Automatic Abusive Language Detection in Urdu Tweets

Maaz Amjad[1], Noman Ashraf[1], Grigori Sidorov[1], Alisa Zhila[2], Liliana Chanona-Hernandez[3], Alexander Gelbukh[1] 1

[1]Instituto Politécnico Nacional, Centro de Investigación en Computación (CIC), Gustavo A. Madero, 07738 Mexico City, Mexico

[2]Independent Researcher, San Francisco, CA 94103, USA

[3]Instituto Politécnico Nacional, Escuela Superior de Ingeniería Mecánica y Eléctrica (ESIME), Gustavo A. Madero, 07340, Mexico City, Mexico

e-mails: maazamjad@phystech.edu, noman@nlp.cic.ipn.mx, Sidorov@cic.ipn.mx, alisa.zhila@ronininstitute.org, lchanonah2100tmp@alumnoguinda.mx, gelbukh@cic.ipn.mx

Abstract: Abusive language detection is an essential task in our modern times. Multiple studies have reported this task, in various languages, because it is essential to validate methods in many different languages. In this paper, we address the automatic detection of abusive language for tweets in the Urdu language. The study introduces the first dataset of tweets in the Urdu language, annotated for offensive expressions and evaluates it by comparing several machine learning methods. The Twitter dataset contains 3,500 tweets, all manually annotated by human experts. This research uses three text representation techniques: two count-based feature vectors and the pre-trained fastText word embeddings. The count-based features contain the character and word n-gram, while the pre-trained fastText model comprises word embeddings extracted from the Urdu tweets dataset. Moreover, this study uses four non-neural network models (SVM, LR, RF, AdaBoost) and two neural networks (CNN, LSTM). The study finding reveals that SVM outperforms other classifiers and obtains the best results for any text representation. Character tri-grams perform well with SVM and get an 82.68% of F1 score. The best-performing words n-grams are unigrams with SVM, which obtain 81.85% F1 score. The fastText word embeddings-based representation yields insignificant results.

Keywords: Twitter corpus; Abusive language detection; Urdu language; Machine learning

[1] Corresponding author: Alexander Gelbukh

1 Introduction

Abusive language detection is an alluring concept. People use language to highlight, depict, elicit, instruct, and urge to inform the nuances of themselves and their worlds [1], some use it for a good cause, and some use it for spite. Impacts of abusive language are detrimental, ranging from short-term emotional reactions (outrage, dread, self-fault, etc.) to long-term mental health effects (low confidence, misery, etc.), causing psychological and medical problems (rest issues, migraine, dietary issues, etc.) [2] [3]. According to the Guardian report[2], abusive language can change human behavior. Although several prevention and intervention strategies were introduced, usage of abusive language on social media increased in recent years.

The task of abusive language detection is widely investigated in languages other than the English language [5-10]. Some studies discussed linguistic aspects and linguistic resources in different languages, such as Arabic [6], German [9], Japanese [10], Indonesian [7], Danish [8], and Portuguese [5]. Although automatic abusive language detection is still in its earliest stage, no study to date investigated abusive language detection with automatic manners on Twitter in Urdu, a local language of Pakistan, having over 230 million worldwide native speakers[3]. In addition, Urdu is viewed as one of the best ten most spoken languages on the planet. According to the point of view of NLP tools, inaccessibility, and the shortage of annotated data [4], Urdu is viewed as a low-resource language. Therefore, the study mainly focuses on counting features (N-gram) and word embeddings as feature vectors for abusive language identification tasks using Urdu tweets.

Abusive language detection is a challenging task. Recently social networks established themselves as the primary platforms for discussion, sharing ideas, and emotions. Being free and accessible, they lack language moderation. While most of the users stay cordial and polite, some occasionally express themselves in a manner that is obscene/profane and even might be rude or offensive to other users. The profane and objectionable content on social media might severely affect the addressee's emotional state and deteriorate life quality. Therefore, automatic abusive language identification is an invaluable measure. Among all, it can blow with a shovel the obscenities or indecent content for increased child protection.

Twitter is recognized as a social network where users can only post short text posts. To mitigate the use of abusive language on its platform, Twitter characterized harmful conduct as an endeavor to molest, threaten, or quietness of another person's voice[4]. Abuse is characterized by physical or psychological maltreatment. For example, I love you, but you are chubby. In this context, the word chubby might be perceived as diminishing one's appearance. Therefore, such reference to human appearance, personality or behavior might be considered a vile aspersion.

[2] https://www.theguardian.com/education/2011/oct/03/researchdemonstrates-language-affects-behaviour

[3] https://www.statista.com/statistics/266808/the-most-spokenlanguages-worldwide/

[4] https://help.twitter.com/en/rules-and-policies/abusive-behavior

Till today, no work identified abusive language using Urdu tweets, and no related corpora were recently gathered to the best of our knowledge. Moreover, no relevant Twitter dataset containing Urdu tweets was recently compiled. Therefore, this study presents the first balanced dataset containing abusive and non-abusive tweets in Urdu to address the automatic detection of abusive tweets.

Furthermore, as Urdu stays a relatively low resource language, we explore how data-intensive approaches such as neural networks and embedding-based text representation perform compared to count-based features and linear classifiers.

This study makes three main contributions, discussed as follows:

- Presents the first dataset in the Urdu language, for the automatic detection of abusive language using Twitter postings in Urdu, manually labeled by experts using given guidelines. This study also clarifies the dataset collection and annotation process that addresses the task of automatic abusive language detection, in Urdu.
- Baseline results utilize five non-neural network models (RF, Ada-Boost, MLP, LR, SVM) and two neural network models (LSTM, 1D-CNN). Three text representations techniques are used: two count-based and the pre-trained fastText word embeddings to identify abusive postings on the Twitter dataset.
- Analyzes the performance of different machine learning and deep learning algorithms on the proposed dataset.

The remaining paper is divided into different sections: the recent studies on the identification of abusive language are highlighted in Section 2. Section 3 examines the guidelines used to create and annotate the dataset. The results obtained in the experimental setup using machine learning (ML) and deep learning (DL) algorithms are presented in Section 4. Section 5 analyzes the performance of various classifiers and explains the results in detail. Finally, in Section 6, this study provides conclusions to our work.

2 Literature Review

This section first discusses the definition of abusive language and subsequently sheds light on existing research in the automatic detection of abusive language.

2.1 Defining and Characterizing Abusive Language

Twitter is characterized as one of the top five social networks, where a substantial number of users experience unethical communication and bullying. Using this platform, users can access a large active Twitter community (more than 330 million) and write a tweet with a maximum of 280-characters[5]. Several recent studies [11-13] highlighted that abusive language, and bullying cases are often reported on Twitter that contain injurious consequences for active tweeter users. Thus, Twitter took some safety measures and described some policies to control the usage of abusive language on its platform. According to the new guidelines, messages from obscure clients who have no profile picture will be removed. If the tweet is detected containing abusive words, this will lead to removing the user account from Twitter[6]. Nonetheless, Twitter must take more robust steps, especially distinguishing harmful tweets in different dialects since individuals utilize different abusive words in other languages.

2.2 Available Methods for Abusive Language Detection

Twitter permits its users to create new profiles, follow existing profiles, send messages and tags (both private and public), uploading status, images and videos. Within minutes and seconds, a single tweet can target a massive number of audiences, primarily through commenting, liking, sharing, and re-tweeting mechanisms. Among all the social media platforms, ordinary people widely use Twitter, yet this social network also attracted politicians, government organizations, and a government media spokesperson to release government statements (i.e., policies). Eventually, it creates a space for ignorant users to spread humiliating, hostile, and infancy comments with high velocity to a broader range of people.

Online platforms started introducing new policies to counterfeiting this issue because young people were a target group for bullying victimization. For example, Instagram started to fight against bullies by introducing shadow banning online abusers (i.e., limiting the user (bully) who used abusive language from publishing new posts or commenting on others' posts). Instagram introduced this system to mitigate cyberbullying events[7]. Likewise, another platform called Ask.fm [7] (an online website that permits its users to ask each other questions without disclosing identity) also introduced new policies to avoid discrepancies between users and other threats (life threats).

Numerous studies discussed abusive language detection and proposed various methods ranging from traditional machine learning to neural network-based models. Two features, such as character-level and word-level representations, were used to detect abusive language [14, 15, 19]. Furthermore, another study [18] highlighted that feature engineering, e.g., n-grams and POS tags, are extremely fruitful in machine learning methods. Other studies [14-18] used various traditional machine learning models, such as support vector machines, random forest, decision tree,

[5] https://www.statista.com/ statistics/282087/number- ofmonthly- active- twitter- users

[6] https://social.techcrunch.com/2017/02/16/twitter-starts-puttingabusers-in-time-out

[7] https://qz.com/1661410/instagram-wants-to-fight-bullies-byshadowbanning-them-and-telling-them-they-are-bullies/

logistic regression, and deep learning models like BERT [32] and Roberta [32], to identify hate speech. Further, several neural network architectures [8, 15, 17, 20] were used to identify the abusive language in Twitter posts. Recent studies [20] concluded deep learning models, such as convolutional neural networks (CNN) along with recurrent neural networks (RNN) [8] outperformed traditional ML classifiers, such as Logistic Regression [8] [19] and SVM [14-17]. Obscenity and offensive language detection tasks focused on languages, such as English [5] [8] [14-19] [20-24], Indonesian [7], Arabic [6], Portuguese [5], German [9], Japanese [10] and Danish [8]. Table 1 summarizes the recent works that examined abusive language detection in different languages.

3 Abusive Tweets Dataset in Urdu

This section explains the steps followed for dataset creation and data annotation. The dataset creation is divided into two stages (i) dataset crawling and (ii) dataset annotation.

3.1 Data Crawling

This study used Twitter API[8] to extract the tweets in the Urdu language using abusive keywords. To crawl tweets from Twitter, some keywords are used that contained only either a word or at least two abusive words. A dictionary containing abusive words and phrases in Urdu was manually created based on the most frequent words used on different social media platforms. The complete list of the keywords used to crawl the abusive tweets can be accessed[9].

This study collected the dataset for 20 months, starting from 01 January 2018 to 30th August 2019. This time interval was chosen primarily due to the General Elections in Pakistan held in July 2018. Typically, during or near election season, supporters of different political parties express their emotions and show antagonistic behavior to each other.

According to a recent report, although abusive language and threats to anyone are not confined to politics[10], some people use abusive and threatening language as a potent weapon for a political campaign.

Similarly, some people use social networks to use vulgar language to support a specific political party. For example, the current prime minister of Pakistan claimed that the daughter of the Ex-prime minister compelled her supporters to abuse him

[8] https://developer.twitter.com/en/docs/tweets/search/apireference/get-search-tweets

[9] https://github.com/MaazAmjad/Abusive_dataset.git

[10] https://www.cbc.ca/news/politics/violence-vandalismcampaign-rise-1.6177269

in public[11].

Table 1
Overview of the recent studies to identify the abusive language in different languages

Comparison of the state-of-the-art in abusive language detection				
Language	Platform	Feature extraction method	Classifier	Reference
English	Twitter	Char n-gram (1-4)	LR, Graph Convolutional Network	[20]
English	NewsGroup	Complement NB	Multinomial Decision Table NB (DTNB), Updateable NB	[22]
English	Twitter	BoW, Char n-grams	SVM, LR, CNN	[15]
English	YouTube	BoW, Word n-grams (2,3,5), Lexical Syntactic Feature	SVM, NB	[16]
English	Twitter	Word unigram	SVM, CNN, BiLSTM	[17]
English	Twitter	Word n-grams (1-8)	SVM	[18]
English	Twitter	Latent Dirichlet Allocation (LDA)	LR	[21]
English	User-generated online comments	Char and Word n-grams	NB, SVM	[14]
English	Twitter	BoW, char n-gram (3-8), word n-grams (1-3)	CNN, RNN, RF, NB, SVM, Gradient Boosted Trees, LR,	[19]
English	Twitter	BoW, word n-grams, hate or non-hate words list	SVM (linear, polynomial, radial)	[23]
English	Twitter, Articles	Abusive and non-abusive word list	Unsupervised learning	[24]
English, Portuguese	Twitter, Blogs	hateword2vec, hate-doc2vec, unigram	NB, SVM	[5]
Arabic	YouTube	Word n-grams	SVM	[6]

[11] https://tribune.com.pk/story/1300546/maryam-nawaz-forcingpml-n-leaders-abuse-public-imran

Indonesian	Twitter	cha and Word n-grams	SVM, NB, RF	[7]
Danish, English	Twitter, Facebook, Reddit	BoW, cha n-grams	LR, BiLSTM	[8]
German	Twitter	Twitter and Wikipedia embedding	CNN	[9]
Japanese	Blogs	Word n-grams (1-5)	SVM	[10]

Such events can induce a wave of anger in supporters of one political party towards other political parties. Thus, this increases the chances for malevolent tweet writing and social violence. Therefore, this period was chosen to extract maximum abusive tweets in the Urdu language.

In the crawling process, 55600 tweets in the Urdu language were retrieved that contained the seed words. The seed words are referred to as the words that were used to crawl the tweets. Although Urdu belongs to the Indo-Aryan language group, and some people believe that Urdu is a camp language, the report[12] contracted that Urdu is a camp language. Nonetheless, Urdu has roots[13] in the Arabic, Persian, and Turkish languages. Therefore, we removed all the crawled tweets that were written in other languages, such as Arabic, Persian, and Turkish. Thus, 47,700 tweets were obtained after the flirtation process, which were sent for the annotation process. In addition, instructions with a task definition and examples were provided to the annotators, particularly concerning the binary class annotation.

3.2 Dataset Normalization

In the dataset normalization process, all the non-Urdu tweets were deleted. Nonetheless, Urdu has roots[14] in the Arabic, Persian, and Turkish languages. Moreover, Urdu contains similar alphabets to these languages. Many tweets were crawled in these languages due to the same hashtags. Therefore, we removed all the crawled tweets that were written in other languages, such as Arabic, Persian, and Turkish. In addition, irrelevant tweet attributes like username, location, date and time, punctuation, uniform resource locator (URL), address, hashtag, emoticons (emojis), and the re-tweet symbol were also removed normalize the dataset and keep only the relevant information. Thus, 47700 tweets were obtained after the flirtation process, which were sent for the annotation process.

[12] https://www.dawn.com/news/681263/urdusorigin-its-not-acamp-language

[13] https://www.ucl.ac.uk/atlas/urdu/language.html

[14] https://www.ucl.ac.uk/atlas/urdu/language.html

3.3 Guidelines for Data Annotation

This study used paid crowdsourcing to label the dataset. This study did not use Amazon Mechanical Turk for crowdsourcing; instead, the Fiverr platform was used to hire annotators, and the annotation process was completed within two months. Moreover, a digital framework was introduced to alleviate human mistakes and accelerate the dataset annotation process. A strict criterion was constructed to recruit annotators:

(a) Indigenous to Pakistan

(b) Familiar with Twitter

(c) Urdu native

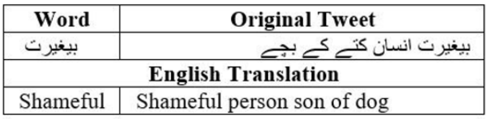
(d) Dissociate from any social, profit, political, non-political party or organization

(e) The annotator should fall within the age group of 20-35 years.

These points were significantly considered to minimize annotation prejudice, especially to annotate politics or election campaign tweets. Furthermore, 16 annotators were recruited for the dataset annotation, which contained 8 males and 8 females: 10 annotators belonged to the age group of 21-25 years, while 4 annotators were between 26-30 years age group, and 2 annotators belonged to 31-35 years. We also considered the educational background of the annotators so that the bias in the dataset could be minimized. The educational background of the annotators (last degree obtained) was as follows: 8 annotators held a bachelor's degree, 4 annotators had a master's degree, and 4 annotators earned a specialized journalism degree.

Forty-seven thousand seven hundred tweets were given to the annotators, and only 3500 tweets fulfilled the annotation process. Thus, the 3500 tweets were annotated as abusive tweets. Tables 1 and 2 show a sample tweet annotated as abusive and non-abusive, respectively.

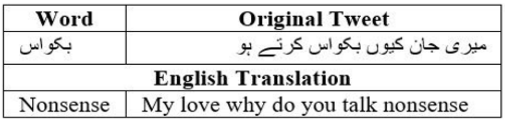
- **Abusive Tweet**: A Twitter post containing words to embarrass or humiliate other Twitter users.
- **Non-Abusive Tweet**: A Twitter post published for other objectives, such as mockery, joke, advertise, undermine, phishing, threatening, sarcasm, etc.



Word	Original Tweet
بیغیرت	بیغیرت انسان کتے کے بچے
English Translation	
Shameful	Shameful person son of dog

Table 1

Abusive tweet



Word	Original Tweet
بکواس	میری جان کیوں بکواس کرتے ہو
English Translation	
Nonsense	My love why do you talk nonsense

Table 2

Non-abusive tweet

3.4 Inter-Annotator Agreement

Calculating the agreement between dataset annotators is crucial for many reasons. First of all, this helps to annotate the dataset correctly. Secondly, bias in the dataset can be mitigated by using the inter-annotator agreement. Therefore, we used Cohan's Kappa Coefficient to quantify the reliability between annotators. As a result, a Kappa coefficient of 90% was accomplished after computing the inter-annotator agreement to collect the first abusive tweets dataset in the Urdu language.

3.5 Dataset Statistics

As a result of the annotation process, 3,500 tweets secured 100% annotator agreement for either abusive or non-abusive (and non-threatening) labels. This rigorous annotation procedure ensured the construction of a reliable dataset of 1,750 offensive tweets and 1,750 non-abusive tweets. Tables 2 show dataset statistics.

Table 2

Dataset statistics

Dataset	Words	Char	Avg Word	Total Tweets
Abusive	26,378	118,512	15	1,750
Non-Abusive	30,709	140,627	17	1,750
Totals	57,087	259,141	16	3,500

4 Experiment Settings

This section discusses the detailed experimental procedure to identify abusive tweets as a binary classification problem in which the task is to assign a label of whether a tweet is abusive or non-abusive. This study is based on neural networks (traditional machine learning), and non-neural networks (deep learning) approaches. Traditional machine learning algorithms, mainly supervised machine classifiers, were used. The machine learning classifiers used for automatic abusive language detection: Logistic Regression (LR), Multilayer Perceptron (MLP), Random Forest (FR), Support Vector Machine (SVM), and Adaboost classifier.

A python-based library, known as Scikit-Learn[15] library, was used to develop machine learning algorithms.

Two deep learning models are implemented with the Keras[16] library: 1-Dimensional Convolutional Neural Network (1D-CNN) and Long Short-Term Memory (LSTM). Eventually, for the training and the evaluation of the classifiers, 10-fold cross-validation was used. Figure 3 illustrates the overall methodology.

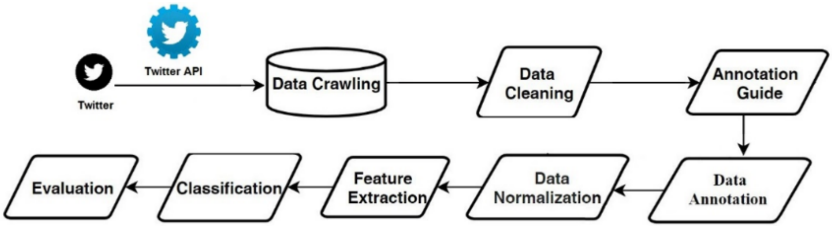


Figure 3
Methodology framework

4.1 Pre-Processing

The dataset was pre-processed to use for the experimental setup. First of all, all the tweets were converted into words (Tokens) using the white space character. Moreover, Western Arabic numerals were used to convert the numerals that followed the Eastern Arabic-Indic numeral system to normalize the entire data. Furthermore, stop words, white space tokens (blanks), punctuation, and bullets were also discarded to clean the dataset. Finally, we removed invalid utf-8 characters in the dataset and used standard utf-8 codification.

4.2 Features Extraction

This study used different text features to investigate the effect of the different types of text representation, namely, count-based (word n-grams and char n-grams) vs. embedding-based features on automatic abusive detection tasks. We considered character n-grams, word n-grams, and their combinations. We generated n-grams up to 6-grams because the previous study [4] reported that higher n-grams show insignificant results. To convert words into numeric features, a TF-IDF weighting scheme was used. We used the maximum number of features when the n-gram space dimension was higher.

For the pre-trained embedding features, we used fastText [26]. It is a neural network that is based on a word2vec algorithm. The word2vec model considers sub-words rather than dealing with entire words. In the training phase of word embeddings, if a word is not present in the dataset, its embeddings can be created by splitting the word into character n-grams.

[15] https://scikit-learn.org/stable/ [16]https://keras.io/about

4.3 Machine Learning Algorithms

For experiments, this research used five machine learning classifiers with all feature types: Multilayer Perceptron (MLP), AdaBoost, Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM). We used the default parameters of all machine learning classifiers.

4.3.1 Logistic Regression

Logistic regression (LR) is a linear model that assumes a linear relationship between input and output. It is based on the sigmoid function that measures a categorical dependent variable (abusive or non-abusive). Moreover, different studies [15, 20, 21] reported that LR provides better results on binary classification problems, particularly automatically detecting abusive language [8, 19].

4.3.2 Random Forest

A random forest (RF) is used for classification and regression tasks based on ensemble learning techniques. It uses bagging and feature randomness on different samples to construct multiple decision trees. It combines these decision trees to create a forest of trees where prediction by the majority is helpful to make an accurate prediction compared with any individual tree. The majority voting of multiple decision trees is used for the classification task, while an average of these decision trees is used in regression problems [19]. This algorithm uses dataset features to build individual decision trees and address variance and over-fitting challenges [13].

Moreover, these features are randomly selected to construct multiple decision trees. Recent studies used a random forest (RF) to classify abusive language [7].

4.3.3 Support Vector Machine

Support Vector Machine (SVM) [27] creates a line or a hyperplane (decision boundary) to separate the data into classes. It is a predictive analysis data classification algorithm used for linear, nonlinear classification, and regression tasks. Moreover, the kernel trick transforms data and finds an optimal boundary (clear margin of separation) between the possible outputs based on data transformation. SVM provides better results in high-dimensional spaces. In other words, SVM effectively performs when the dimensions are higher than the dataset instances [7]. Furthermore, different studies [5, 18, 19] reported that the SVM algorithm outperformed other classifiers in automatically identifying abusive language [10].

4.3.4 Ada-Boost

The boosting algorithm [28] is a supervised machine learning algorithm that combines different algorithms and re-assigns the weights to the input data. For the Ada-Boost algorithm, misclassified instances are crucial because the task is to make a robust classifier by combining different algorithms to make accurate predictions. Moreover, the adaptive boosting algorithm provides higher weights only to relevant features. One of the limitations of using the AdaBoost algorithm is over-fitting. This is due to the noise present in the dataset. In other words, if features are not relevant (noisy dataset), Adaboot will not make accurate predictions. Different studies [29] revealed that using the Adaptive Boosting algorithm can effectively detect abusive language compared to other machine learning algorithms.

4.3.5 Multilayer Perceptron

A multilayer perceptron (MLP) [30] is a feed-forward neural network that generates outputs from a set of inputs. To train the MLP model, a technique called back-propagation is used that assigns weights to the neurons present in the neural network. Furthermore, this neural network consists of three layers (i) an input layer, (ii) a hidden layer, and (iii) an output layer, which is fully connected. The dataset samples are given as inputs in the input layer. Then, the dot product is used between the input samples and the weights to input the hidden layer. The output is given as input to the activation function so that the final output of the hidden layer is obtained. In the last stage, the dot product of the output of the activation function and the weights are measured, which are fed to the final layer to predict the final output. A recent study also reported that MLP showed good performance in classifying abusive language [31].

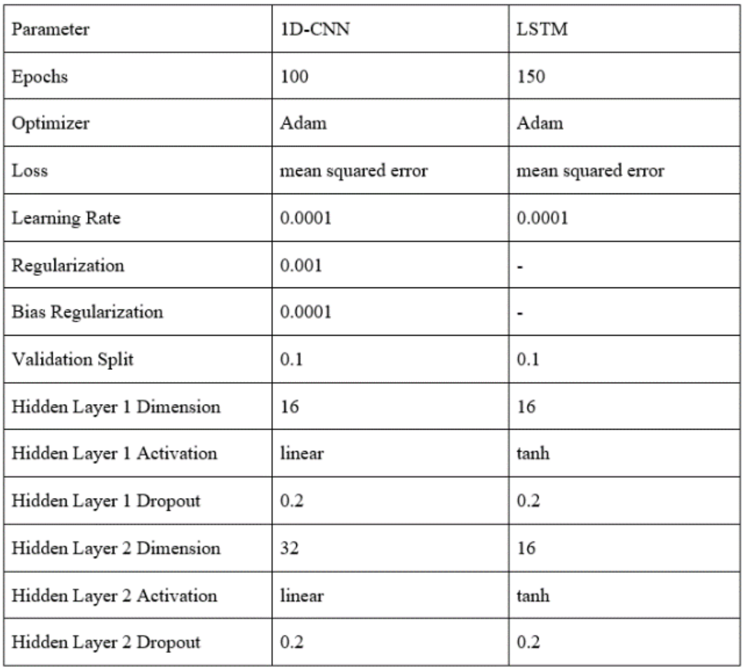
4.4 Deep Learning Classifiers

CNN and RNN are used to investigate abusive language detection tasks in Urdu. Figure 4 shows the information of deep learning parameters: all layers, parameters, and their values used for the experiments.

4.4.1 Convolutional Neural Network

A convolutional neural network (CNN) is a deep learning neural network used to solve various classification problems. This neural network typically comprises several layers where every hidden layer in the neural network contains neurons and biases. In addition to this, the dot product of the samples and the weights in the input layer is given to the activation function, which is present in the second layer. The activation function in individual neurons measures the dot product of the input samples and their weights and then adds a bias to the weighted sum. Like Multilayer Perceptron, Convolutional Neural Networks also use back-propagation to build a neural network. It is essential to mention that back-propagation reduces the error by re-assigning different values to the weights in each layer, starting from the final layer to the first layer of the neural network [17]. Moreover, this neural network is also effective in memory consumption along with dimensionality reduction.

Initially, CNN was used for image processing tasks (2D data) and video processing tasks (3D matrix). However, recent studies [9, 15, 19] also used CNN for text classification tasks using text-based features (1D matrix) [17]. CNN is extremely efficient in extracting relevant and distinctive features and making accurate predictions. Furthermore, unlike feed-forward networks, Convolutional Neural Network is computationally efficient. Figure 5 shows the architecture of 1D-CNN that was used for automatic abusive language detection.



Parameter	1D-CNN	LSTM
Epochs	100	150
Optimizer	Adam	Adam
Loss	mean squared error	mean squared error
Learning Rate	0.0001	0.0001
Regularization	0.001	-
Bias Regularization	0.0001	-
Validation Split	0.1	0.1
Hidden Layer 1 Dimension	16	16
Hidden Layer 1 Activation	linear	tanh
Hidden Layer 1 Dropout	0.2	0.2
Hidden Layer 2 Dimension	32	16
Hidden Layer 2 Activation	linear	tanh
Hidden Layer 2 Dropout	0.2	0.2

Figure 4

Deep learning parameters

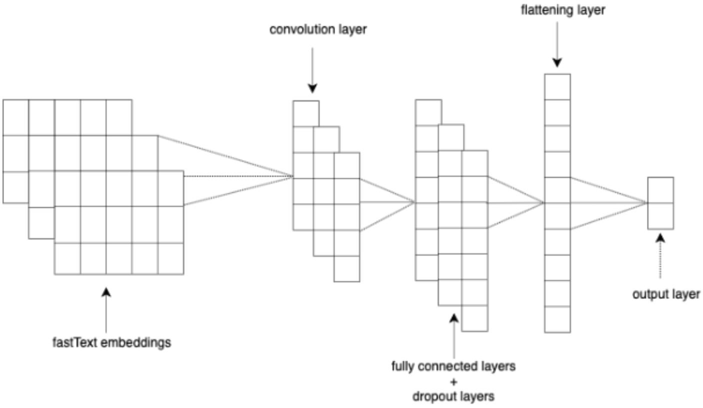


Figure 5

CNN model architecture

A pre-trained embeddings model, known as fastText embedding, is extracted from Urdu tweets to train the convolutional neural network. Subsequently, the 1D-CNN classifier receives these embeddings as input. The convolutional neural network contained two fully connected and a convolution layer. The filter size in the convolution layer was set to 8, and the window size of the kernel was fixed to 1. Moreover, this neural network is trained ten times using 100 epochs, and to avoid overfitting, a dropout is employed in all layers of the neural network. The mean accuracy of 10 iterations is used to acquire the CNN results in identifying abusive tweets.

4.4.2 Long Short-Term Memory Networks

Another deep learning algorithm, known as Long Short-Term Memory (LSTM) [8], was introduced to tackle the limitation of order dependence in sequence prediction projects, like speech recognition and machine translation [17, 19].

The LSTM model was also trained on fastText embedding extracted from Urdu tweets like the convolutional neural network. The LSTM contained two fully dense layers, and for training, each iteration had 150 epochs, and 10-fold cross-validation was employed. Initially, all the word vectors are normalized after each update, and a dropout layer between the hidden and output layer is used. The architecture of our proposed LSTM model is shown in Figure 6.

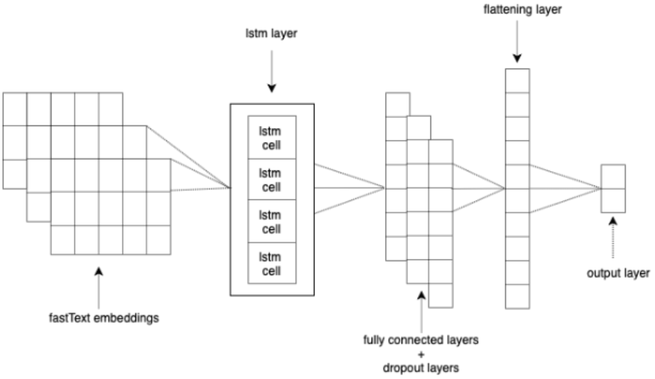


Figure 6
LSTM model architecture

4.5 Evaluation Metrics

This is a balanced dataset. For comparative analysis, two types of techniques were used in the study. For machine learning, five machine learning algorithms were used. Two neural networks were employed in the experimental setting for deep learning. Therefore, the algorithm’s selection is deemed to be appropriate for this task. All the proposed models are evaluated based on standard metrics, including F-measure, accuracy, and (ROC) curve.

5 Results and Analysis

We ran experiments using three text representations: two count-based features (word and char n-grams) and pre-trained fastText word embeddings. This study used fastText embeddings because this embedding represents each word as the sum of the n-gram vectors rather than learning vectors for each word. This embedding contains word vectors for 157 languages learned on Wikipedia and Crawl and addresses the issue of out of vocabulary (OOV). For example, boxer and boxing are employed in distinct contexts, and capturing the underlying commonality of both words is challenging. Therefore, this embedding addresses this problem by dividing the words into character n-grams. Furthermore, as compared to the BERT, fastText embeddings are exceptionally quick and can be trained on more than one billion words in less than ten minutes using a normal multicore CPU.

Moreover, we generated n-grams up to 6-grams for words characters and words. However, the results of word n-grams started to decrease after 4-grams. This is why the results of 5 and 6-word grams are not provided. The results of character n-grams, fastText, and word n-gram are shown in Tables 4, 5, and 6, respectively. The feature column in these tables represents the maximum number of features used to train the machine learning classifiers to distinguish abusive and non-abusive tweets. Moreover, character n-grams and word n-grams were extracted using the TF-IDF weighting scheme.

The results show that SVM outperformed other classifiers and achieved the highest accuracy of 82.37% and F_1 score of 82.68% with char tri-gram features. We only investigated the linear kernel and noticed that its performance was sufficiently high for the baseline experiments. Moreover, LR performed best on all char n-gram features. On the other hand, RF performed worst on the same features. Furthermore, SVM also achieved the best results using word unigram features, slightly less than the highest results. It had an F1 score of 81.85% and an accuracy of 81.27%. However, all the other machine learning models performed worst on bi-gram, tri-gram, and the combination of word n-gram features. Figure 7 illustrates the ROC curve, and Figure 8 represents the confusion matrix of SVM to differentiate between abusive and non-abusive tweets.

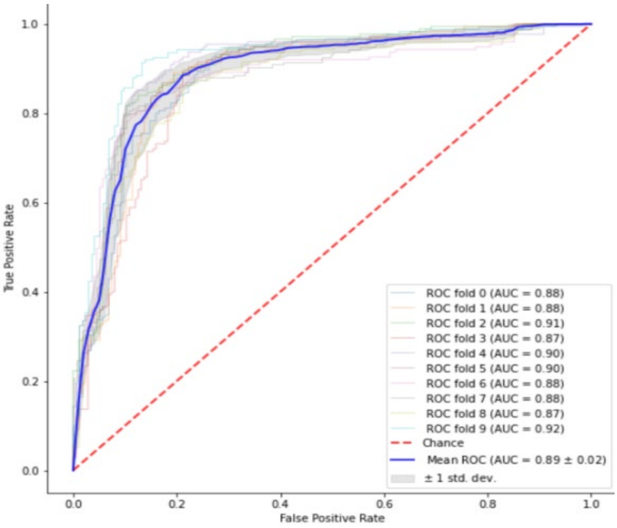


Figure 7
ROC curve for best performing model (SVM)

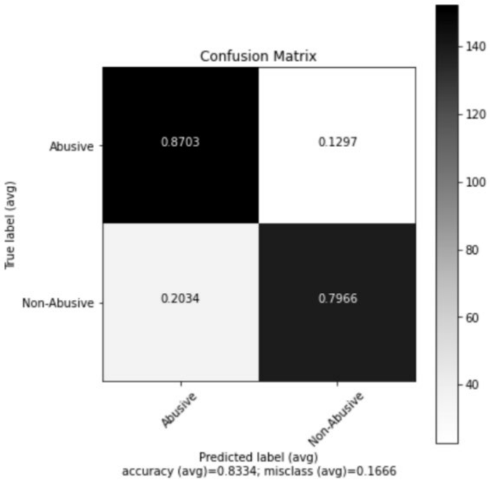


Figure 8
Confusion matrix for best performing model (SVM)

Deep learning algorithms like 1D-CNN and LSTM using fastText pre-trained word embeddings could not achieve the highest results for abusive language detection. Because of the limited training data, most of the words are not present in the fastText vocabulary. Another primary reason is that we used random vectors for out-of-vocabulary words. As a result, most of the vectors were diluted with bias. Moreover, it seems that the performance of deep learning classifiers improves as the dataset size increases. Overall, our results align with state-of-the-art efforts in abusive language detection and illustrate that there is still a great deal of room for growth.

Conclusions

Automatic threat language detection in English or other European languages is a challenging task that is widely examined. Nonetheless, as far as we can ascertain, there is no investigation into automatic abusive language detection in Urdu using Twitter postings. This paper contains a two-fold contribution. First, we collected and annotated the first corpus of tweets in the Urdu language for automatic abusive language detection. The corpus contained 3500 tweets that passed through pre-processing and rigorous manual annotation. Second, we compared the potential of various text representations for automated abusive language detection in Urdu tweets and ran a series of experiments with five different classification algorithms.

The experiment results demonstrate that SVM consistently obtained improved outcomes for both count-based feature types; the word unigrams and character tri-gram got better results than other n-gram features. Moreover, the fastText pre-trained word embeddings for Urdu obtained comparatively low results than the n-gram features. It might be because of the limited corpus size required to pre-train the embedding model, as well as a high number of out-of-vocabulary words that are likely to be present in abusive tweets. These baseline results will serve as a reference point for evaluating classification techniques developed by other researchers in the future. We aim to increase the dataset size and use transformers-based techniques to address abusive language detection in Urdu using Twitter postings for future research.

Table 4
Identification of abusive tweets using char-level features (TFIDF-based)

Feature set	# of Features	–	Classifiers				
			LR	MLP	AdaBoost	RF	SVM
3-gram	9491	P	86.51	79.00	83.31	83.39	86.09
		R	78.68	77.48	74.85	81.14	79.65
		Acc	83.17	78.40	79.85	82.45	83.34
		F_1	82.37	78.14	78.73	82.18	82.68
4-gram	29,138	P	87.03	80.44	85.67	83.17	86.16
		R	78.45	77.82	76.80	81.25	77.88
		Acc	83.37	79.40	81.94	82.37	82.65
		F_1	82.47	79.04	80.94	82.14	81.75
5-gram	59,460	P	86.66	80.26	86.30	81.25	86.23
		R	77.42	79.65	75.77	82.28	75.60
		Acc	82.74	80.00	81.85	81.62	81.74
		F_1	81.75	79.91	80.67	81.73	80.51
6-gram	90,573	P	86.62	76.67	86.06	77.11	86.45

		R	74.00	81.31	70.17	81.42	71.02
		Acc	81.25	78.22	79.37	78.57	79.94
		F_1	79.77	78.87	77.24	79.16	77.95
combination (3-6)-gram	188,662	P	86.82	81.88	85.23	84.17	86.17
		R	78.00	78.22	75.82	80.80	76.97
		Acc	83.05	80.42	81.28	82.77	82.28
		F_1	82.14	79.95	80.14	82.41	81.27

Table 5
Identification of abusive tweets using word-level features (TFIDF-based)

Feature set	# of Features	–	Classifiers	
			1D-CNN	LSTM
fastText	300	P	79.47	79.45
		R	79.37	77.42
		Acc	79.42	78.68
		F1	79.39	78.39

Table 6
Identification of abusive tweets using word-level features (TFIDF-based)

Feature set	# of Features	–			Classifiers		
			LR	MLP	AdaBoost	RF	SVM
unigram	6,671	Acc	82.31	77.40	81.25	81.25	82.82
		P	86.30	77.60	87.49	84.13	86.65
		R	76.85	77.08	72.97	77.08	77.65
		F_1	81.27	77.26	79.54	80.40	81.85
bigram	28,929	Acc	76.42	74.71	68.02	70.42	74.17
		P	83.65	72.23	89.23	66.32	85.19
		R	65.77	80.51	41.08	83.20	58.51
		F_1	73.60	76.09	56.14	73.76	69.32
trigram	38,006	Acc	69.11	58.05	54.57	56.40	65.45
		P	81.60	55.03	83.68	53.82	81.63
		R	49.37	88.40	11.31	90.11	39.94
		F_1	61.43	67.82	19.79	67.39	53.54
4-gram	37,577	Acc	64.54	51.42	52.94	51.48	64.57
		P	80.60	50.78	51.52	50.81	79.71
		R	38.34	90.39	99.42	92.57	39.20
		F_1	51.92	65.03	67.87	65.60	52.46
combination (1-4)-gram	111,183	Acc	81.25	80.05	81.34	79.40	78.02
		P	86.55	82.16	87.33	85.20	86.77
		R	74.05	76.91	73.31	71.25	66.17
		F_1	79.77	79.38	79.68	77.55	75.03

Acknowledgments

The work was supported by the Mexican Government through the grant A1-S-47854 of the CONACYT, Mexico and grants 20211784, 20211884, and 20211178 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors utilize the computing resources provided by the CONACYT through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico.

References

[1] W. Schwartz: Descriptive Psychology, and the Person Concept*:* Essential Attributes of Persons and Behavior, Academic Press, 2019

[2] Y. Urano, R. Takizawa, M. Ohka, H. Yamasaki, H. Shimoyama: Cyberbullying victimization and adolescent mental health: The differential moderating effects of intrapersonal and interpersonal emotional competence, Journal of Adolescence, Vol. 80, 2020, pp. 182-191

[3] Y. Zhu, W. Li, J. E. O'Brien, T. Liu: Parent-child attachment moderates the associations between cyberbullying victimization and adolescents' health/mental health problems: An exploration of cyberbullying victimization among Chinese adolescents, Journal of Interpersonal Violence, 2019

[4] M. Amjad, G. Sidorov, A. Zhila, H. Gómez-Adorno, I. Voronkov, A. Gelbukh: Bend the truth: Benchmark dataset for fake news detection in Urdu language and its evaluation, Journal of Intelligent & Fuzzy Systems, Vol. 39, No. 2, 2020, pp. 2457-2469

[5] R. Pelle, C. Alcântara, V. P. Moreira: A classifier ensemble for offensive text detection, Proceedings of the 24th Brazilian Symposium on Multimedia and the Web, 2018, pp. 237-243

[6] A. Alakrot, L. Murray, N. S. Nikolov: Towards accurate detection of offensive language in online communication in Arabic, Procedia Computer Science, Vol. 142, 2018, pp. 315-320

[7] M. O. Ibrohim, I. Budi: A dataset and preliminaries study for abusive language detection in Indonesian social media, Procedia Computer Science, Vol. 135, 2018, pp. 222-229

[8] G. I. Sigurbergsson, L. Derczynski: Offensive language and hate speech detection for Danish, arXiv preprint arXiv:1908.04531, 2019

[9] J. M. Schneider, R. Roller, P. Bourgonje, S. Hegele, G. Rehm: Towards the automatic classification of offensive language and related phenomena in German tweets, 14th Conference on Natural Language Processing KONVENS, 2018, p. 95

[10] T. Ishisaka, K. Yamamoto: Detecting nasty comments from BBS posts, Proceedings of the 24th Pacific Asia Conference on Language, Information

and Computation, 2010, pp. 645-652

[11] G. Sterner, D. Felmlee: The social networks of cyberbullying on Twitter, International Journal of Technoethics (IJT), Vol. 8, No. 2, 2017, pp. 1-15

[12] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, A. Vakali: Mean birds: Detecting aggression and bullying on Twitter, Proceedings of the 2017 ACM on Web Science Conference, 2017, pp. 13-22

[13] V. Balakrishnan, S. Khan, T. Fernandez, H. R. Arabnia: Cyberbullying detection on Twitter using Big Five and Dark Triad features, Personality and Individual Differences, Vol. 141, 2019, pp. 252-257

[14] Y. Mehdad, J. Tetreault: Do characters abuse more than words?, Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2016, pp. 299-303

[15] J. H. Park, P. Fung: One-step and two-step classification for abusive language detection on Twitter, arXiv preprint arXiv:1706.01206*,* 2017

[16] Y. Chen, Y. Zhou, S. Zhu, H. Xu: Detecting offensive language in social media to protect adolescent online safety, 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, 2012, pp. 71-80

[17] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar: Predicting the type and target of offensive posts in social media, arXiv preprint arXiv:1902.09666, 2019

[18] P. Rani, A. K. Ojha: KMI-coling at SemEval-2019 task 6: exploring N-grams for offensive language detection, Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 668-671

[19] Y. Lee, S. Yoon, K. Jung: Comparative studies of detecting abusive language on Twitter, arXiv preprint arXiv:1808.10245, 2018

[20] P. Mishra, M. D. Tredici, H. Yannakoudakis, E. Shutova: Abusive language detection with graph convolutional networks, arXiv preprint arXiv:1904.04073, 2019

[21] G. Xiang, B. Fan, L. Wang, J. Hong, C. Rose: Detecting offensive tweets via topical feature discovery over a large-scale twitter corpus, Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, pp. 1980-1984

[22] A. H. Razavi, D. Inkpen, S. Uritsky, S. Matwin: Offensive language detection using multi-level classification, Canadian Conference on Artificial Intelligence, Springer, Berlin, Heidelberg, 2010, pp. 16-27

[23] P. Burnap, M. L. Williams: Us and them: identifying cyber hate on Twitter across multiple protected characteristics, EPJ Data Science, Vol. 5, 2016, pp. 1-15

[24] H. S. Lee, H. R. Lee, J. U. Park, Y. S. Han: An abusive text detection system based on enhanced abusive and non-abusive word lists, Decision Support Systems, Vol. 113, 2018, pp. 22-31

[25] J. Cohen: A coefficient of agreement for nominal scales, Educational and Psychological Measurement, Vol. 20, No. 1, 1960, pp. 37-46

[26] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov: Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics, Vol. 5, 2017, pp. 135-146

[27] N. Rusnachenko, N. Loukachevitch, E. Tutubalina: Distant supervision for sentiment attitude extraction, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), 2019, pp. 1022-1030

[28] C. Ying, M. Qi-Guang, L. Jia-Chen, G. Lin: Advance and prospects of adaboost algorithm. Acta Automatica Sinica, Vol. 39, No. 6, 2013, pp. 745-758

[29] R. E. Schapire: Explaining Adaboost, Empirical Inference, Springer, Berlin, Heidelberg, 2013, pp. 37-52

[30] D. W. Ruck, S. K. Rogers, M. Kabrisky: Feature selection using a multilayer perceptron, Journal of Neural Network Computing, Vol. 2, No. 2, 1990, pp. 40-48

[31] M. W. Gardner, S. R. Dorling: Artificial neural networks (the multilayer perceptron): A review of applications in the atmospheric sciences, Atmospheric Environment, Vol. 32, No. 14-15, 1998, pp. 2627-2636

[32] Amjad. M, Zhila. A, Sidorov. G, Labunets. A, Butt. S, Amjad. H. I, Vitman. O, Gelbukh. A: UrduThreat@ FIRE2021: Shared Track on Abusive Threat Identification in Urdu, In Forum for Information Retrieval Evaluation, 2021, pp. 9-11