Urdu Abusive Language Detection using Machine Learning

Muhammad Owais Raza[1], Qaisar Khan[2], Ghulam Muhammad Soomro[3]

[1] *Mehran University of Engineering and Technology, Indus Hwy, Jamshoro, Sindh 76062, Pakistan*
[2] *Sunway University, 5 Jalan University, Bandar Sunway, 47500 Petaling Jaya, Selangor, Malaysia*
[3] *Mehran University Institute of Science, Technology and Development, Indus Hwy, Jamshoro, Sindh 76062, Pakistan*

Abstract

The growing popularity of user-generated material on social media has increased the quantity of offensive language used online. The tendency of user-generated material on social media is growing, giving rise to offensive language on these platforms. The offensive language negatively impacts individuals and affects society as a whole, which is why it is a dire need of time to identify vulgar remarks in languages used online. 'Urdu' is one of the many languages used on the internet that faces the same issue. Manually labeling the text as abusive on social media platforms is unattainable due to the production of a large amount of daily content. Therefore, automation (machine learning) is used to create the solution. This study uses machine learning algorithms, namely logistic regression, bagging algorithms, decision trees, and artificial neural networks (ANN), to detect abuse in the text. The F1 score is used as the primary metric, along with accuracy, precision, recall, and AUC-ROC, to measure the performance. Based on the evaluation, the bagging and logistic regression perform equally with an 83% F1 score. However, logistic regression is better for this use case because it is computationally less expensive and requires less effort than the bagging classifier.

Keywords

Machine Learning, NLP, Urdu Abuse Detection, Python, Logistic Regression

1. Introduction

Historically, mass communication mediums were utilized under ethical and moral obligations dictated by societal standards. In this digital age, the wide acceptance of social media continues to be fueled by the prevalence of internet connection and mobile technologies, particularly smartphones and tablets [2]. The growing opportunities to express opinions online have given a high rise in hate speech and offensive language. Studies show that people may use offensive language online that affects other people's feelings [4]. The internet's secrecy has a detrimental effect on the population, encouraging obscene language, disparaging phrases, poisonous, unpleasant, and abusive language on the web, specifically social media. This derogatory content on the internet can be aggressive and harmful. It can erode people's self-esteem, inflict suicidal thoughts, and compel them to wipe out their social media existence. Due to this rise of cyberterrorism, cyberbullying, and widespread usage of derogatory language on the internet, identifying hate speech has become a critical component of anti-bullying measures for social media platforms [5]. The manual detection and removal of hate speech and undesirable information is a time-consuming process, owing to the vastness of the web and the growing number of internet users. The work gets harder considering the anonymity of online users. Hence, it is high time for technologies and approaches to rapidly detect abusive language on social media platforms and eradicate the spread of hate speech.

Forum for Information Retrieval Evaluation, December 13-17, 2021, India
EMAIL: owais.leghari@hotmail.com (A. 1); qaisar.k@imail.sunway.edu.my (A. 2); soomrogm95@gmail.com (A. 3)
ORCID: 0000-0002-3065-385X (A. 1); 0000-0001-7903-0277 (A. 2); 0000-0002-9327-9674 (A. 3)
https://github.com/owais4321/Urdu-Abusive-Language-Detection-using-Machine-Learning


Urdu is South Asia's resource-scarce language [6]. Compared to resource-rich languages such as English, a few annotated corpora are available for different NLP applications. The lack of linguistic resources such as stemmers and annotated corpora complicates and inspires study. Studying abusive language detection in Urdu [23] presents several difficulties. There is a dearth of sufficient annotated corpora and Urdu text preprocessing tools. This study presents different techniques to detect the abusive language in Urdu using different machine learning techniques and discusses the challenges and solutions.

2. Related Work

The rise in the use of social media has given birth to numerous problems, one of which is abusive behavior on social media platforms. Each platform has its policies to detect such behavior. For example, Twitter defines abusive behavior as an attempt to intimidate, harass, or silence someone's voice. Based on this definition, Twitter can classify a tweet as abusive or non-abusive. Recently, the computational linguistics community has focused on detecting abusive language and hate speech from various online social media platforms, such as Twitter [7], [8]. Early identification of many social abnormalities, such as hate speech, cyberbullying [9], [10], trolling [11], false news [12], rumor [13], fake profile identification [14], and sexism [15], has been a current trend in social media-based research. In [22] researchers have detected threatening tweets in urdu. Different researchers used different techniques to identify inappropriate text online. Researchers in [16] employ a variety of machine learning approaches that include support vector machines, decision trees, instance-based and rule-based, and algorithms from the WEKA toolkit used to identify bully-specific language patterns and built rules for automatically detecting cyberbullying content. In [17], researchers use a variety of classifiers to detect cyberbullying, including support vector machines, naive bayes, random forest, JRip, J48, k-nearest neighbors, sentence pattern extraction architecture (SPEC), and convolutional neural network (CNN). The results indicate that CNN outperforms other classifiers by over 11% in F-score. However, a challenge in this study is that there is no limitation for the language on these social media platforms. It is easier to create a machine learning model to detect abusive language for English because of resources. However, when it comes to languages like Urdu, the resources are low, and the process is laborious.

3. Dataset

The dataset is adapted from HASOC abusive and threatening language detection in Urdu competition [20][21]. The dataset was gathered and labeled in the natural language and text processing laboratory at the center for computing research of Instituto Politécnico Nacional. The collector and annotator of the datasets are native Urdu speakers. The dataset used in the study contains two columns tweet and target with 2400 rows. Each row represents a tweet and its corresponding labels. There are two labels, 0 and 1, '0' represents a neutral text while '1' shows abusive text. Table 1 shows the distribution of the dataset on the label count. We have two labels in the dataset, abusive and non-abusive, and 1187 abusive tweets and 1213 non abusive tweets, which balances the dataset.

Table 1
Label Count

Label	Count
Abusive	1187
Non-Abusive	1213

4. Algorithms

The problem we are tackling in this research is a binary classification problem, and the following algorithms were used:

1. Logistic Regression
2. Decision Tree
3. Bagging classifier
4. Neural Network

4.1. Logistic Regression

Logistic regression is a powerful technique of simulated results of a binary classification. It is used to allocate data to a discrete label, and being a classification method, it relies on probability [16]. The logistic regression function is represented by equation 1.

$$g(z) = \frac{1}{1 - e^{z}} \tag{1}$$

The prediction function is represented by equation 2

$$h\theta(x) = g(\theta^{T}x) = \frac{1}{1 + e^{-\theta^{T}x}} \tag{2}$$

The value of θ has a unique significance; it indicates the likelihood that $h\theta(x)$ is 1 [17].

In this study, all the hyperparameters for the logistic regression are kept default because of getting the best results.

4.2. Decision Tree

The decision tree is a technique for successfully supervising inductive learning through the generation of rules from data. Typically, one event might trigger two or more subsequent events, each with a distinct outcome. The decision tree's structure is top-down as a result of this characteristic. This structure resembles a flowchart. Every branch of a tree corresponds to a new decision result. The children node on each node corresponds to the corresponding attribute test. This child node's ID, generated from decision-making algorithms, is passed on [18].

4.3. Bagging Classifier

The bagging technique (bootstrap aggregation) generates a collection of classifiers. A bootstrapped duplicate of the original dataset is created for each classifier by randomly selecting N instances with replacement. When a new input is desired to be classified, the number of classifiers that anticipate the instance's class value is counted for every label state. The number of votes and the state with the most votes are projected to win the instance. In this study, we are using bagging for bootstrapping different logistic regression classifiers.

4.4. Artificial Neural Network

A neural network comprises a linked group of artificial neurons that analyses data in a connectionist fashion. In general, an ANN is a self-organizing system that fine-tunes its organization in response to external or internal data that flows through the network throughout the learning process. They are typically used to describe complex connections between inputs and outputs or to deduce patterns from data. ANN has been successfully utilized in a variety of applications. For instance, ANNs have been successfully utilized in predictions, handwritten character recognition, and assessing home values [19].

5. Methodology

The methodology of this study is shown in figure 1. The methodology for the research consists of 6 steps:

1. **Importing Dataset:**

The first step is to import the respective dataset, so the abuse language dataset was imported.

2. **Cleaning Dataset:**

After importing the dataset, ambiguities are searched. Then, tokenizing takes place, removing any punctuation marks, unique characters, and numbers in the data using regular expression. Next is to remove any stopwords, which are high-frequency words with low semantic importance. The remaining data is cleaned data.

3. **Extracting Features:**

To extract features, TF-IDF is used. The TF-IDF algorithm (term frequency-inverse document frequency) is an enhancement to the DF technique. It is a type of statistical technique used to determine the significance of a word within a file collection. The significance of a word is related to its frequency of occurrence in the text and inversely related to its frequency of occurrence in the total document collection.

$TF_{i,j}$ is the rate of input W_i in document X_j, as shown by equation (3)

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

IDF_i is represented by equation 4

$$IDF_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|} \tag{4}$$

Here $|D|$ is the total number of files and d_j indicates the total number of occurrences of a word. $TFIDF_{i,j}$ is represented by equation 5.

$$TFIDF_{i,j} = TF_{i,j} \, X \, IDF_i \tag{5}$$

4. **Splitting Dataset:**

Using test train split for creating two sets of datasets—one for training and the other for testing. The train test ratio is 75:25. That is, 75% of the dataset is used for training and 20% for evaluating results.

5. **Apply Machine Learning Algorithm:**

In this step, machine learning algorithms are applied to the training set to create a classification model. The algorithms used in this study are logistic regression, decision tree, bagging ensemble classifier, and ANN.

6. **Evaluate Machine Learning Model:**

The last step is where inference is performed on 25% of the dataset to determine how well the classifiers are performed. The classification metrics used in this study are accuracy, precision, recall, F1 score, and AUC ROC.

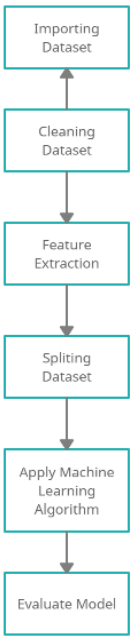

Figure 1: Methodology Flowchart

6. Results

6.1. Evaluation Metrics

The evaluation parameter taken for this study is accuracy, precision, recall, F1 score, and AUC ROC. To decide the best-performing model, the key parameter used is the F1 score.

6.1.1. Accuracy

Accuracy is an evaluation parameter that shows the degree to which a classifier fits all classes. It is helpful since it treats all classes equally. It is measured as the fraction of correct predictions to all predictions. Equation 6 represents accuracy in mathematical terms.

$$\frac{I_p + I_n}{(T_p + T_n + F_p + F_n)} \tag{6}$$

T_p in equation 6 represent true positive, T_n is true negative, F_p is false positive, and F_n is false negative.

6.1.2. Precision

The precision is determined as the fraction of Positive samples accurately classified to all positive cases classified. Precision is a measure that shows a model's predictive accuracy in classifying a sample as positive. It is determined by equation 7.

$$\frac{I_p}{(T_p + F_p)} \tag{7}$$

T_p in equation 7 represent true positive and F_p is false positive.

6.1.3. Recall

The recall is calculated as the ratio of correctly classified positive samples to all accessible, positive occurrences. The recall parameter specifies the model's ability to detect positive samples. The higher the recall, the more positive samples are discovered. Equation 8 represents mathematical representation.

$$\frac{{}^{1}p}{(T_p + F_n)} \tag{8}$$

T_p in equation 6 represent true positive and F_n is false negative.

6.1.4. F1 Score:

The F1 score is a statistic that indicates how accurate a model is on a given dataset. It is used to assess binary classification systems that categorize examples as either positive or negative. Equation 9 represents mathematical representation.

$$\frac{2 * precision * recall}{precision + recall} \tag{9}$$

6.1.5. AUC ROC:

The AUC - ROC curve is a benchmarking tool for classification problems using a variety of threshold values. The receiver operating characteristic (ROC) curve denotes the extent or measure of separability. In contrast, the area under the curve (AUC) indicates the level or degree of separability. It indicates the degree to which the model is adept at differentiating between classes. The larger the AUC, the more accurately the model predicts 1 class as '1' and 0 class as '0'. For example, the greater the AUC, the more accurate the model discriminates between abusive and non-abusive texts.

6.2. Accuracy, Precision, and Recall

Three fundamental parameters for deciding which classifier performed best are accuracy, precision, and recall. The classifier was created using all four algorithms. Table 2 shows the accuracy, precision, and recall value for all the algorithms.

Table 2
Accuracy, Precision, and Recall Evaluation Table

Algorithm	Accuracy	Precision	Recall
Decision Tree	74.6	71.5	83.7
Logistic Regression	83.6	89.4	77.1
Bagging Classifier	83.6	88.5	78.1
ANN	79.1	74.1	70.5

According to the table, the best performing algorithm is bagging that showed an accuracy of 83.6, the same as logistic regression. However, it outperformed logistic regression in precision and recall values having 88.5 and 78.1, respectively. The least performing model was a decision tree with accuracy, precision, and recall of 74.6, 71.5, and 83.7, respectively. There is not much difference between logistic regression and bagging classifier, which shows the use of extra effort in bagging classifier is not worth it. The ANN was the second least performing model with 79.1%, 74.1%, and 70.5% accuracy, precision, and recall.

6.3. F1 Score:

The main parameter used in this study is the F1 score, which is considered a reliable metric in classification tasks due to the involvement of both precision and recall. Figure 2 represents the F1 score for all the algorithms. F1 score for bagging and logistic regression are the same, 83%. Due to the extra effort being put into the bagging classifier, logistic regression is considered the better choice for the task. It provides the same F1 score with less effort.

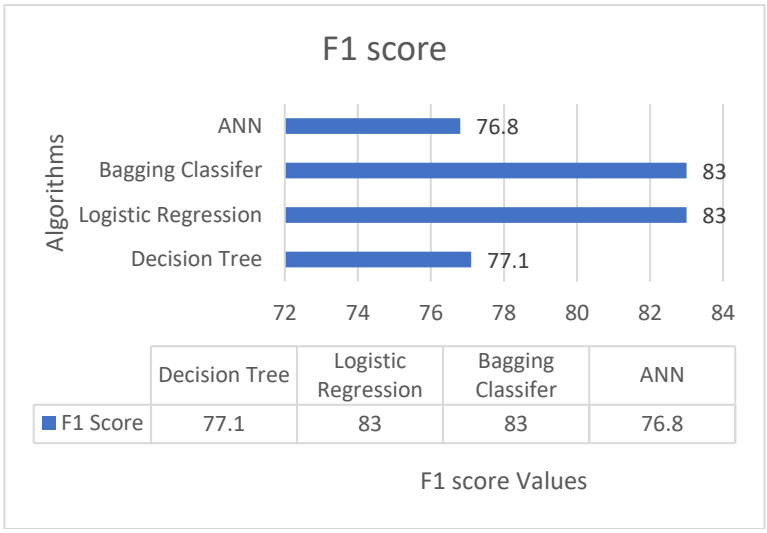

Figure 2: F1 score Bar Chart

6.4. AUC ROC:

The area under the curve for the receiver operation curve is an important parameter to judge the algorithm's performance. The AUC ROC values for each of the algorithm is shown in Figure 3. Looking at figure 3, we can see that bagging has the best value of 91.4, followed by logistic regression with 90.7%, which is not a significant difference compared to the extra efforts put into the bagging classifier. ANN also proved to be a good algorithm with a value of 87%. The algorithm with the lowest value of AUC ROC is the decision tree having 75.5% AUC ROC.

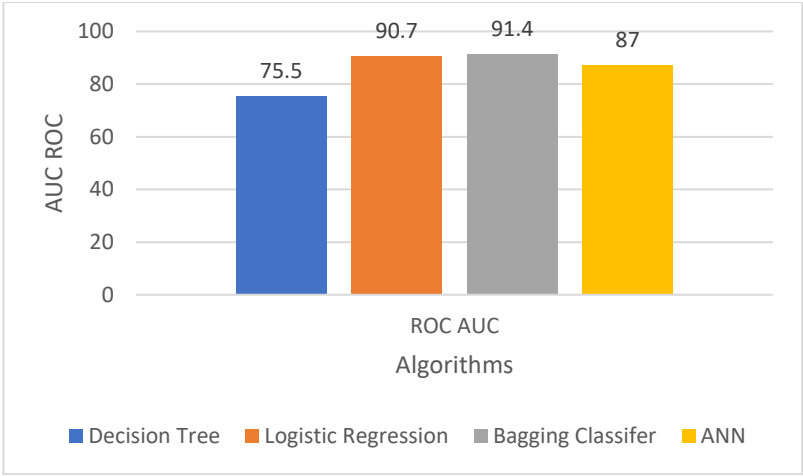

Figure 3: AUC ROC Bar Chart

6.5. ROC Curve:

The receiver operating characteristic (ROC) curve illustrates the relationship between TPR and FPR at various categorization levels. Reduce the classification threshold, and more items are classified as positive, increasing both true and false positives. Figure 4 shows the ROC curve for the decision tree, logistic regression, bagging classifier, and ANN. The more the area under the curve, the better the model. According to curves, it can be seen that the best performing models are logistic regression and bagging classifier. Although the ROC curve is not bad, compared to logistic regression and bagging classifier, it does not cut the best algorithms for this case due to the limitation of the dataset.

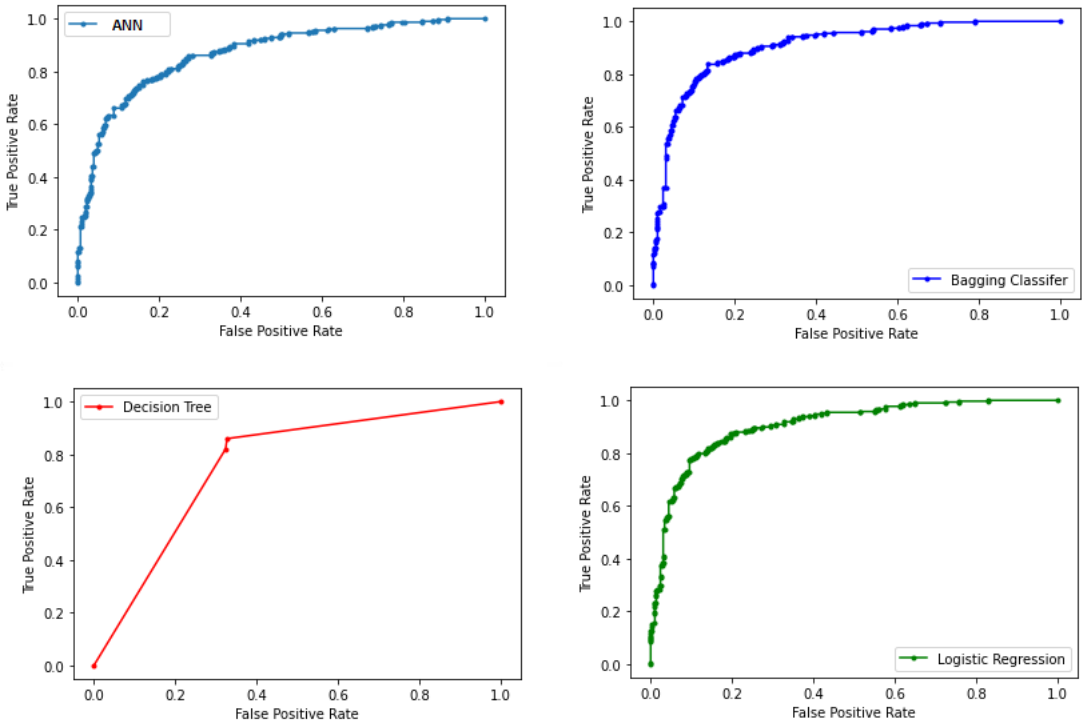

Figure 4: ROC Chart for All the Algorithms

7. Conclusion

To eradicate the problem of the use of abusive language on social media, machine learning is employed to detect abusive remarks in Urdu tweets. The dataset used in this study was obtained from the text processing laboratory at the center for computing research of Instituto Politécnico Nacional. The detection of abusive language in Urdu text is performed as a classification task. The dataset has only two labels which makes it a binary classification problem. In order to solve this problem, the algorithm chosen were logistic regression, decision tree bagging algorithm, and ANN, which could work well on binary classification. Accuracy, precision, recall, F1 score, and AUC ROC were used as evaluation metrics. The key parameter for deciding the best algorithm was the F1 score. Based on the F1 score, logistic regression, and bagging performed equally well. However, logistic regression was chosen as the best performing model with an 83% of F1 score. ANN did not perform well due to the limitation of the dataset. All these evaluations were made on 25% of the test split.

For future work, different embedding layers will be trained on Urdu data. Different pre-trained models will be tuned for this use case to improve accuracy. After acquiring a model with better results, a chrome extension can be created to detect abusive Urdu text on social media.

8. References

[1] H. Mubarak and K. Darwish, "Arabic offensive language classification on Twitter," in Proceedings of the International Conference on Social Informatics, pp. 269–276, National Research Council of Pisa, Pisa, Italy, May 2019.

[2] E. Abozinadah, Detecting Abusive Arabic Language Twitter Accounts Using a Multidimensional Analysis Model, George Mason University, Fairfax, VA, USA, 2017.

[3] Nayel, Hamada A., and H. L. Shashirekha. "DEEP at HASOC2019: A Machine Learning Framework for Hate Speech and Offensive Language Detection." FIRE (Working Notes). 2019.

[4] K. Stapleton, "Swearing and perceptions of the speaker: a discursive approach," Journal of Pragmatics, vol. 170, pp. 381–395, 2020.

[5] . de Gibert, O., Perez, N., Garc´ıa-Pablos, A., Cuadros, M.: Hate Speech Dataset from a White Supremacy Forum. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). pp. 11–20. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/W18-5102

[6] Akhter, Muhammad Pervez, et al. "Exploring deep learning approaches for Urdu text classification in product manufacturing." Enterprise Information Systems (2020): 1-26.

[7] H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection", IEEE Access, vol. 6, pp. 13825-13835, 2018

[8] J. M. Schneider, R. Roller, P. Bourgonje, S. Hegele and G. Rehm, "Towards the automatic classification of offensive language and related phenomena in German tweets", Proc. 14th Conf. Natural Lang. Process. (Konvens), pp. 95, 2018.

[9] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surv. (CSUR), vol. 51, no. 4, pp. 1-30, 2018.

[10] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," PLOS One, vol. 14, no. 8, p. e0221152, 2019.

[11] M. Tomaiuolo, G. Lombardo, M. Mordonini, S. Cagnoni, and A. Poggi, "A survey on troll detection," Future Internet, vol. 12, no. 2, p. 31, 2020.

[12] T. Jiang, J. P. Li, A. U. Haq, A. Saboor, and A. Ali, "A Novel Stacking Approach for Accurate Detection of Fake News," IEEE Access, vol. 9, pp. 22626-22639, 2021.

[13] A. Kumar, V. Singh, T. Ali, S. Pal, and J. Singh, "Empirical evaluation of shallow and deep classifiers for rumor detection," in Proc. ICACM 2019, in Advances in Computing and Intelligent Systems, in Algorithms for Intelligent Systems, 2020, pp. 239-252.

[14] S. R. Sahoo and B. Gupta, "Real-time detection of fake account in twitter using machine-learning approach," in Proc. CICT 2019, in Advances in computational intelligence and communication technology, in Advances in Intelligent Systems and Computing, vol. 1086, 2021, pp. 149-159.

[15] E. W. Pamungkas, V. Basile, and V. Patti, "Misogyny detection in twitter: a multilingual and cross-domain study," Inf. Process. Manag., vol. 57, no. 6, p. 102360, 2020.

[16] Shah, Kanish, et al. "A comparative analysis of logistic regression, random forest and KNN models for the text classification." Augmented Human Research 5.1 (2020): 1-16.

[17] Wang, Peng, et al. "Classification of proactive personality: Text mining based on weibo text and short-answer questions text." IEEE Access 8 (2020): 97370-97382.

[18] Chen, Caixia, Liwei Geng, and Sheng Zhou. "Design and implementation of bank CRM system based on decision tree algorithm." Neural Computing and Applications 33.14 (2021): 8237-8247.

[19] El-Mahelawi, Jamal Khamis, et al. "Tumor Classification Using Artificial Neural Networks." International Journal of Academic Engineering Research (IJAER) 4.11 (2020).

[20] Amjad, Maaz, Alisa Zhila, Oxana Vitman, Sabur Butt, Hamza Imam Amjad, Grigori Sidorov, Alexander Gelbukh. "Overview of the shared task on threatening and abusive detection in Urdu at fire 2021." In CEUR Workshop Proceedings. (2021).
[21] Amjad, Maaz, Alisa Zhila, Oxana Vitman, Sabur Butt, Hamza Imam Amjad, Grigori Sidorov, Alexander Gelbukh. "UrduThreat@ FIRE2021: Shared Track on abusive threat Identification in Urdu." In Forum for Information Retrieval Evaluation. (2021).
[22] Amjad, Maaz, Noman Ashraf, Alisa Zhila, Grigori Sidorov, Arkaitz Zubiaga, and Alexander Gelbukh. "Threatening Language Detecting and Threatening Target Identification in Urdu Tweets." IEEE Access. (2021).
[23] Maaz Amjad, Noman Ashraf, Grigori Sidorov, Alisa Zhila, Liliana Chanona-Hernandez, Alexander Gelbukh. "Automatic Abusive Language Detection in Urdu Tweets." Acta Polytechnica Hungarica. (2021).