Detection of Abusive Records by Analyzing the Tweets in Urdu Language Exploring Transformer Based Models

Sakshi Kalra[a], Yash Bansal[a] and Yashvardhan Sharma[a]

[a]*Department of Computer Science and Information Systems, Birla Institute of Technology and Science Pilani, Pilani Campus, Rajasthan, India*

Abstract

As social media platforms grow in popularity and importance, the consequences of their misuse become more severe. Numerous posts containing abusive language directed at specific users worsen users' experiences on such platforms. In this paper, we look at the task of detecting Abuse in the Urdu Language. We experiment with different machine learning algorithms and Transformer based models to achieve the best results on this one-of-a-kind task of Abusive language detection in Urdu. We got accuracy equal to 0.93607 on the test dataset using the soft voting technique with the help of 3 transformer based-techniques such as Urduhack, BERT, and XLM-Roberta.

Keywords

Abusive Language Detection, Hate Speech, Label Classification, Versions of BERT, HASOC

1. Introduction

With the advent of social media, anti-social and abusive behavior has become a prominent occurrence online. Undesirable psychological effects of abuse on individuals make it an important societal problem of our time [1]. Pew Research Centre, in its latest report on online harassment [2], revealed that 40% of adults in the United States had experienced abusive behavior online, of which 18% have faced severe forms of harassment, e.g., that of sexual nature. These statistics stress the need for automated detection and moderation systems. Hence, a new research effort on abusive language detection has sprung up in NLP in recent years.

Online communities, social media enterprises, and technology companies are investing heavily and encouraging research in this area by organizing tasks and workshops. One such community is FIRE, which has been actively organizing the HASOC tasks since 2019 [3]. The Urdu language has more than 230 million speakers worldwide with vast social networks and digital media representation.[4] This paper will contain details regarding the subtask A - Abusive language using Twitter tweets in Urdu language of Abusive and Threatening Language Detection Task in Urdu. This is a binary classification task in which participating systems are required to classify tweets into two classes, namely: Abusive and Non-Abusive.

Forum for Information Retrieval Evaluation, December 13-17, 2021, India

- Abusive This Twitter post contains any abusive content.
- Non-Abusive This Twitter post does not contain any abusive or profane content.

2. Related Work

Techniques for abuse detection have gone through several stages of development, starting with extensive manual feature engineering and then turning to deep learning. Early approaches experimented with feature extraction from speech text like a bag of words or n-grams [5], lexical and linguistic features [6] and, and user-specific features, such as age [7]. With the advent of deep learning, the trend shifted, with great work focusing on neural architectures for abuse detection. Initially witnessing an extensive use of CNNs [8] and then moving on to LSTMs [9]. Most recently, the use of pre-trained transformer-based architectures such as BERT [10] has given state-of-the-art results. [11] describe the first shared task for fake news detection in the Urdu language. The dataset consists of news articles from five domains with 900 annotated articles for the training and 400 annotated news articles for the testing part. In this shared task, nine teams submitted their results, and the best performing system achieved an F-score value of 0.90. Authors in [3] introduced a new dataset for classifying threatening and non-threatening language in the Urdu language. The recommended dataset comprises 3,564 tweets manually annotated by human specialists. They applied different models based on Machine and Deep Learning-based techniques. They compared the three forms of text representations. Their research reveals that an MLP classifier with the combination of word n-gram features outperformed other classifiers. [19], [20] has also performed well in the Abusive language detection in the Urdu language.

3. Dataset

The datasets for the tasks are provided by the organizers of HASOC '21[12] [1] and the code is available in the github repository [2] The data consists of tweets in Urdu annotated for a binary classification task: Abusive, Non-Abusive. Abusive - This Twitter post contains any abusive content. Non-Abusive - This Twitter post does not contain any abusive or profane content. Table 1 lists the statistics of the dataset. According to Twitter, the definition describes abusive comments toward individuals or groups to harass, intimidate, or silence someone else's voice. The dataset was collected and annotated in Natural Language and Text Processing laboratory at the Center of Computing Research of Instituto Politécnico Nacional, Mexico, by Ph.D. candidate Maaz Amjad, a native Urdu-speaker [13].

4. Proposed Techniques and Algorithms

The paper describes various approaches and draws out a comparison between them. The first approach extracts N-grams features from the tweets, which are weighted according to TF-IDF values. Then, models using machine learning algorithms are trained upon these features. Fig

[1]https://www.Urduthreat2021.cicling.org/home
[2]https://github.com/Kalra-Sakshi/Abusive-HASOC.git

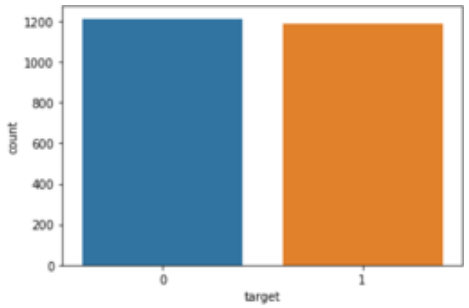

Figure 1: Training set distribution in the Urdu Dataset (1 is Abusive,0 is Non-Abusive)

2 shows the proposed architecture using machine learning-based techniques such as Logistic Regression, Random Forest Classifier, and Support Vector Machine. The second approach uses pre-trained transformer-based models and their associated tokenizers. Three pre-trained models are used for this task. Urduhack Roberta-Urdu-small [14]: Trained on news data from Urdu news resources in Pakistan BERT (checkpoint : bert-base-multilingual-cased [15]) : Trained on 104 different languages XLM-Roberta [16]: Trained on 2.5TB of newly created clean CommonCrawl data in 100 languages. Fig 3 shows the proposed architecture using transformer-based techniques.

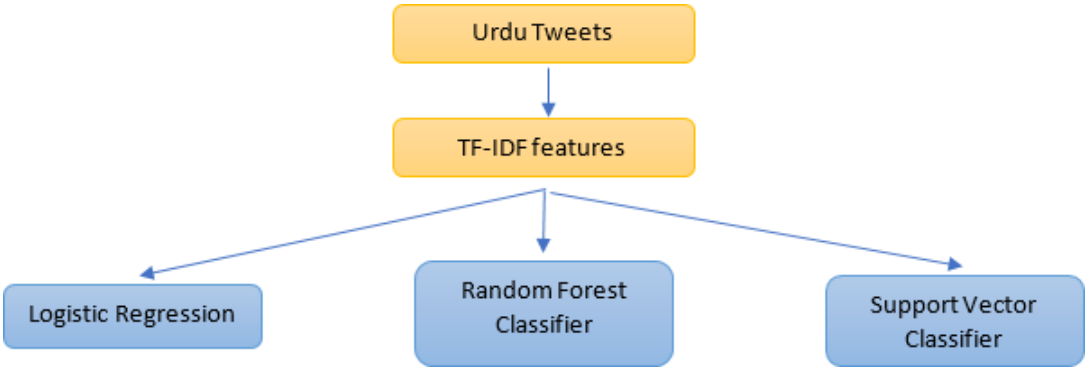

Figure 2: Proposed Architecture Based on Various Machine Learning based Algorithms

5. Experimental Work

The primary evaluation metric for evaluating the applied machine-Learning and Transformer based models is the F1 score, and ROC AUC is the secondary evaluation metric used.

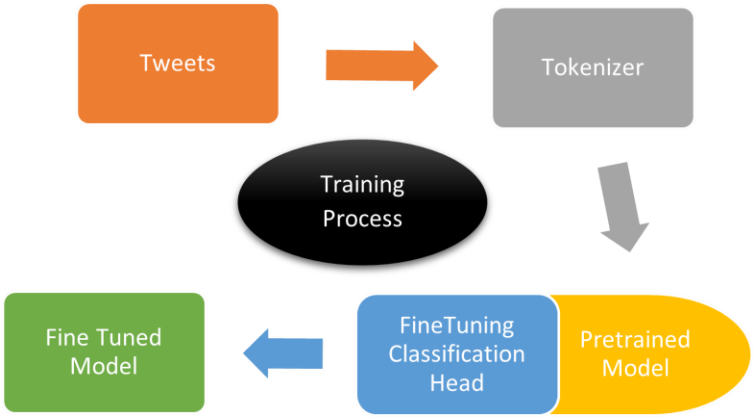

Figure 3: Proposed Architecture based on Various Transformer based Algorithms

Table 1
Hyperparameters used in the task of Abusive Language Detection in the Urdu Language

Hyperparameter	Description
Learning Rate	5e-5-5e-6
Number of Epochs	2,3,4
Batch Size	2,4,8,16

5.1. Logistic Regression, Support Vector Classifier, Random Forest Classifier

Here, we use three machine learning algorithms: Logistic Regression, Support Vector Classifier, and Random Forest Classifier available in the 'scikit-learn' package. While training, a 5-fold grid search is performed on the entire train dataset to find the best set of hyperparameters.

5.2. TRANSFORMER BASED MODELS

For initial experimentation, pre-processing is carried out in Normalization [17], but results without the Normalization are significantly better. Hyper-parameter tuning for the models is carried out using RAY TUNE. Population-Based Training scheduler is used for all three models, with train batch size in [2,4,8,16]. The learning rate was set to a uniform log distribution between 5e-6 and 5e-5. Table 1 and 3 lists the Hyperparameter description. For the multilingual Bert and Urduhack model, train epochs are selected between 2,3,4. Given the large size of XML-Roberta, train epochs are fixed at 2. Finally, soft voting is carried out, taking the average of each model's output scores and predicting the target class.

Table 2
Results obtained on the test set were made public at the end of the competition

Algorithm	Weighted-F1	ROC-AUC
Logistic Regression	0.8038	0.8927
SVM	0.8036	0.8925
Random Forest Classifier	0.7899	0.8390

Table 3
Hyperparameters used in the Urdu Threatening Language Detection

Hyperparameter	Value
Learning rate	4.4391e-05
Number of train epochs	2
Training batch size	4

Table 4
Model performances on the public and private Data

Evaluation Parameters	Public	Private
F1 Score	0.8393	0.8685
ROC-AUC	0.9340	0.9350

Table 5
Hyperparameters of three-Transformer based models without normalization of the tweets

Model	Learning Rate	Number of Train Epochs	Train Batch Size
Urduhack	1.976e-05	2	16
BERT	8.1528e-06	2	4
XLM-Roberta	2.09411e-05	2	8

6. Results and Evaluations

The following results are obtained on the test set made public at the end of the competition and described in Table 2. [18] All models used the best parameters obtained through a 5-fold grid search. Submission for the competition has been made using the Urduhack model with Normalization and results are listed in Table 4. Further soft voting is carried out using the three transformer-based models without Normalized the tweets, using the following parameters listed in Table 5. The following results are obtained on the entire test set listed in Table 6.

7. Conclusions and Future Work

This paper started with experimentation using classical machine learning models such as Logistic Regression, SVM, and Random Forest Classifier. We then moved on to leveraging

Table 6
Results obtained on the entire test set using Soft-Voting Technique

Result	Weighted F1	ROC-AUC
Soft-Voting	0.86424	0.93607

recent advances in large-scale Transformer-based pre-trained language models. The larger pre-trained models still outperform the classical models while performing well. Pre-processing performed using the UrduHack library did not necessarily yield better results, which could lead to why punctuations and diacritics add information valuable to Abuse detection. Our model is getting 0.9340 accuracies on the public data with normalization of the tweets and 0.9360 without normalization. For future work, we can try out different multilingual transformer-based models to get a more robust model.

References

[1] E. R. Munro, The protection of children online: a brief scoping review to identify vulnerable groups, Childhood Wellbeing Research Centre (2011).

[2] M. Duggan, Online harassment 2017 (2017).

[3] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, A. Gelbukh, Threatening language detecting and threatening target identification in urdu tweets, IEEE Access (2021).

[4] Fire 2021, urduabuse2021, URL: https://www.urduthreat2021.cicling.org/home#h.5r0apwv33dhk.

[5] A. Gaydhani, V. Doma, S. Kendre, L. Bhagwat, Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach, arXiv preprint arXiv:1809.08651 (2018).

[6] N. D. Gitari, Z. Zuping, H. Damien, J. Long, A lexicon-based approach for hate speech detection, International Journal of Multimedia and Ubiquitous Engineering 10 (2015) 215–230.

[7] M. Dadvar, D. Trieschnigg, R. Ordelman, F. de Jong, Improving cyberbullying detection with user context, in: European Conference on Information Retrieval, Springer, 2013, pp. 693–696.

[8] J. H. Park, P. Fung, One-step and two-step classification for abusive language detection on twitter, arXiv preprint arXiv:1706.01206 (2017).

[9] A. Bisht, A. Singh, H. Bhadauria, J. Virmani, et al., Detection of hate speech and offensive language in twitter data using lstm model, in: Recent trends in image and signal processing in computer vision, Springer, 2020, pp. 243–264.

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

[11] M. Amjad, G. Sidorov, A. Zhila, A. F. Gelbukh, P. Rosso, Overview of the shared task on fake news detection in urdu at fire 2020., in: FIRE (Working Notes), 2020, pp. 434–446.

[12] Urdufake-data, URL: https://www.cicling.org/urdufake-data/2021/.

[13] Urdu-hack-soc2021data, URL: https://ods.ai/competitions/urdu-hack-soc2021/data.

[14] Urduhack library, URL: https://github.com/urduhack/urduhack.
[15] m-bert(multilingual bidirectional encoder representations from transformers), ???? URL: https://huggingface.co/transformers/multilingual.html#bert.
[16] (xlm-roberta),URL:https://huggingface.co/transformers/multilingual.html#xlm roberta.
[17] Normalization, URL: https://docs.urduhack.com/en/stable/reference/normalization.html.
[18] URL:https://drive.google.com/file/d/19G9ntBaDCGnf765ELctEX2ZPmbCvyy1G/view?usp=sharing.
[19] M.Amjad, A.Zhila, O.Vitman, S.Butt, H.I.Amjad, G.Sidorov, A. Gelbukh. "UrduThreat@ FIRE2021: Shared Track on abusive threat Identification in Urdu." In Forum for Information Retrieval Evaluation. (2021).
[20] M.Amjad, N.Ashraf, G.Sidorov, A.Zhila, L. Chanona-Hernandez, A. Gelbukh. "Automatic Abusive Language Detection in Urdu Tweets." Acta Polytechnica Hungarica. (2021)

A. Online Resources

The implementation of different pre-trained BERT-models are available at

- Huggingface.