Abusive and Threatening Language Detection from Urdu Social Media Posts: A machine learning approach

Abhinav Kumar[1], Sunil Saumya[2] and Pradeep Kumar Roy[3]

[1]*Department of Computer Science & Engineering, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India*
[2]*Department: Computer Science & Engineering, Indian Institute of Information Technology Dharwad, India*
[3]*Indian Institute of Information Technology Surat, Gujarat*

Abstract

Urdu is spoken by approximately 230 million people throughout the world, and it has a sizable social media and digital media following. However, none of the possible efforts to detect abusive and threatening postings from Urdu social media posts has been suggested to the best of our knowledge. This study explores the usability of conventional machine learning and deep learning models to identify abusive and threatening messages on Urdu social media. The proposed ensemble-based machine learning model performed promising in the identification of abusive and threatening language from Urdu social media posts. The suggested ensemble-based model achieved a weighted F_1-score of 0.81, the accuracy of 0.81, a ROC of 0.90 for abusive language identification, and a weighted F_1-score of 0.81, accuracy of 0.85, and ROC of 0.81 for threatening language identification.

Keywords

Abusive content, Social media, Deep learning, Hate speech,

1. Introduction

The impact of social media platform misuse has grown in tandem with the expansion and prominence of these platforms [1, 2, 3]. Numerous posts, in the example, contain abusive language directed at specific users, therefore detracting from the communication experience on such platforms, while others contain genuine threats that might put platform users in danger [4, 5, 6, 7, 8]. Several works [9, 10, 11, 12, 13, 14, 15] have been proposed by researchers to identify hate speech from English, Hindi, and code-mixed Dravidian social media posts. Kumari and Singh [13] presented a model based on convolutional neural networks for detecting hate, obscenity, and abusive language in English and Hindi tweets. To recognize hatred, offensive, and profanity in English, Hindi, and German tweets, Mishra and Pal [14] developed an attention-based bidirectional long-short-term memory network. Mujadia et al. [15] developed an ensemble-based model comprised of a support vector machine, random forest, and Adaboost classifiers to identify hate content in tweets written in English, Hindi, and German. Saumya et al. [12] experimented with several conventional machine learning and deep learning models for the

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

abhinavanand05@gmail.com (A. Kumar); sunil.saumya@iiitdwd.ac.in (S. Saumya); pradeep.roy@iiitsurat.ac.in (P. K. Roy)

CEUR Workshop Proceedings (CEUR-WS.org)

hate speech identification from Dravidian social media posts. They found character N-gram features with conventional machine learning classifiers performing better than the complex deep learning models.

Urdu is spoken by over 230 million people worldwide, and it has a significant social media and digital media presence. But to the best of our knowledge, none of the potential work has been proposed to identify abusive and threatening posts from Urdu social media posts. This paper explores the usages of different conventional machine learning and deep learning models for the identification of abusive and threatening Posts from Urdu social media posts. The proposed models are validated with the dataset published in CICLing 2021 track FIRE 2021 workshop [16]. The task is divided into two subtasks: (i) Sub-task A is concerned with identifying the abusive language in Twitter tweets written in Urdu. This is a binary classification task that must classify tweets into one of two categories: abusive or non-abusive, (ii) Sub-task B focuses on identifying Threatening language in Urdu-language tweets. This is a binary classification task in that one must categorize tweets into two categories: Threatening and Non-Threatening.

The rest of the sections are organized as follows: Section 2 discusses the proposed methodology in detail. Section 3 lists the findings and finally the paper concluded in Section 4.

2. Methodology

This section discusses the proposed methodology in detail. For Task-A: Abusive language identification, we submitted four different models: (i) Ensemble (SVM + LR + RF) (Ensemble of Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF)), (ii) Dense Neural Network (DNN), (iii) Ensemble (Three variant of BERT ((a) BERT bert-base-arabic[1], (b) distilbert-base-multilingual-cased[2], and (c) bert-base-multilingual-cased[3]), and (iv) Support Vector Machine (SVM).

To provide input to the models, one-to-six gram character-level TF-IDF features were used for Ensemble (SVM + LR + RF), Dense Neural Network (DNN), and SVM models. As the deep learning-based models are very sensitive to the chosen hyper-parameters, we performed extensive experiments to choose the best-suited hyper-parameters for the proposed models. After extensive experiments by varying the number of layers, learning rate, batch size, epochs, loss function, and optimizer, for dense neural network (DNN) model, we found four layers with 4,096, 1,024, 128, and 2-neurons performed best with a dropout rate of 0.2, binary cross-entropy as a loss function and Adam as the optimizer, a batch size of 32, and a learning rate of 0.001.

Similarly, for the Ensemble (Three variants of BERT), we found a learning rate of $2e^{-5}$ and a batch size of 32 with 20 epochs of training performed best. After training each BERT model individually, the prediction probability for both the classes was averaged class-wise to get the final class prediction.

In the case of Task-B of threatening language identification, we submitted two models: (i) Ensemble (SVM + LR + RF) and (ii) AdaBoost. In this case, also, we used one-to-six gram character-level TF-IDF features to train the system.

[1] https://huggingface.co/asafaya/bert-base-arabic

[2] https://huggingface.co/distilbert-base-multilingual-cased

[3] https://huggingface.co/bert-base-multilingual-cased

Table 1
Results of different models for Abusive and Threatening language identification

Task	Models	Class	Precision	Recall	F_1-score	Accuracy	ROC
Task-A (Abusive)	Ensemble (SVM + LR + RF)	Not-Abusive	0.77	0.86	0.82	0.81	0.90
		Abusive	0.85	0.76	0.80		
		Weighted Avg.	0.81	0.81	0.81		
	DNN	Not-Abusive	0.75	0.81	0.78	0.78	0.83
		Abusive	0.81	0.74	0.77		
		Weighted Avg.	0.78	0.78	0.78		
	Ensemble (Three variant of BERT)	Not-Abusive	0.79	0.77	0.78	0.79	0.85
		Abusive	0.78	0.80	0.79		
		Weighted Avg.	0.79	0.79	0.79		
	SVM	Not-Abusive	0.76	0.87	0.81	0.80	0.89
		Abusive	0.86	0.74	0.80		
		Weighted Avg.	0.81	0.80	0.80		
Task-B (Threatening)	Ensemble (SVM + LR + RF)	Not-Threatening	0.85	0.99	0.91	0.85	0.81
		Threatening	0.77	0.22	0.35		
		Weighted Avg.	0.84	0.85	0.81		
	AdaBoost	Not-Threatening	0.85	0.96	0.90	0.83	0.74
		Threatening	0.60	0.26	0.37		
		Weighted Avg.	0.81	0.83	0.81		

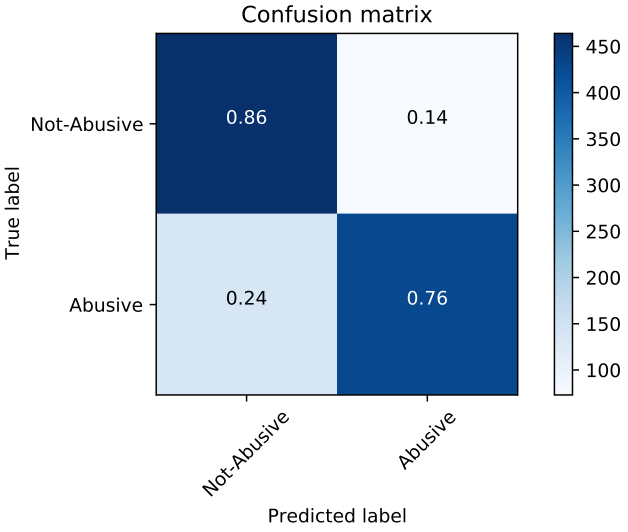

Figure 1: Confusion matrix of ensemble of SVM, LR, and RF classifiers for Abusive language detection

In the case of conventional machine learning classifier, Sklearn python library[4] is used with the default parameters. To implement dense neural network, Keras[5] with TensorFlow[6] as a back-end whereas for BERT models, Huggingface[7] library is used.

[4] https://scikit-learn.org/stable/
[5] https://keras.io/
[6] https://www.tensorflow.org/
[7] https://huggingface.co/

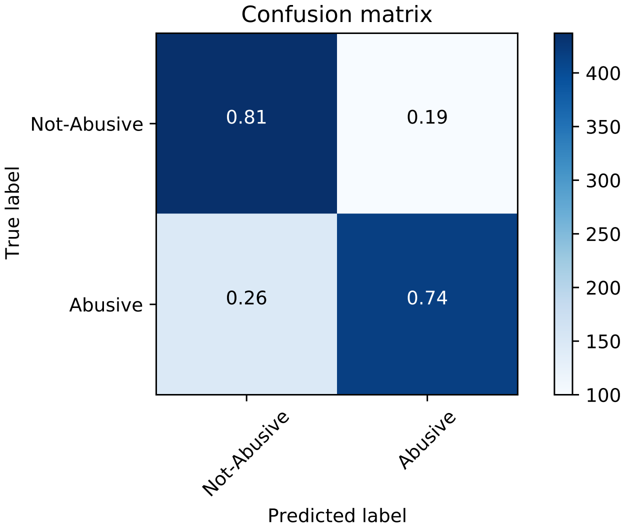

Figure 2: Confusion matrix of Dense Neural Network (DNN) model for Abusive language detection

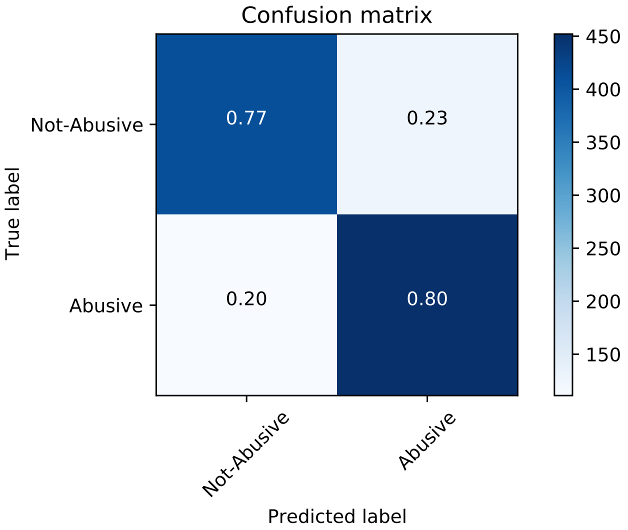

Figure 3: Confusion matrix of BERT ensemble model for Abusive language detection

3. Results

The performance of the proposed model is measured in terms of precision, recall, $F_1 - score$, accuracy, and ROC value. The results for all the submitted models for Task-A and Task-B are listed in Table 1. For Task-A, the proposed Ensemble (SVM + LR + RF) model achieved a weighted precision, recall, and F_1-score of 0.81, an accuracy of 0.81, and a ROC value of 0.90. The confusion matrix for Ensemble (SVM + LR + RF) model can be seen in Figure 1. The DNN

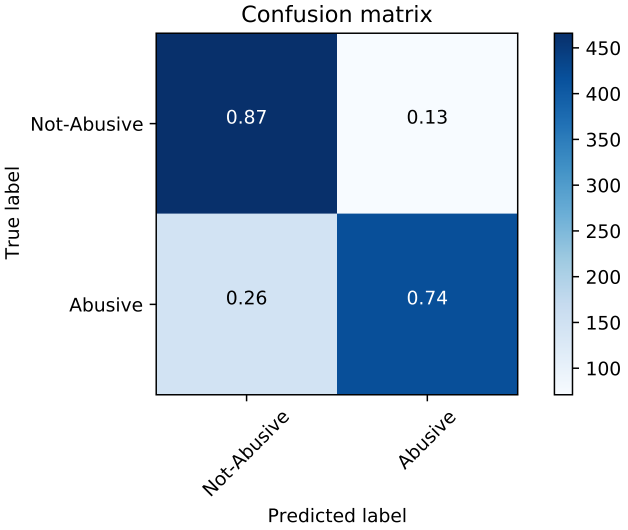

Figure 4: Confusion matrix of SVM classifier for Abusive language detection

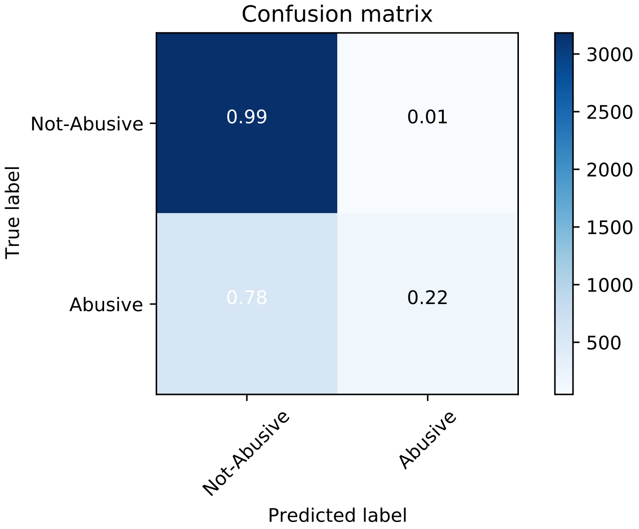

Figure 5: Confusion matrix of ensemble of SVM, LR, and RF classifiers for Threatening language detection

model achieved a weighted precision, recall, and F_1-score of 0.78, an accuracy of 0.78, and a ROC value of 0.83. The confusion matrix for the DNN model can be seen in Figure 2. The Ensemble (Three variants of BERT) model achieved a weighted precision, recall, and F_1-score of 0.79, an accuracy of 0.79, and a ROC value of 0.85. The confusion matrix for the Ensemble (Three variants of BERT) model can be seen in Figure 3. The SVM classifier achieved a weighted precision of 0.81, recall of 0.80, F_1-score of 0.80, an accuracy of 0.80, and a ROC value of 0.89.

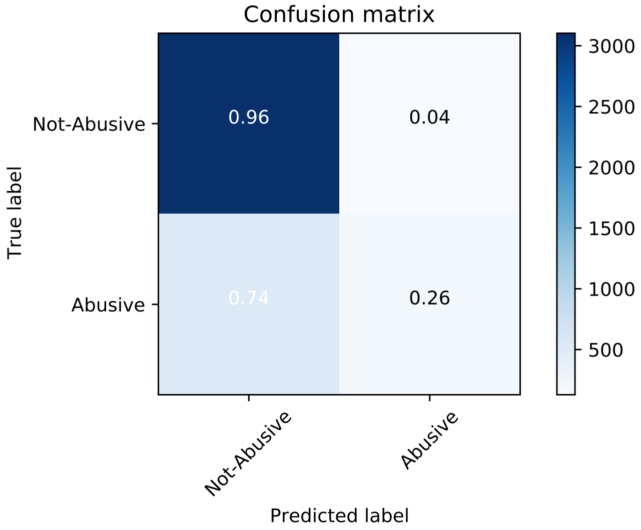

Figure 6: Confusion matrix of AdaBoost classifier for Threatening language detection

The confusion matrix for the SVM classifier can be seen in Figure 4.

For Task-B, the Ensemble (SVM + LR + RF) model achieved a weighted precision of 0.84, recall of 0.85, F_1-score of 0.81, accuracy of 0.85, and a ROC value of 0.81. The confusion matrix for Ensemble (SVM + LR + RF) model for threatening language identification can be seen in Figure 5. The AdaBoost classifier achieved a weighted precision of 0.81, recall of 0.83, F_1-score of 0.81, accuracy of 0.83, and a ROC of 0.74. The confusion matrix for the AdaBoost classifier for Threatening language identification can be seen in Figure 6.

4. Conclusion

Urdu is widely used on social media and in the digital world. Users publish a disproportionate number of abusive and threatening Urdu social media posts. Various ensemble-based models based on machine learning and deep learning are proposed in this study. The suggested Ensemble-based approach, which combines support vector machines, logistic regression, and random forest classifiers, outperformed all other models in terms of identifying abusive and threatening language in Urdu social media postings. For abusive language identification, the proposed ensemble-based model achieved a weighted $F1$-score of 0.81, accuracy of 0.81, and ROC of 0.90, while for threatening language identification, it achieved a weighted $F1$-score of 0.81, accuracy of 0.85, and ROC of 0.81.

References

[1] J. P. Singh, A. Kumar, N. P. Rana, Y. K. Dwivedi, Attention-based lstm network for rumor veracity estimation of tweets, Information Systems Frontiers (2020) 1–16.

[2] A. Kumar, J. P. Singh, Location reference identification from tweets during emergencies: A deep learning approach, International journal of disaster risk reduction 33 (2019) 365–375.

[3] A. Kumar, J. P. Singh, Y. K. Dwivedi, N. P. Rana, A deep multi-modal neural network for informative twitter content classification during emergencies, Annals of Operations Research (2020) 1–32.

[4] H. Rizwan, M. H. Shakeel, A. Karim, Hate-speech and offensive language detection in roman Urdu, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2512–2522.

[5] M. M. Khan, K. Shahzad, M. K. Malik, Hate speech detection in roman Urdu, ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 20 (2021) 1–19.

[6] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butt, H. Imam Amjad, O. Vitman, A. Gelbukh, Overview of the shared task on threatening and abusive detection in Urdu at Fire 2021, in: FIRE (Working Notes), CEUR Workshop Proceedings, 2021.

[7] M. Amjad, N. Ashraf, G. Sidorov, A. Zhila, L. Chanona-Hernandez, A. Gelbukh, Automatic abusive language detection in Urdu tweets, Acta Polytechnica Hungarica (2021).

[8] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, A. Gelbukh, Threatening language detection and target identification in Urdu tweets, IEEE Access 9 (2021) 128302–128313.

[9] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@ HASOC-FIRE2020: Fine tuned BERT for the hate speech and offensive content identification from social media., in: FIRE (Working Notes), 2020, pp. 266–273.

[10] A. Kumar, S. Saumya, J. P. Singh, NITP-AI-NLP@ HASOC-Dravidian-CodeMix-FIRE2020: A machine learning approach to identify offensive languages from Dravidian code-mixed text., in: FIRE (Working Notes), 2020, pp. 384–390.

[11] P. K. Roy, A. K. Tripathy, T. K. Das, X.-Z. Gao, A framework for hate speech detection using deep convolutional neural network, IEEE Access 8 (2020) 204951–204962.

[12] S. Saumya, A. Kumar, J. P. Singh, Offensive language identification in Dravidian code mixed social media text, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 36–45.

[13] K. Kumari, J. P. Singh, AI ML NIT Patna at HASOC 2019: Deep learning approach for identification of abusive content., in: FIRE (Working Notes), 2019, pp. 328–335.

[14] A. Mishra, S. Pal, IIT Varanasi at HASOC 2019: Hate speech and offensive content identification in Indo-European languages., in: FIRE (Working Notes), 2019, pp. 344–351.

[15] V. Mujadia, P. Mishra, D. M. Sharma, IIIT-Hyderabad at HASOC 2019: Hate speech detection., in: FIRE (Working Notes), 2019, pp. 271–278.

[16] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butt, H. Imam Amjad, O. Vitman, A. Gelbukh, UrduThreat@ FIRE2021: Shared track on abusive threat identification in Urdu, in: Forum for Information Retrieval Evaluation, 2021.