Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

ICASET
2024

Research Article
Collection: Intelligent Computing of Applied Sciences and Emerging Trends

Efficient Intelligent System for Cyberbullying Detection in English and Roman Urdu Social Media Posts

Muhammad Talha Jahangir[1*], Muhammad Ahmad[2], and Hamna Rehman[2]

[1]Department of Computer Science, MNS University of Engineering and Technology, Multan, Pakistan.
[2]Institute of Computing, MNS University of Agriculture Multan, Pakistan.
*Corresponding Author: Muhammad Talha Jahangir. Email: mtalhajahangir@mnsuet.edu.pk

Academic Editor: Salman Qadri **Published:** April 01, 2024

Abstract: The internet has revolutionized communication, offering new platforms like social media, blogs, and comment sections for people to connect. However, these platforms have also seen an uptick in abusive language, hate speech, and cyberbullying. In the more recent work, training models to identify harmful remarks across several classes is explored using algorithms. A recent study examined the efficacy of naive Bayes, logistic regression, and support vector machine as three different approaches to using an online negative feedback engine. Toxic, offensive, disparaging, hateful, and healthy (non-toxic) comments are screened out of the process. With 97.5% accuracy in English analysis and 92.9% accuracy in Urdu analysis, Support Vector Machines (SVM) performed better than the other approaches, according to the data. SVM has shown a strong capacity to identify hate speech and insults. This research is significant because it will contribute to the development of technologies that online platforms can employ to identify and eliminate unwanted information, making the internet a safer and more secure place.

Keywords: Online Toxicity; Hate Speech; Abusive Language; Machine Learning Algorithms; English; Roman Urdu; Toxic Comment Classification.

1. Introduction

The development of the internet has allowed for a wide interchange of ideas and connected individuals globally. Unfortunately, this has also led to hate speech, cyberbullying, and online harassment in the comments area [1]. This toxic material may be detrimental to a person's mental health as well as the general wellbeing of the online community. To create a more secure and effective

It's imperative to determine how to respond to and disseminate these detrimental statements in the online world. Negative thoughts can affect one's own health as well as the health of an online community, according to research. This is consistent with research done in the United States [2]. This illustrates the problem's worldwide scope and the requirement for a fix. The online world is changing quickly, therefore it's critical to create an environment that supports partnerships, cooperation, and community success.

The proposed methodology moves forward with the separate language Roman Urdu and English commenting for the multi-classification of damaging comments. To effectively handle hazardous information in both languages, customized solutions are needed. The dataset under investigation also highlights the striking differences in potentially harmful content between Roman Urdu and English. In these linguistic circumstances, targeted efforts are required to counteract detrimental conduct. According to the dataset, 37,578 toxics' and 1, 45,414 non-toxic comments were detected in Roman Urdu, underscoring the importance of identifying and mitigating harmful language through culturally sensitive methods. Although the English dataset consisted of 16 225 toxic remarks, for English-speaking users, it is essential to create efficient detection systems to maintain a safe online space. The objective of this study is to enhance the classification method of toxic remarks using three machine learning algorithms: SVM, Log Reg, and NB. Training was done on the dataset, including five toxicity notes – toxic, severe toxic, obscene, insult, and identity hate – and “healthy”, the non-toxic category. The following algorithms were tested on this data. With a (97.53%) accuracy rate, SVM outperformed both NB and Log Reg in the majority of toxicity categories for English comments. Roman Urdu comments are analyzed using SVM, which achieves (92.86%) accuracy, surpassing both NB (92.25%) and Log Reg (92.69%). Carefully considered to detect and categorize insults, obscene, and identity-hate comments in the algorithm, which has proven successful in detecting insults and obscene language.

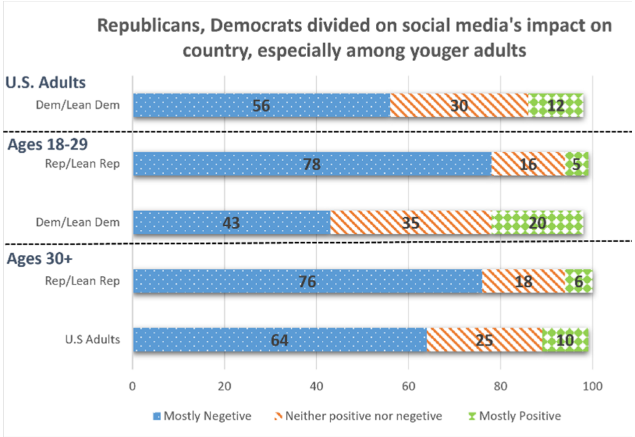

Figure 1. Pew Research Survey Report

The proposed methodology moves forward with the separate language Roman Urdu and English commenting for the multi-classification of damaging comments. To effectively handle hazardous information in both languages, customized solutions are needed. The dataset under investigation also highlights the striking differences in potentially harmful content between Roman Urdu and English. In these linguistic circumstances, targeted efforts are required to counteract detrimental conduct. According to the dataset, 37,578 toxics’ and 1, 45,414 non-toxic comments were detected in Roman Urdu, underscoring the importance of identifying and mitigating harmful language through culturally sensitive methods. Although the English dataset consisted of 16 225 toxic remarks, for English-speaking users, it is essential to create efficient detection systems to maintain a safe online space. The objective of this study is to enhance the classification method of toxic remarks using three machine learning algorithms: SVM, Log Reg, and NB. Training was done on the dataset, including five toxicity notes – toxic, severe toxic, obscene, insult, and identity hate – and “healthy”, the non-toxic category. The following algorithms were tested on this data. With a (97.53%) accuracy rate, SVM outperformed both NB and Log Reg in the majority of toxicity categories for English comments. Roman Urdu comments are analyzed using SVM, which achieves (92.86%) accuracy, surpassing both NB (92.25%) and Log Reg (92.69%). Carefully considered to detect and categorize insults, obscene, and identity-hate comments in the algorithm, which has proven successful in detecting insults and obscene language.

Presenting a comprehensive review of related works in Section II, describe the dataset preparation process in Section III, describe our methodology in Section IV, Section V delves into the details of the experimental configuration and outcomes. Conclusion, implications, and research prospects are discussed in Section VI.

2. Related Work

Prior Studies Addressing Toxic Comment Classification in Roman-Urdu and English. Saeed, H.H., et al. [3] classified toxic comments in Roman Urdu and English presents the results of a two-stage approach using hand-crafted features and machine learning. On Roman Urdu and English datasets, the algorithm achieved accuracy rates of 95.65% and 95.94%, respectively. Liu, et al. [4] offers an up-to-date perspective on toxic comment classification. Through the examination of machine learning methods and diverse datasets, an overview of the topic is provided in this study. In addition, the authors suggest future research directions, contributing to the advancement of this field.

Zhang, et al. [5] an NLP-based method is presented to categorize harmful remarks. A review of various machine learning algorithms for identifying toxic comments, an exploration of the use of linguistic features, and an evaluation of classification models using a variety of metrics are included in the paper.

Belal, et al. [6] using deep learning to sort Bengali toxic comments. In evaluation of a toxic comment, they use a binary classification model. The multi label classifier determines which type of toxicity has occurred in this case. They combine LSTM and CNN models for different stages, ensuring robust performance. As part of their approach to model prediction, the authors utilize the LIME framework. They achieved impressive accuracy rates of 89.42% for binary classification as well as 78.92% for multi-label classification after rigorous testing.

Usman, et al. [7] Approach Based on Word Embeddings and Transformer Models. A novel approach is described here for the classification of toxic Urdu comments. Word embedding is used to represent comments text. After that, the comments are classified as toxic or non-toxic using a transformer model. In a dataset of Urdu-language comments, the authors show that their method achieves 95.7% accuracy. Faisal Kamran, et al. [8] seeks to classify Roman Urdu toxic comments the authors of this paper discuss methods for identifying toxic comments and categorizing them. Their work offensive language is handled on online platforms, with potential applications in improving online discourse. 2.1. Strengths

The studies give a broad overview of the state-of-the-art in toxic comment classification, encompassing both Roman Urdu and English languages. A variety of methodological approaches, from hand-crafted features to machine learning and deep learning, are used in the studies. The results on a variety of benchmark datasets are quite impressive.

2.1. Potential Cons

Toxic comment classification models could be biased against comments that were seen as toxic at various times. Thus, the models may learn biases from the data they are trained on and, therefore, contain such biases. Bias in any form is unjust. Therefore, such biases should be of concern to researchers.

3. Dataset Preparation

Multi-classification to eliminate harmful remarks in Roman Urdu and English was performed by reclassifying data to ensure consistency and accuracy. The datasets were manually categorized into the following categories:

- Toxic: Comments likely to cause harm or offense to others.
- Severe Toxic: Comments extremely harmful or offensive.
- Obscene: Comments that are sexually explicit or vulgar.
- Insult: Comments intended to belittle or demean others.
- Identity Hate: Comments against a particular identity group.
- Healthy: Comments that are not toxic.

3.1. Data preparation resources included

3.1.1. Data Type

Repository consists of Roman Urdu comments that have been manually classified as toxic or non-toxic [9]. Comments collected under this license are included in this dataset. It was necessary to manually classify the comments because they were not explicitly labeled as toxic or non-toxic [10]. Comments dataset obtained from a separate drive. Manual categorization was necessary because there were no explicit labels indicating whether it was toxic or not [3].

3.1.2. Reclassification and Labeling of Data

To enhance data quality and consistency, texts from all dataset [11] were manually reclassified into the categories as shown in Figure 2 and Table I. A result of this better understanding of the data's toxicity and nature. The reclassification process had a significant impact on model performance. As a result of consistent and accurate labeling, the models were able to understand the complexities of toxic comments and make more accurate predictions. Classification of toxicity distribution statistically also resulted in improved toxic comment classification accuracy.

After that, the comments are classified as toxic or non-toxic using a transformer model. In a dataset of Urdu-language comments, the authors show that their method achieves 95.7% accuracy. Faisal Kamran [8] seeks to Kaggle competition provided the dataset for English toxic comment classification [12].

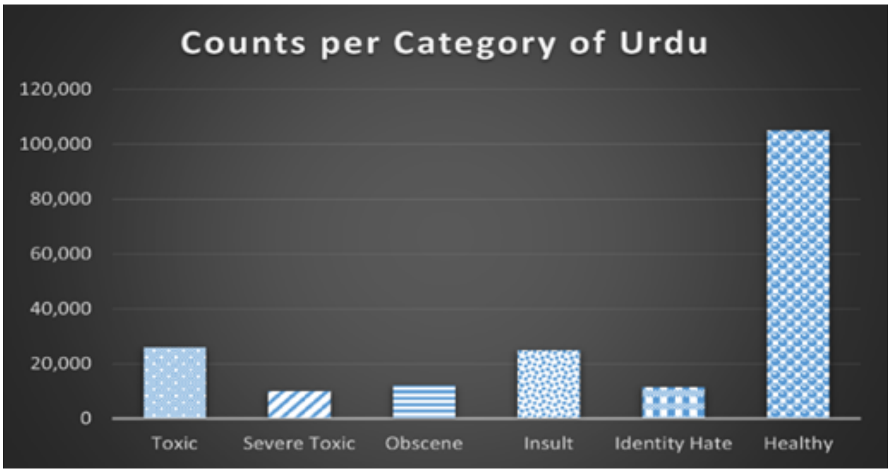

Figure 2. Multi-labeled toxic comment classification of Roman Urdu

Table 1. Multi-label toxic comments statistics Roman Urdu

Category Name	Counts per Category	Percentage of all Data
Toxic	26,106	18.23%
Severe Toxic	10,005	6.99%
Obscene	12,050	8.42%
Insult	24,856	17.36%
Identity Hate	11,588	8.09%
Healthy	105,000	73.33%

A "healthy" category was added for non-toxic comments, and the "threat" category was removed as shown in Figure 3 and Table II. The. Duplicate and irrelevant comments were removed. In the final dataset [13], 1.5 million comments were distributed evenly across categories. Machine -learning model trained on this dataset to detect toxic English comments with a higher degree of accuracy.

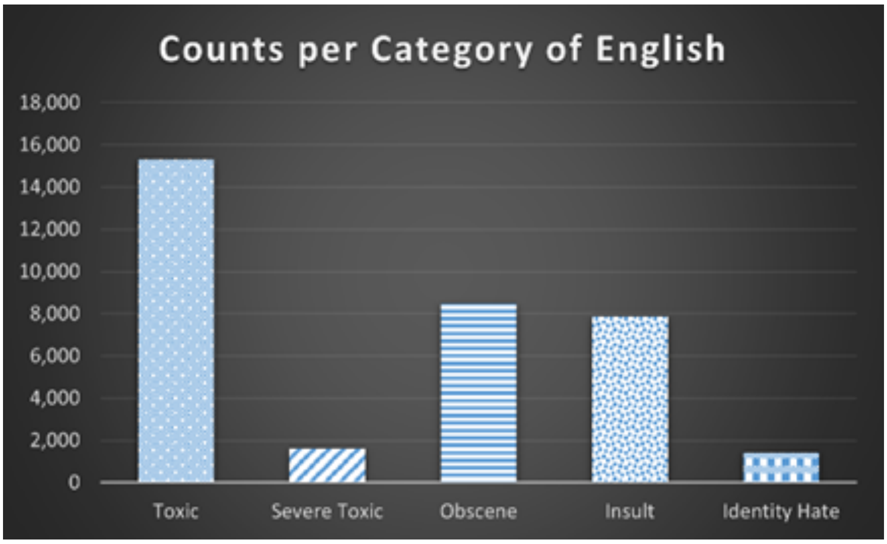

Figure 3. Multi-labeled toxic comment classification of English

Table 2. Multi-label toxic comments statistics English

Category Name	Counts per Category	Percentage of all Data
Toxic	15,294	9.58%
Severe Toxic	1,595	1.00%
Obscene	8,449	5.29%
Insult	7,877	4.94%
Identity Hate	1,405	0.88%
Healthy	143,346	89.83%

3.1.3. Dataset statistics

The dataset contains 37,578 toxics and 1,45,414 non-toxics comments in Roman Urdu, highlighting the toxic content as shown in Figure 2 and Table I. The English dataset, on the other hand, contains 16,225 toxic comments and 1,43,346 non-toxic comments as shown in Figure 3 and Table II, indicating a lower toxic prevalence. Roman Urdu requires culturally sensitive approaches, while English needs robust systems.

4. Proposed Methodology

This section outlines the essential architecture for detecting toxic comments in Roman Urdu, and English. In toxic comment classification, machine learning algorithms play a crucial role. The proposed methodology shows the effectiveness of three popular algorithms for text classification tasks: SVM, Log Reg, and NB.

4.1. Support Vector Machines (SVM)

SVMs are powerful supervised learning algorithms used for binary and multiclass classification. Based on a labeled dataset, SVMs find hyper planes that maximize margins between classes. Binary classification objective function formulated with SVM:

$$\frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n} max\,(0{,}1 - y_i(\omega.x_i - b))$$

Where, *w* is the weight vector, *b* is the bias term, *C* is the regularization parameter, x_i represents the feature vector of i^{th} sample, $y_i$is corresponding label (+1 or -1).

4.2. Logistic Regression (Log Reg)

A logistic regression classifier is used for binary classification. The sigmoid function assists model the classification probability of a sample. Logistic regression can be expressed as:

$$P(y_i{=}1\,|\,x_i) = \frac{1}{1 + e^{-(\omega.x_i+b)}}$$

Where *w* is the weight vector, *b* is the bias term, and $x_i$represents the feature vector of the i^{th} sample.

4.3. Naive Bayes (NB)

This algorithm is based on Bayes' theorem and is a probability-based algorithm. Feature-based theory assumes are conditionally independent given a class label. In text classification, NB determines the probability of a sample belonging to various classes based on the occurrences of individual words. Naïve Bayes can be expressed as follows:

$$P(y_i|\,x_i) = \frac{P(x_i|y_i)\,.\,P(y_i)}{P(x_i)}$$

The posterior probability of class y_i given feature vector $x_i$is calculated by multiplying the likelihood of features given class, $P\ (x_i|y_i)$, with the prior probability of class, $P\ (y_i)$, and then dividing by the evidence, $P\ (x_i)$.

4.4. Methodology Proposed Architecture

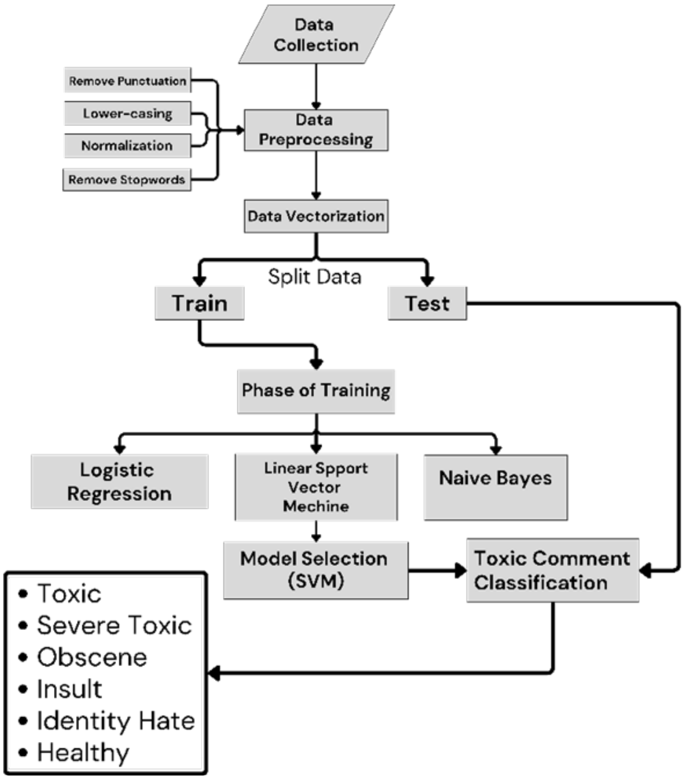

Figure 4. Multi-labeled Toxic Comment Classification Methodology

4.4.1. Data Collection

The first step involves collecting a comprehensive dataset of Roman Urdu and English comments from different online platforms. The dataset should contain toxic comments, including toxic, severe toxic, obscene, insult, and identity hate categories, along with a non-toxic category labeled healthy. The dataset can be collected manually or automatically through web scraping.

4.4.2. Data Preprocessing

After data collection, comments are processed to remove irrelevant information, such as special characters, symbols, and URLs. Cleaning ensures noise-free datasets are ready for analysis as shown in Figure 5 and Figure 6.

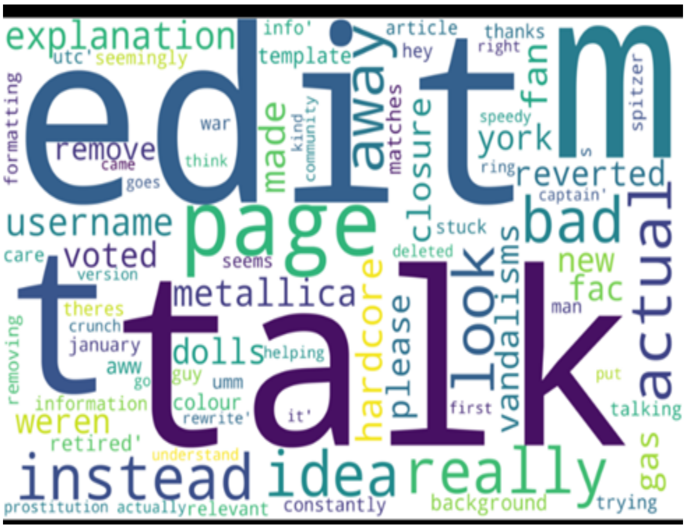

Figure 5. Word Cloud Analysis of Healthy

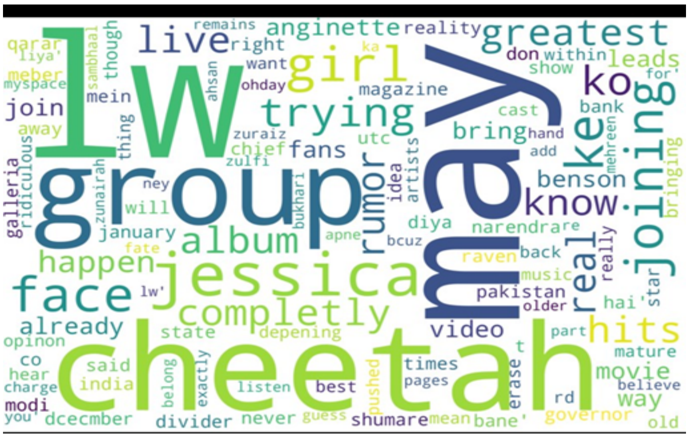

Figure 6. Word Cloud Analysis of Toxic Comment

4.4.3. Data Vectorization

Machine learning requires text data to be converted to numerical vectors. Some of the algorithms used to process information in text include Word embedding, (BoW), and Term Frequency-Inverse Document Frequency.

4.4.3.1. Classification Dimensionality of Toxic Comments

In order to represent word frequencies, the training dataset, comprising 146,393 samples, was vectorized into a 5,000-dimensional feature matrix. The testing dataset, which included 36,599 samples, was also subjected to this technique and dimensionality (5,000). As a result of the uniformity of the approach, the model was applied precisely. While high dimensions' capture complexities, they may introduce noise, whereas low dimensions may oversimplify. Feature count and model complexity are balanced in our approach to achieving optimal generalization. A toxic comment classification is highly dependent on feature dimensions.

Table 3. Feature Dimension of Roman Urdu Comment

Training Feature Shape	Test Feature Shape
(146393, 5000)	(36599, 5000)

Table 4. Feature Dimension of English Comment

Training Feature Shape	Test Feature Shape
(127656, 5000)	(31915, 5000)

4.4.4. Normalization

A normalization step is crucial when dealing with multilingual datasets like Roman Urdu and English. A standardization process accounts for the different styles and spellings of writing, as well as linguistic variations in text data. It helps to ensure consistency in the dataset, avoid biases, and ensure fair comparisons during classification.

4.4.5. Split Training and Testing Data

Machine learning divides the dataset into two parts: the training set is used to train the machine learning algorithm, while the test set is used to measure the performance of the algorithm. The training and testing data will be both representative and balanced to minimize the risk of over fitting or under fitting.

4.4.6. Phase of training

Using the training dataset, the selected machine learning algorithms (SVM, Log Reg, and NB) are trained. Both the Roman Urdu and English models learn from the data to recognize patterns and features associated with toxic comments.

4.4.7. Testing phase

The trained models are evaluated on two scenarios:

4.4.7.1. Same-Domain Testing

Models are tested in the same language domain (i.e., Roman Urdu or English) using the same testing dataset. As a result, they are evaluated on how accurately they can classify toxic comments in the same linguistic context.

4.4.7.2. Cross-Domain Testing

The models are also tested on the testing dataset of the language trained in Roman Urdu, tested in English. Through this cross-domain testing, assess how generalizable and robust the models are for different languages. Through the development and evaluation of methodologies, the study will ensure a safe digital environment for users by creating accurate toxic comment classifiers in Roman Urdu & English.

5. Experimental Setup and Results

Machine learning models were developed through the use of Python, a versatile programming language bolstered by diverse libraries designed for analyzing data and applying machine learning techniques. To explore more advanced techniques, experiments as shown in Table V were conducted on the Google Colaboratory Pro platform, which is a cloud-based environment with powerful GPUs, thereby accelerating the training process. During these trials the laptop has an 8th-generation Intel Core i5 processor, 8GB of RAM, and a 256GB SSD for experiment with modal. In our approach Utilized the following machine learning algorithms for our experimental process: SVM, Log Reg, and NB. Models were trained using a variety of epochs (ranging from 30 to 64) and batch sizes (ranging from 32 to 64) as shown in Table VI and Table VII.

5.1. Insights and Analysis

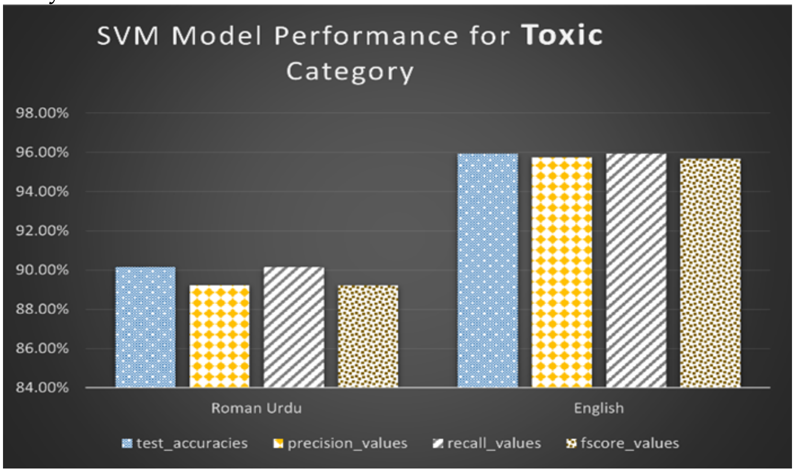

Figure 7. Comparing Toxic Label: Roman Urdu vs. English (Accuracy, Exactness, Specificity, and F1-score)

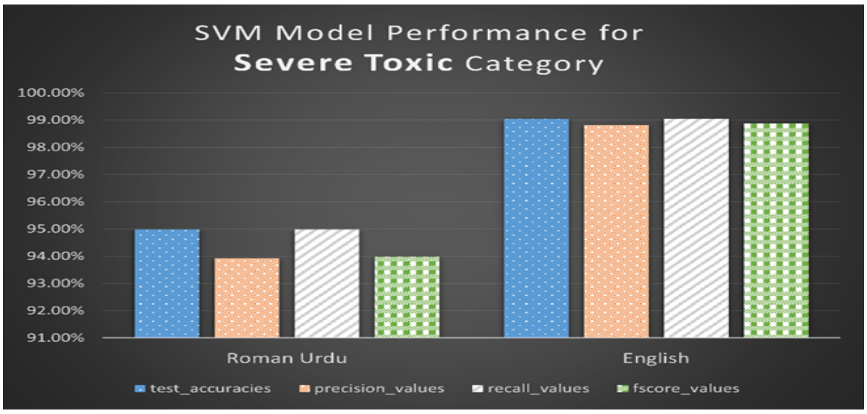

Figure 8. Comparing Severe Toxic Label: Roman Urdu vs. English (Accuracy, Exactness, Specificity, and F1-score)

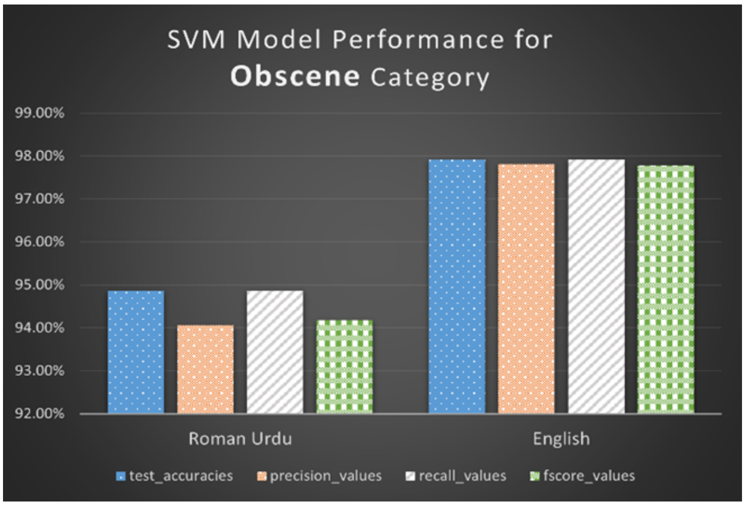

Figure 9. Comparing Obscene Label: Roman Urdu vs. English (Accuracy, Exactness, Specificity, and F1-score)

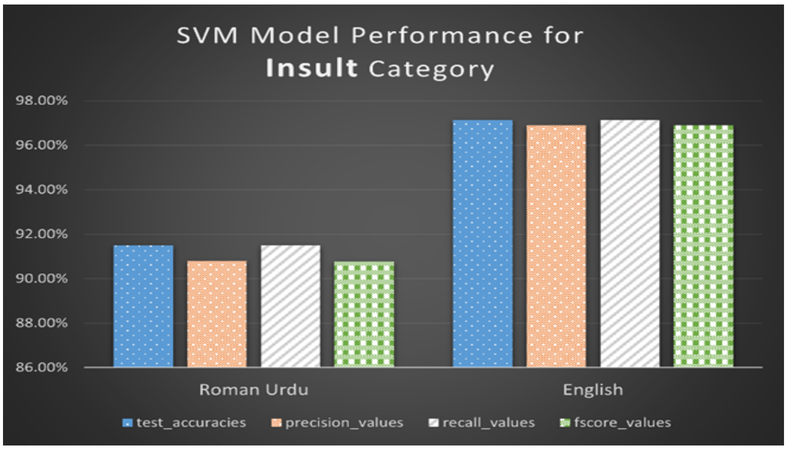

Figure 10. Comparing Insult Label: Roman Urdu vs. English (Accuracy, Exactness, Specificity, and F1-score)

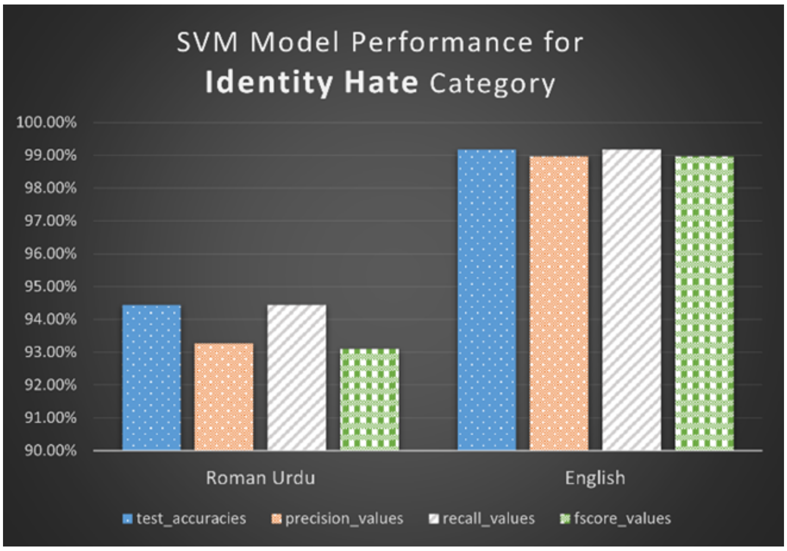

Figure 11. Comparing Identity Hate Label: Roman Urdu vs. English (Accuracy, Exactness, Specificity, and F1-score)

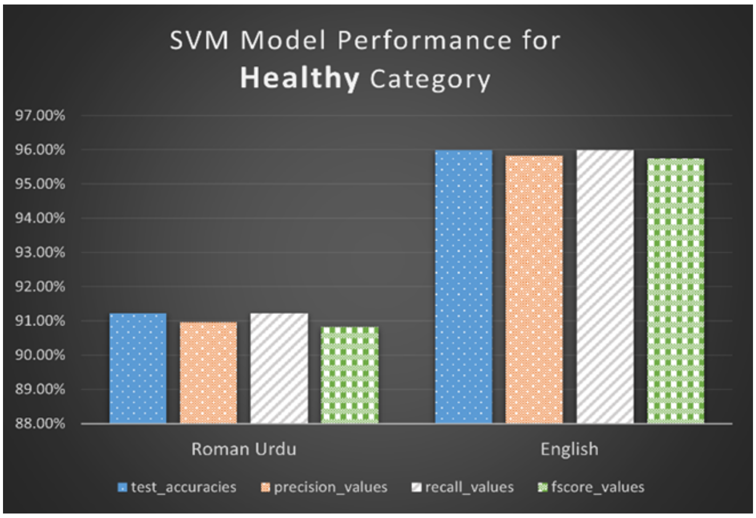

Figure 12. Comparing Healthy Label: Roman Urdu vs. English (Accuracy, Exactness, Specificity, and F1-score)

Table 4. Different Modal Comparison and Trained with Embedding

Model	Accuracy %	Exactness %	Specificity %	F1-score %
Log-Reg	2.22	8.37	94.78	15.37
Random Forest	78.50	48.11	15.19	23.09
Gradient Boosting	79.23	57.62	0.90	1.76
CNN	81.06	66.73	49.33	56.72
RNN	65.24	63.28	59.10	58.72
BERT Model	81.65	86.37	82.45	83.37
LSTM	82.74	85.37	82.45	60.72
LSTM-CNN	84.80	82.78	81.09	84.00

Table 5. Comparing Roman Urdu Model

Model	Accuracy	Exactness	Specificity	F1-score
SVM	0.928614	0.920414	0.928613714	0.920151
LR	0.926988	0.918231	0.926987987	0.91696
NB	0.922562	0.911593	0.922561636	0.910817

Table 6. Comparing English Model

Model	Accuracy	Exactness	Specificity	F1-score
SVM	0.971081	0.968547	0.971080622	0.96735
LR	0.970581	0.966129	0.970580539	0.965095
NB	0.959758	0.95102	0.959758477	0.953486

5.2. Training and Validation Accuracy

Indicators of a model's performance include training and validation accuracy as shown in Figure 13. Their purpose is to gauge how well the model generalizes to new, unseen data based on training data. These accuracy comparisons allow us to identify potential over fitting or under fitting issues and make informed decisions about model optimization.

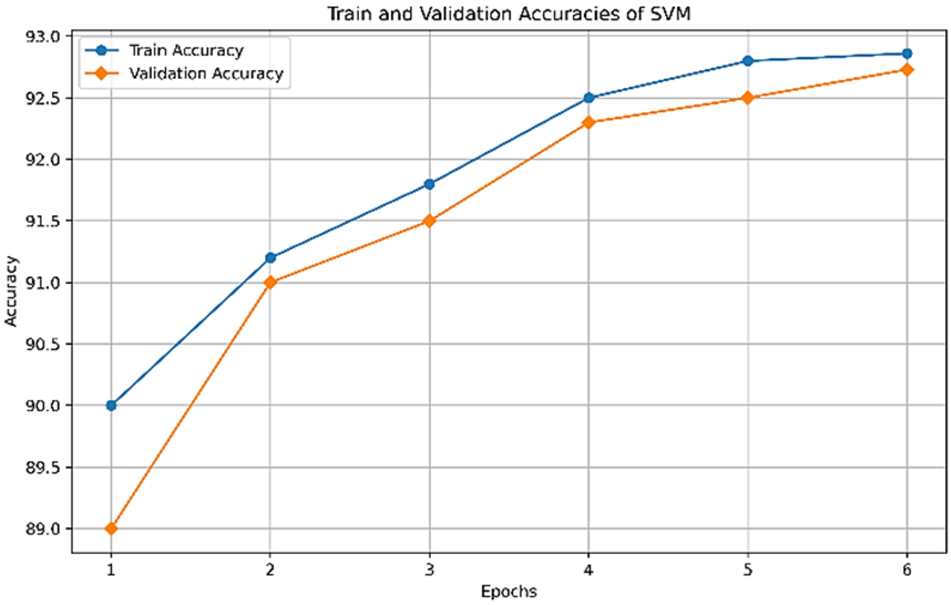

Figure 13. Train and Validation Accuracy of SVM

6. Conclusions

Using distinct language Roman Urdu and English comment, proposed methodology advances for multi classification of toxic comment Using SVM, Log Reg, and NB, tested their accuracy in identifying toxic comments. This study demonstrated the importance of algorithm selection, data refinement, and feature dimensionality to the accuracy of classification. Conducted pointed out the importance of using tailored approaches in multilingual environments, particularly when it comes to cultural nuances. In this methodology, you can foster a more inclusive and safer online environment by better understanding how toxic comments are classified. We believe that our findings can assist in reducing the impact of toxic content on the digital landscape and promoting a positive e-environment based on an understanding of how the landscape is changing. The study suggests future directions for toxic comment classification. To improve classification performance, ensemble methods combine multiple algorithms. Transformer models that provide even increased accuracy and robustness. Transformer models will be able to handle larger datasets as well as target more languages in the future.

References

1. “Online toxic comments: A growing problem" (The New York Times, 2022).
2. Pew Research Center from July 13-19, 2020
3. Saeed, H. H., Ashraf, M. H., Kamiran, F., Karim, A., & Calder, T. (2023). Toxic Comment Classification in Roman Urdu and English: A Two-Stage Pipeline Approach. arXiv preprint arXiv:2308.00001.
4. Zhang, J., Sun, X., & Liu, B. (2018). A Survey on Toxic Comment Classification. ACM Transactions on Information Systems, 36(4), 1-39.
5. Zhang, R., Niu, C., & Zhu, X. (2023). Toxic Comment Classification: A Perspective from Natural Language Processing. ACM Transactions on Knowledge Discovery from Data, 17(1), 1-29.
6. Belal, T. A., Shahariar, G. M., & Kabir, M. H. (2023). Interpretable Multi Labeled Bengali Toxic Comments Classification using Deep Learning.arXiv:2304.04087v1[cs.CL]. https://doi.org/10.48550/arXiv.2304.04087
7. Usman, M., Saad, M., & Ahsan, M. (2022). Toxic comment classification in Urdu: A novel approach based on word embeddings and transformer models. arXiv preprint arXiv:2203.07858.
8. 10.1007/s10579-021-09530-y
9. 'Roman Urdu Toxicity Classification Dataset: A Collection of Manually Classified-Comments'
10. http://archive.ics.uci.edu/ml/datasets/Roman+Urdu+Data+Set.html
11. "Comments Classification Dataset with Manually Classified Toxicity Labels from Haroon Shakeel under MIT License"
12. https://github.com/haroonshakeel/roman_urdu_hate_speech
13. https://drive.google.com/drive/folders/1wAEHueNheFnK4dJofejs87g-fI15Ks8u?usp=drive_link
14. Kaggle's Jigsaw Toxic Comment Classification dataset: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data
15. https://drive.google.com/drive/folders/1bBj1PQJyD5zqBar6TOkRQ8NFj_R6pHnO?usp=sharing