced a model for threatening expression identification and then target identification from threatening

expression. It is a kind of hierarchical classification task. Three objectives are planned to design an automated and effective threatening expression identification model.

- To develop an accurate system for the identification of threatening expression in Urdu (Nastaliq writing style) language.
- Regarding threatening expression; design a target identification framework to distinguish between individual and group.
- The proposed framework should be based on an automated feature-generation technique in contrast to hand-crafted features.

The remaining part of the paper is organized as follows; challenges with under-resource languages are presented in Sect. 2, followed by Sect. 3, in which related work is described. Section 4 presents the description of proposed model and experiments are discussed and analyzed in Sect. 5. Section 6 provides the conclusion of the research work and future prospective in this domain.

2 Challenges with Under-Resource Languages

While dealing with resource-poor languages, we encounter the following challenges.

- Lack of annotated datasets.
- Several essential resources and accurate text processing toolkits are not available, especially for Urdu, and Bengali languages etc.
- Some languages have multiple scripts, like Urdu has Nastaliq/Arabic style or Latin/Roman style.
- Social media users usually use multiple scripts while sharing their opinions. Issue of code-mixing.
- Pertinent language models (pre-trained) are scarcely available.

3 Related Work

In this section, a brief review of prior studies handling abusive and threatening expression identification in under-resource languages. It is not an easy task to filter out unwanted content from social media posts. The first model was introduced [16] in 2021 to address the task of threatening expression and target identification in Urdu. The word and char n-grams, and FastText are combined with some Machine Learning (ML) and Deep Learning (DL) models, but their dataset has improper annotations. Then another study [17] proposed a detection model for Urdu threat records but there was an issue of highly imbalanced dataset. Similarly, another detection model is developed by Das et al. [18] to handle abusive and threatening text detection in Urdu. They utilized transformer model with XGboost and obtained 54% f1-score for threat record detection. Then another detection model was proposed for abusive and threatening expression detection in Urdu [19] by exploiting word n-grams with word2vec model but their propose solution only achieved 49.31% f1-score. A recent study [20] proposed an ensemble model-based system for threat records detection and obtain 73.99% f1-score.

Some research studies focused on the design of identification models related to abusive expression detection in Urdu. Hussain et al. [21] introduced an offensive expression

identification model in Urdu language using Facebook posts. They incorporated ensemble model with word2vec and obtained 88.27% accuracy on balanced dataset. Likewise, another framework [22] is introduced for Twitter platform by utilizing char and word n-grams, and FastText embeddings and obtained 82.68% f1-score. For Roman and Nastaliq Urdu, a significant framework is proposed for abusive expression detection. They utilized bag-of-words with ML and DL models and got 96% accuracy with Convolutional Neural Network (CNN) model. A recent study [23] used char and word n-grams, and BERT model for the detection of offensive content and their proposed system obtained 86% f1-score.

In literature, we found lack of annotated corpus for threatening content and target identification task. Furthermore, majority of approaches are based on hand-crafted features and lacking in automated feature engineering. The comparison between ML and DL models are rarely handled in the literature.

4 Proposed Model

In this section, proposed framework is described in detail. The pipeline for the design of proposed framework is presented in Fig. 4. Here, the complete flow of the process adopted in the design of proposed framework is defined.

4.1 Problem Definition

We address the task of threatening expression detection as a binary classification problem. Here, two-level of classification is performed; for the first level, text is categorized into threatening or not-threatening. For the second level, the threatening expressions are further categorized into individual or group category.

4.2 Dataset Collection and Annotation

Twitter platform was chosen to collect the tweets from Pakistani Twitter accounts. There is an annotation issues with the prior dataset [16] available for this task. Therefore, we designed a new dataset and crawled the tweets using Twitter API. The time period is chosen from August 2020 to August 2022 due to uncertainty and un-stability in the politics of Pakistan. At first, we designed a lexicon of seed words containing 250 keywords in total. This lexicon helped us to identify the relevant tweets from Pakistani Twitter accounts. Some example keywords are listed in the Table 1.

After that cleaning process is employed to the crawled data and the steps of cleaning process are described in the left part of Fig. 5. After cleaning the dataset, it was shared to three annotators for the annotation purposes. Two level of annotations are performed. In the first level, threatening vs not-threatening and in the second level, threatening tweets are further categorized into “individual or group” for target tagging. Some example guidelines are presented in the Table 2. The annotators are chosen by following some criteria described by the study [24].

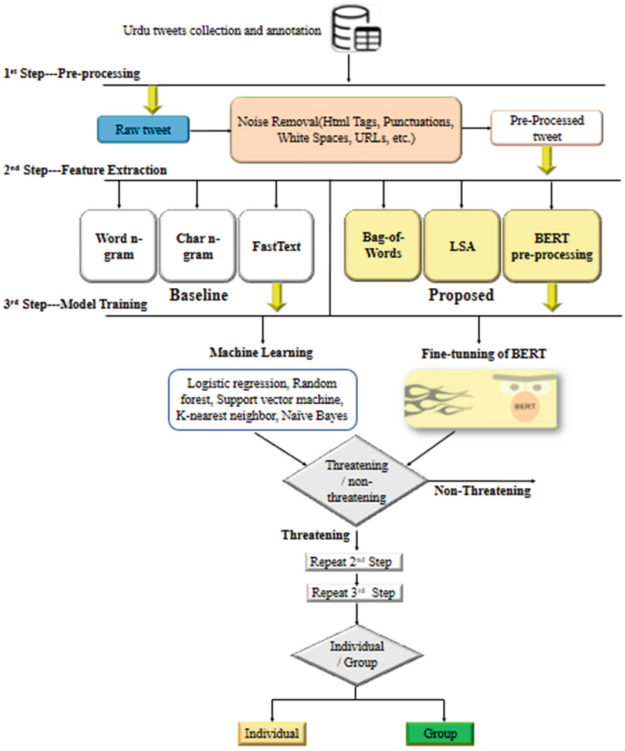


Fig. 4. Proposed pipeline for threatening expression and target identification

Table 1. Examples of Seed words

Urdu	Translation	Urdu	Translation	Urdu	Translation
ٹکڑے	Pieces	افسوسناک موت	Sad death	خون کے آنسو	Tears of blood
ہلاک	Killed	ٹـــانـگیں	Break the legs	عبرت کا نشان	Sign of lesson
لاشیں	Corpses	چھـــورا	Stabbing	خون پی جاؤں	Drink blood
دھمکی	Threat	چیر دوں	Tear	جان سے مار دو	Kill with soul
جنازہ	Funeral	الٹا لٹکا	Hang upside	سر تن سے جدا	Head separated from body

4.3 Pre-processing

Following pre-processing steps are applied on the dataset:

- Removal of punctuations, mentions, hashtags, numbers, HTML tags, and URLs.
- Emoji/Emoticons are replaced with relevant text.
- Stop words removal (not for transformers)

An example of all pre-processing steps are presented in the Table 3 for the understanding of the readers.

Table 2. Examples (guidelines) for annotating the tweets

S#	Urdu	Translation	Level 1	Level 2
01	لعنتی کتے عـــوام کا حـــال دیکھو لوڈشیڈنگ کو روکو خدا کی قسم ہم تم کو زندہ جلا دیں گے	Cursed dogs, look at the condition of the people, stop the load shedding, by God we will burn you alive	Threatening	Individual
02	دونوں کی آنکھیں نکال	I will take out the eyes of both	Threatenin	Group
03	ابھی کچھ کچھ امید زندہ ہورہی ہے	There is some hope now	N o n - threatening	NA
04	تم کتے کی وہ دم ھو جو بارہ سال بند رکھی پھر بھی ٹیڑی یہ کام تمھارہ مرشد ھی کرتا تھا	You are the dog's tail that was tied for twelve years, yet your mentor used to do this	N o n - threatening	NA

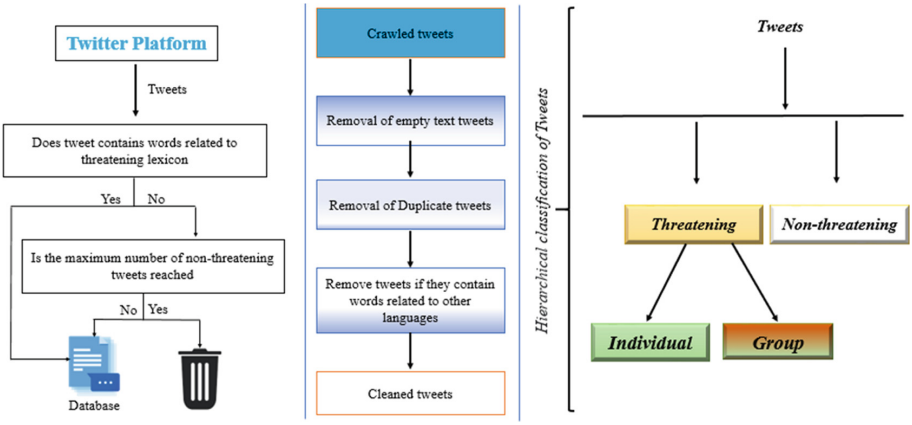


Fig. 5. Process of data collection, cleaning and classification

4.4 Urdu-RoBERTa Model

In this study, we exploited the strength of Urdu-RoBERTa model for the identification of threatening expression and target from the tweets. The RoBERTa model has already proved its effectiveness for several NLP tasks [25, 26]. There are several benefits of RoBERTa including fast development, fewer data requirements, and contextual feature generation, etc. It is a pre-trained transformer model proposed by the researchers [27] and mainly based on the BERT transformer model. Here, we used the Urdu-RoBERTa model (https://huggingface.co/urduhack/roberta-urdu-small) by fine-tuning some important hyperparameters for the threatening expression identification task. The word count and cloud representation of the annotated dataset is presented in Fig. 6.

Fine-Tuning Process: Some steps are usually required in the fine-tuning of any transformer model. The input data must be transformed into a pre-defined format to make it compatible for the RoBERTa architecture. After pre-processing, we need to apply data transformation step.

Table 3. Demonstration of pre-processing steps

Urdu Text Preprocessing	
1. Punctuations Removal	
Before	After
ہم آنکھیں نکال لیں گے نواز شریف کی.....پھر نابینا باپ کو لے کر بیٹھی رہو گی.....جھوٹی اور مکار عورت	ہم آنکھیں نکال لیں گے نواز شریف کی پھر نابینا باپ کو لے کر بیٹھی رہو گی جھوٹی اور مکار عورت We will take out the eyes of Nawaz Sharif then you will sit with the blind father a lying and deceitful woman
2. Stop word Removal	
دشمن کے دانت کھٹے اور جسم مردہ ہونگے انشاءاللہ The teeth of the enemy will be sour and the body will be dead, God willing	دشمن دانت کھٹے جسم مردہ انشاءاللہ enemy's teeth sour, body dead, God willing
3. Replace Emoji with Corresponding Text	
😡تم خبیث لوگ بچ نہیں سکو گے چاہے کچھ بھی کر لو You evil people will not escape no matter what you do	مشتعل تم خبیث لوگ بچ نہیں سکو گے چاہے کچھ بھی کر لو You evil people will not be saved no matter what you do, enraged
4. Hashtag, Numbers and Link Removal	
چن چن کے ماریں گے تمہارے گھروں میں گھس کر ماریں گے ;128514#&http://t.co/cOU2WQ5L4q	چن چن کے ماریں گے تمہارے گھروں میں گھس کر ماریں گے

Fig. 6. Word count and cloud representation of the annotated dataset

Data Transformation: It is a mandatory step for fine-tuning process. First of all, uncased RoBERTa tokenizer is applied to each tweet to break it into word tokens. After that, each sentence is appended by the [CLS] token at the start and the [SEP] token at the end. Then each token of the tweet is mapped to an index. As maximum length of a tweet is 280 characters, therefore we chose 64 and 128 sequence lengths for the experimental setup. The process of addition of attention masks is performed to fed input to RoBERTa base classifier.

Classification: We applied 80–20 data split on the dataset. The 20% is used for testing and the 80% part is further divided into 90–10, in which 90% is actually used for training and 10% is used for validation. A single layer is appended on the top of Urdu-RoBERTa base model for the binary classification task. For fine-tuning, we chose Grid Search technique to find the optimum values of hyper-parameters. The list of hyper-parameters and their values are presented in the Table 4. The optimizer function is utilized for updating all the parameters of each epoch.

Catastrophic Forgetting and Overfitting: As RoBERTa is pre-trained in a generic prospective on a big corpus and it needs appropriate fine-tuning with important hyper-parameters for a new learning. The new learning could encounter the issues of Catastrophic Forgetting, and Overfitting. Every transformer is prone to catastrophic forgetting. We dealt this issue by exploring a range of learning rates and concluded that higher learning rates encounter convergence failures and best results are obtained with learning rate of 2e−5. To deal with the issue of overfitting, we monitor loss value on the validation dataset and found that 5 epochs are appropriate to save the fine-tuning process from over and under fitting.

Table 4. List of hyper-parameters and their ranges

Hyperparameters	Grid Search
Sequence length	64, 128
Batch size	8, 16, 32
Learning rate	1e−4, 1e−5, 2e−5, 3e−4, 3e−5, 5e−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

4.5 Experimental Setup

The detail of benchmark and the comparable models are presented below. We chose following ML models because they demonstrated state-of-the-art performance in several NLP tasks [28, 29].

Benchmark: Amjad et al. study [16].

Comparable Models

- Latent Semantic Analysis (100) + Logistic Regression
- Latent Semantic Analysis (100) + Random Forest
- Latent Semantic Analysis (100) + Support Vector Machine

- Latent Semantic Analysis (100) + Naive Bayes
- Latent Semantic Analysis (100) + K-nearest neighbor
- Bag-of-words + Logistic Regression
- Bag-of-words + Random Forest
- Bag-of-words + Support Vector Machine
- Bag-of-words + Naive Bayes
- Bag-of-words + K-nearest neighbor

In total, two feature engineering techniques (latent semantic analysis and bag-of-words) and five ML models are utilized. In addition, the classifiers' performance is evaluated using standard accuracy, precision, recall and macro f1-score measures.

5 Results and Analysis

In this section, two types of experiments are performed to fine-tune the Urdu-RoBERTa model for threatening expression and target identification tasks. The proposed framework is compared with a baseline and ten comparable models.

5.1 Fine-Tuning Urdu-RoBERTa and Comparison with Baselines (Threatening Identification)

Six RoBERTa classifiers are trained, validated and tested using fine-tuning process employed to design an effective identification model for threatening expression. Two sequence lengths (64 and 128) and three batch sizes are tried with other hyper-parameters (listed in Table 4) for fine-tuning. The training and validation loss obtained by applying sequence lengths of 64, and 128 with batch sizes of 8 and 16 are presented in Fig. 7. It is clearly visible that training loss decreases continuously from epoch 1 to 5, indicates continuous learning of Urdu-RoBERTa in the training phase, but Fig. 7 shows that validation loss started decreasing up to the 3rd epoch and then started increasing up to 5th epoch. This signals that further training and validation may leads to overfitting.

Next, the performance of fine-tuned RoBERTa is compared with a baseline [16] and comparable models. From the applied ML models, logistic regression presented best performance as compared to random forest, support vector machine, naïve bayes and k-nearest neighbor. Therefore, we added only best results from the comparable models as shown in Table 5. The performance is presented in accuracy, precision, recall, and macro f1-score. Among the baseline, char 5-g presented highest metric values. i.e. f1-score is 85.83% and accuracy is 86.25%. In contrast, the proposed fine-tuned RoBERTa (with sequence length of 64 and batch size of 8) outperformed the baseline and the comparable models, by providing 87.8% f1-score and 87.5% accuracy. This leads to 2% improvement in f1-score and 1.25% in accuracy. Thus fine-tuned Urdu-RoBERTa proved its effectiveness for the identification of threatening expression on the Twitter network. The performance of FastText embedding in baseline is worst here, although it demonstrated state-of-the-art performance for other NLP tasks.

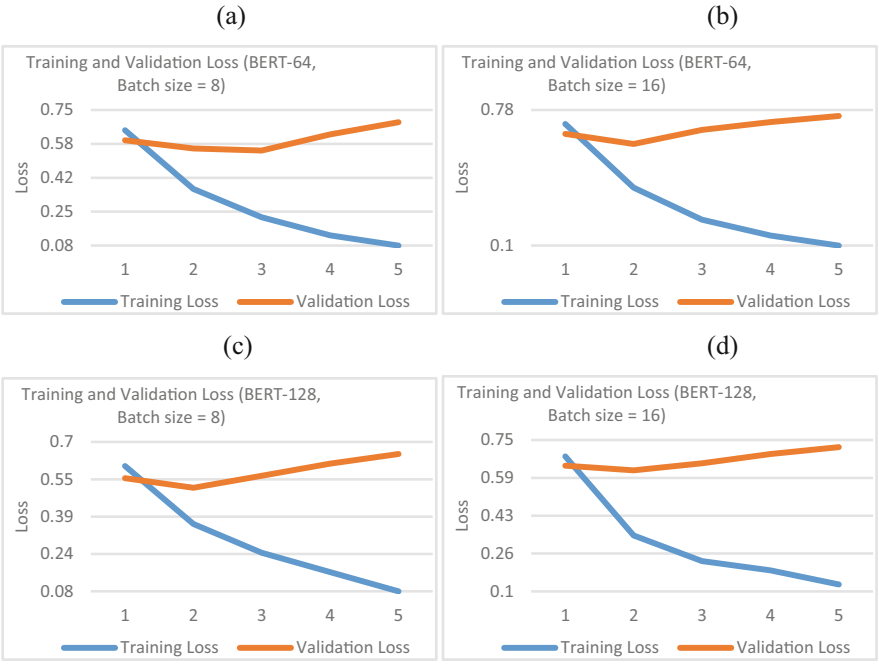


Fig. 7. Training & validation losses employing SL of 64 including a) batch-size of 8, and b) batch-size of 16, and SL of 128 including c) batch-size of 8, and d) batch-size of 16

5.2 Fine-Tuning Urdu-RoBERTa and Comparison with Baselines (Target Identification)

Here, we performed fine-tuning of Urdu-RoBERTa for the target identification task and compared it with baseline [16] and comparable models. Again six classifiers are trained, validated and tested using grid search fine-tuning process for five epochs and results are presented in Table 6. We employed the same procedure and range of parameters as presented in Sect. 5.1. The sequence lengths of 64 and 32 and batch size of 8, 16 and 32 are explored and results are demonstrated. The performance of baseline and comparable models are also added in the Table 6.

For target identification task, all evaluation metrics indicate that fine-tuned Urdu-RoBERTa with sequence length of 128 and batch size of 8 outperformed the baseline and comparable models by achieving the benchmark accuracy of 82.5% and 83.20% f1-score. The Urdu-RoBERTa with sequence length of 64 and batch size of 8 did not demonstrated highest performance for target identification task but presented comparable performance to bag-of-words + logistic regression model. Among the baseline, combined word (1–2-3) grams demonstrated better performance as compared to other features for target identification task. Again FastText embeddings did not perform well. As a whole, proposed framework improves accuracy by 0.58% and f1-score by 0.41% as compared to state-of-the-art baseline.

Table 5. Performance comparison of fined-tuned RoBERTa with baseline and comparable models (Threatening vs non-Threatening)

Type	Features	Accuracy	Precision	Recall	F1-score
Baseline [16]	Word uni-gram	83.83	81.51	80.17	80.83
	Word bi-gram	82.08	81.97	82.64	82.30
	Word tri-gram	75.42	80.39	67.77	73.54
	Word combined (1-2-3)	82.92	80.30	87.60	83.79
	Char 1-g	68.75	69.16	68.59	68.87
	Char 2-g	79.58	82.14	76.03	78.97
	Char-3-g	80.00	84.11	74.38	78.94
	Char 4-g	84.16	87.38	80.16	83.62
	Char 5-g	**86.25**	**89.28**	**82.64**	**85.83**
	Char 6-g	85.83	87.82	83.47	85.59
	Char 7-g	82.50	83.19	81.81	82.50
	Char 8-g	82.08	82.50	81.81	82.15
	Char combined (1–8)	80.00	74.82	90.90	82.08
	FastText	59.17	63.64	54.69	58.82
Proposed	Bag of words	83.75	81.54	87.60	84.46
	Latent Semantic Analysis	81.25	79.23	85.12	82.07
	BERT-64 (8)	**87.5**	**86.4**	**89.26**	**87.8**
	BERT-64 (16)	83.75	85.34	81.82	83.54
	BERT-64 (32)	84.58	87.5	80.99	84.12
	BERT-128 (8)	**87.5**	**89.57**	**85.12**	**87.29**
	BERT-128 (16)	84.17	86.73	80.99	83.76
	BERT-128 (32)	82.92	80.3	87.6	83.79

In the end, we conclude our experiments by summarizing three advantages of the proposed framework. First, it improves the identification performance for both tasks (threatening and target identification) in comparison with benchmark with substantial margin. Second, the proposed pipeline is based on automated feature generation method in contrast to hand-crafted features (baseline). Third, the state-of-the-art Urdu-RoBERTa language model is capable to capture the threatening language context realistically in the Urdu and experiments proved its effectiveness.

Table 6. Performance comparison of fined-tuned RoBERTa with baseline and comparable models (Target Identification)

Type	Feature Set	Accuracy	Precision	Recall	F1-score
Baseline [16]	Word uni-gram	80.33	81.51	80.17	80.83
	Word bi-gram	82.08	81.97	82.64	82.30
	Word tri-gram	75.42	80.39	67.77	73.54
	Word combined (1-2-3)	81.92	80.30	87.60	82.79
	Char 1-g	60.83	68.08	50.00	57.65
	Char 2-g	69.16	72.13	68.75	70.40
	Char 3-g	70.83	72.30	73.43	72.86
	Char 4-g	70.83	73.77	70.31	72.00
	Char 5-g	69.16	71.42	70.31	70.86
	Char 6-g	67.50	71.18	65.62	68.29
	Char 7-g	60.83	66.66	53.12	59.13
	Char 8-g	61.66	68.75	51.56	58.92
	Char combined (1–8)	72.50	79.24	65.62	71.79
	FastText	54.17	51.67	54.39	52.99
Proposed	Bag-of-words	80.33	81.51	80.17	80.83
	Latent Semantic Analysis	68.33	68.57	75.00	71.64
	BERT-64 (8)	79.17	80	81.25	80.62
	BERT-64 (16)	75	81.48	68.75	74.58
	BERT-64 (32)	75	79.31	71.88	75.41
	BERT-128 (8)	**82.5**	**85.25**	**81.25**	**83.2**
	BERT-128 (16)	79.17	81.97	78.13	80
	BERT-128 (32)	76.67	83.33	70.31	76.27

6 Conclusion

This study addressed the task of threatening expression and target identification for an under-resource language. First, hate speech is described in detail with several definitions presented by the literature. Then, various forms of hate speech are summarized and threatening expression (type of hate speech) is defined. After that, challenges dealing with under-resource languages are discussed. Urdu language with Nastaliq style is chosen to demonstrate the steps of the proposed methodology for threatening expression and target identification task. The process of dataset collection and then annotation are described with examples. After that, data cleaning and pre-processing steps are demonstrated on real tweets. The Urdu-RoBERTa model is used with grid search fine-tuning to capture the actual context of threatening expression and their target identification. The issues of catastrophic forgetting and overfitting are highlighted and their solutions

are discussed. After that, implementation detail of fine-tuning process is described and results are compared with a benchmark and ten comparable models. The proposed system outperformed the benchmark and comparable models for both tasks (threatening expression and target identification). Thus, the proposed system and its findings may assist law and enforcement organizations to detect and filter-out this kind of material from social media platforms.

Regarding future prospects, researchers will encounter following challenges: First, there will be an issue of interpretability because each low-resource language has different way of creating context to describe an opinion. Second, appropriate categorization of various types of hate speech is challenging in low-resource languages as their definitions overlap. Third, designing an efficient code-mixed content identification framework for low-resource language is not an easy task.

Acknowledgments. 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 (HSE University). Moreover, this research was supported in part by computational resources of HPC facilities at HSE University.

References

1. Chhabra, A., Vishwakarma, D.K.: A literature survey on multimodal and multilingual automatic hate speech identification. Multimed. Syst. 1–28 (2023)
2. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (2017)
3. Delgado, R., Stefancic, J.: Images of the outsider in American law and culture: can free expression remedy systemic social ills. Cornell L. Rev. **77**, 1258 (1991)
4. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. (CSUR) **51**(4), 1–30 (2018)
5. Youtube. YouTube hate policy. https://support.google.com/youtube/answer/2801939?hl=en. 2019
6. Twitter. Twitter_Hate Definition. https://support.twitter.com/articles/.2017
7. De Gibert, O., et al.: Hate speech dataset from a white supremacy forum. arXiv preprint arXiv: 1809.04444 (2018)
8. Andročec, D.: Machine learning methods for toxic comment classification: a systematic review. Acta Universitatis Sapientiae, Informatica **12**(2), 205–216 (2020)
9. Malmasi, S., Zampieri, M.: Challenges in discriminating profanity from hate speech. J. Exp. Theor. Artif. Intell. **30**(2), 187–202 (2018)
10. Thompson, N.: Social Problems and Social Justice. Bloomsbury Publishing (2017)
11. Chen, Y., et al.: Detecting offensive language in social media to protect adolescent online safety. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing. IEEE (2012)
12. Ashraf, N., et al.: Individual vs. group violent threats classification in online discussions. In: Companion Proceedings of the Web Conference 2020 (2020)
13. Jiang, L., et al.: Intelligent control of building fire protection system using digital twins and semantic web technologies. Autom. Constr. **147**, 104728 (2023)
14. Mazari, A.C., Boudoukhani, N., Djeffal, A.: BERT-based ensemble learning for multi-aspect hate speech detection. Cluster Comput. 1–15 (2023)

15. Nawaz, A., et al.: Extractive text summarization models for Urdu language. Inf. Process. Manag. **57**(6), 102383 (2020)
16. Amjad, M., et al.: Threatening language detection and target identification in Urdu tweets. IEEE Access **9**, 128302–128313 (2021)
17. Kalraa, S., Agrawala, M., Sharmaa, Y.: Detection of Threat Records by Analyzing the Tweets in Urdu Language Exploring Deep Learning Transformer-Based Models (2021)
18. Das, M., Banerjee, S., Saha, P.: Abusive and threatening language detection in Urdu using boosting based and BERT based models: a comparative approach. arXiv preprint arXiv:2111.14830 (2021)
19. Humayoun, M.: Abusive and threatening language detection in Urdu using supervised machine learning and feature combinations. arXiv preprint arXiv:2204.03062 (2022)
20. Mehmood, A., et al.: Threatening URDU language detection from tweets using machine learning. Appl. Sci. **12**(20), 10342 (2022)
21. Hussain, S., Malik, M.S.I., Masood, N.: Identification of offensive language in Urdu using semantic and embedding models. PeerJ Computer Science **8**, e1169 (2022)
22. Amjad, M., et al.: Automatic abusive language detection in Urdu tweets. Acta Polytechnica Hungarica 1785–8860 (2021)
23. Saeed, R., et al.: Detection of offensive language and its severity for low resource language. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **22**, 1–27 (2023)
24. Malik, M.S.I., Cheema, U., Ignatov, D.I.: Contextual embeddings based on fine-tuned Urdu-BERT for Urdu threatening content and target identification. J. King Saud Univ.-Comput. Inf. Sci. 101606 (2023)
25. Malik, M.S.I., et al.: Multilingual hope speech detection: a robust framework using transfer learning of fine-tuning RoBERTa model. J. King Saud Univ.-Comput. Inf. Sci. **35**(8), 101736 (2023)
26. Rehan, M., Malik, M.S.I., Jamjoom, M.M.: Fine-tuning transformer models using transfer learning for multilingual threatening text identification. IEEE Access (2023)
27. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
28. Younas, M.Z., Malik, M.S.I., Ignatov, D.I.: Automated defect identification for cell phones using language context, linguistic and smoke-word models. Expert Syst. Appl. **227**, 120236 (2023)
29. Malik, M.S.I., Imran, T., Mamdouh, J.M.: How to detect propaganda from social media? Exploitation of semantic and fine-tuned language models. PeerJ Comput. Sci. **9**, e1248 (2023)