Pakistan Journal of Engineering and Technology, PakJET
ISSN (p): 2664-2042, ISSN (e): 2664-2050
Volume: 5, Number: 2, Pages: 198- 203, Year: 2022

Detecting Cyberbullying in Roman Urdu Language Using Natural Language Processing Techniques

Fahad Rasheed[1], Mehmoon Anwar[1], and Imran Khan[2]

[1]University of Engineering and Technology, Taxila, Pakistan
[2]International Islamic University, Islamabad, Pakistan

Corresponding author: Fahad Rasheed (e-mail: fahadrasheed677@gmail.com).

Received: 01/05/2022, Revised: 10/08/2022, Accepted: 10/09/2022

***Abstract*- Nowadays, social media platforms are the primary source of public communication and information. Social media platforms have become an integral part of our daily lives, and their user base is rapidly expanding as access is extended to more remote locations. Pakistan has around 71.70 million social media users that utilize Roman Urdu to communicate. With these improvements and the increasing number of users, there has been an increase in digital bullying, often known as cyberbullying. This research focuses on social media users who use Roman Urdu (Urdu language written in the English alphabet) to communicate. In this research, we explored the topic of cyberbullying actions on the Twitter platform, where users employ Roman Urdu as a medium of communication. To our knowledge, this is one of the very few studies that address cyberbullying behavior in Roman Urdu. Our proposed study aims to identify a suitable model for classifying cyberbullying behavior in Roman Urdu. To begin, the dataset was designed by extracting data from twitter using twitter's API. The targeted data was extracted using keywords based on Roman Urdu. The data was then annotated as bully and not-bully. After that, the dataset has been pre-processed to reduce noise, which includes punctuation, stop words, null entries, and duplication removal. Following that, features are extracted using two different methods, Count-Vectorizer and TF-IDF Vectorizer, and a set of ten different learning algorithms including SVM, MLP, and KNN was applied to both types of extracted features based on supervised learning. Support Vector Machine (SVM) performed the best out of the implemented algorithms by both combinations, with 97.8 percent when implemented over the TF-IDF features and 93.4 percent when implemented over the CV features. The proposed mechanism could be helpful for online social apps and chat rooms for the better detection and designing of bully word filters, making safer cyberspace for end users.**

Keywords -- Cyberbullying, Roman Urdu, social media, SVM, TF-IDF.

I. INTRODUCTION

With the passage of time, the worth of social media and internet site for info gathering has increased, as smartphones, laptops, tablets, and other time-saving devices that are useable anywhere with internet access have come into the market. These technological advancements have a significant impact on services such as communication, information gathering, payments, and shopping. Prior to the technological advancements, we used these services either directly or indirectly through acquaintances, coworkers, or family members. Likewise, we rely on digital platforms to communicate and connect with friends and colleagues, but it is difficult to filter out the negative impact of these platforms and keep them secure from such threats.

One of the major issues faced on social media platforms is un-ethical behavior of individuals which in other words can be referred to as cyberbullying. Cyberbullying is one of the most impactful negative aspects of the advancement in social media platforms. There are clear evidences that bullying can cause serious and long-term problems in individuals and according to World Health Organization (WHO) 'Health Behavior in School-Aged Children Survey', the overall frequency of perpetrators across the 35 countries involved was 11%, while bullies accounted for another 11%. Research revealed that approximately four children were harassed. Many of the adverse psychological and physical effects of suicide, depression, anxiety, cutting, negative feelings and psychosomatic symptoms were reported. There is plenty of work being done on automatic cyberbullying detection based on English language but there is still a lot of gaps on addressing different regional languages-based cyberbullying. Pakistan alone has over 70 million online social media users who communicate in Roman Urdu (Urdu written in English). With these stats, there arise an undeniable

need of a detection method that can detect cyberbullying based on Roman Urdu.

As of now, many researchers have published various techniques and methods for analyzing data related to bullying, specifically cyberbullying, and sentiment analysis using Machine Learning, Natural Language Processing, and Artificial Intelligence-based tools. All prior researches addressed the problem in hand in their respective languages and as we discussed, most of the research is based on English language. Our research study focuses on online media users who communicate in Roman Urdu (Urdu language written in English alphabets). In this paper, the major objectives were as follow:

- To Design a benchmark dataset for cyberbullying in Roman Urdu.
- To identify the best performing classifier for our dataset.

we addressed the issue of cyberbullying behavior on Twitter platform, where users use Roman Urdu as medium of their communication, which is essential for few reasons. Firstly, very limited research is done on detection of cyberbullying based in Roman Urdu. Secondly, there is also an absence of a benchmark dataset available in Roman Urdu, and our designed dataset and examination technique can help the social media interaction networks to construct a better filter mechanism to filter out the hatred content, which eventually will help to create a safe cyber-space where social media users are safe from these kinds of attackers and bullying.

II. RELATED WORK

Many researchers have contributed to this cause, with the majority of their findings relating to determining crime patterns from social media or Internet-based apps using sentiment analysis. In the majority of cases, the research was based on a single factor such as harassment, personal data leaks, extremist affiliations, and so on. The majority of the research was conducted in English, with very little research conducted on cyberbullying in Roman Urdu. The previous approaches differ from ours in several ways, most notably in the language used, so a brief summary of their research and findings is presented. Almutairi, Amjad Rasmi, and Muhammad Abdullah Al-Hagery (2021) presented a detection and classification model for cyberbullying on Twitter in the Arabic context in Saudi Arabia. They took the data from Twitter and classified it using PMI and SVM. PMI achieved 50% of the F1-score, whereas SVM achieved 82% [1]. Another research based on crime detection was done by H. Bouma et al., who used an anomaly detection technique to Turkish Kurdish data and 4daagse data (name of corpus utilized) [2]. The system was tested on Twitter, and the findings demonstrated that it can effectively evaluate messages and detect changes in sentiments. As of now, many researchers have published various techniques and methods for analyzing data related to bullying, specifically cyberbullying, and sentiment analysis using Machine Learning, Natural Language Processing, and Artificial Intelligence-based tools, as we reviewed, Kenedy Dende's research where he tried to produce a system being capable of categorizing and classifying data into positive, neutral and negative [3]. This research designed a sentiment analysis model embedded into a web-based application which is based on Naïve Bayes that classified sentiments into positive, neutral and negative. This research was more focused on building a cooperative association between intelligence agencies and the victim. Raja Ashok Bolla also did research on crime pattern recognition, collecting over 100,000 crime-related tweets over a 20-day span [4]. He used Emotion Analysis methods on these tweets to determine the crime concentration of a place based on geographic assessment. His sentiment classification was divided into two categories: binary sentiment classification and multi-class sentiment classification and he used Naive Bayes, Maximum Entropy, and Support Vector Machine to detect crime pattern in a location. In a data article in William et al. (2020) created the CLICK-ID dataset, which contains nearly 15,000 articles from twelve different news sites in Indonesia. The performance was then validated using Bi-LSTM and CNN models, yielding effective accuracy results [5]. Qian, Yu, et al. (2019) investigated the availability of massive amounts of data on everything online. The author proposed a method called "clustering annotation classification" in this work, which is a 3-Phase process for recognizing business events from internet headlines and articles. The retrieved data is utilized to categorize probable business events from online business events news headlines and leads after further relevant data is extracted and processed in a different sort of business events cluster. They discovered that WR clustering performs the best with an average precision of 64.08 percent, recall of 69.70 percent, and F-value of 63.62 percent [6]. By demonstrating posts authored by bullies, victims, and witnesses of online bullying, Van Hee et al., worked on automatic cyberbullying identification in social media content. To determine automatic cyberbullying detection, they used linear support vector machines (SVM) and ran a series of binary classification studies. After gathering data from ASKfm, English and Dutch corpora were produced. Binary classification tests utilizing SVM implemented in LIBLINEAR for the automatic identification of cyberbullying were carried out using Scikit-learn, a machine learning framework for Python. According to the findings, AUC scores are more resistant to data imbalance than recall, precision, or F score [7]. Kareem et al. (2019) suggested a technique for detecting false news in Pakistani media. They gathered the information from some of Pakistan's most prominent news websites. They used several algorithms for it, and as a consequence, they discovered KNN performing best with an accuracy of 70 percent [8].

In the Spanish language, cybernetic harassment in OSNs was proposed (Mercado et al., 2019). Sentiment analysis techniques such as bag of words, sign and number removal, tokenization, and stemming were used in this work. The Agustn Gravano (SDAL) database was used, which had a vocabulary of 2880 terms from the University of Buenos Aires' Faculty of Exact and Natural Sciences, Argentina [9]. Similar lexicon-based research is also carried out in foreign languages, as seen by the studies presented here. An approach for detecting hostile tweets in Spanish was published by (Gómez-Adorno et al., 2018). To balance the classification distribution, they used an oversampling methodology to train a logistic regression classifier using

linguistic patterns, aggressive words lexicon, and multiple types of n-grams. On training data, the classifier achieved an F-measure score of 42.85 for aggressive class, but the approach performed badly on testing data [10]. The lexical approach was used to evaluate Arabic language social media comments from YouTube and Twitter, which were based on a corpus of cyberbullying and violent phrases. The weighted function was used to categorize the comments into three categories: mild, medium, and strong. The method was shown to be effective in identifying the majority of cyberbullying comments [11].

There is a handful amount of research conducted on cyberbullying using deep learning algorithms. The authors hypothesised that deep learning algorithms would outperform the challenge. The GloVe840 word embedding approach, in conjunction with BLSTM, produced the best results on the dataset, which had 35,787 labelled tweets, with accuracy, precision, and F1 values of 92.60%, 96.60%, and 94.20%, respectively [12]. Another technique named as CNN-CB a convolutional neural network (CNN)-based model that adds semantics through word embedding, outperforms traditional content-based cyberbullying detection by 95% [13].

III. METHODOLOGY

The proposed methodology provides a clear depiction of our work. It briefly explains the course of the research, as shown in Fig. 1. As it can be seen, the very first phase is to get data from Twitter using API. The dataset of tweets was then built with the essential information and characteristics for the categorization into bully and not-bully.

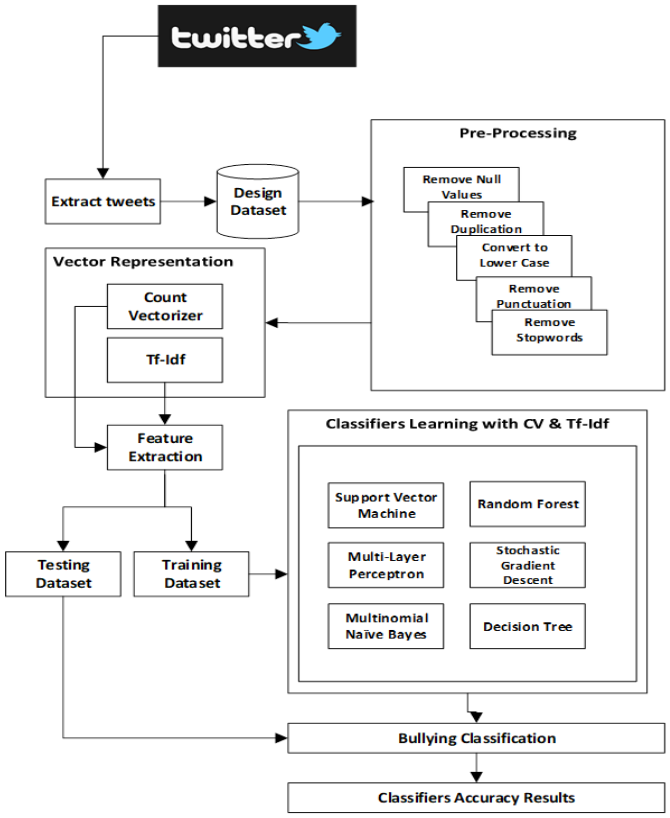

FIGURE 1. Proposed Methodology of Work

The next step in our methodology is pre-processing, which involves cleaning and removing noise from the dataset to make it more effective for analysis. After cleaning the dataset, we convert it into numerical or vector form using two different methods, CV and Tf-Idf, and afterwards the features of both kinds are split into training data, that is used for learning the classifier, and testing data, which is used for categorization testing, and finally we review the outcomes and make a comparison of the algorithms which performed better predicated on features extracted from both strategies.

A. DATASET DESIGNING

Data is the very first and the most important part for any research. Our dataset construction and designing were challenging because of the language we selected and the content available was very limited. We extracted the data (Tweets) from Twitter using Twitter's API. We collected 55,000 tweets. The tweets were extracted using keywords. To avoid any privacy violations, we retrieved only publicly available material using Twitter API and in accordance with Twitter network privacy and policies. For collecting the data based on specific keywords and language, we need a Twitter API Key which is basically the authorization key for an individual to fetch data. In order to collect data, we designed a chunk of code which asks for different keywords and the number of tweets as input and return the data based on the filters we applied. Afterwards the tweets were annotated as bully or not-bully. The tweets with bullying were labeled as true while the not-bully were labeled as false. The dataset was verified by native language speakers (mostly friends and colleagues) for the accurate labeling of each tweet and the tweet was marked as true or false based on vote. If a tweet gets 2 or more votes out of 3, only then it was marked as a specific category. All the tweets having mixed or balanced voting were discarded. The final dataset consists of 50000 tweets with 25000 bully and 25000 non-bully tweets.

B. DATA COLLECTION AND LABELLING

Gathering data is a vital step in Natural Language Processing. For our research, we were supposed to collect data from Twitter platform that include tweets based on cyberbullying in Roman Urdu. For using Twitter's data, we need to get the API access which grants you the permission and authority to fetch that data and use it in your research work. API stands for Application Programming Interface which basically act as a bridge between two applications and help process the request of one to another and vice versa. You may read and write Twitter data using the Twitter API. As a consequence, you may use it to produce tweets, read profiles, and access the data of your followers, as well as a vast number of tweets on specified topics and in specific regions. After collecting tweets data from Twitter via Twitter's API, we must design the dataset to include the necessary features for any dataset. The dataset must be labelled by Roman Urdu speakers so that bully and non-bully tweets can be distinguished. In our research, the data was labeled by our friends and colleagues. A Tweet was only labeled as a bully or not bully if it gets a higher number of votes from a total of an odd number. For example, a

Tweet is said to be a bully if it gets 2 out of 3 votes. All the data on which there was a conflict was deleted from the dataset.

C. DATA PRE-PROCESSING

Pre-processing is a Natural Language Processing task that involves cleaning the data in order to increase text classification performance using learning algorithms. It requires several cleaning techniques to make the dataset more productive for feature extraction and classification. The transformation of data into features that can be used to build a suitable classification model. It also improves the efficiency of learning algorithm training. Some of these steps include converting the text to lower case, removing stop words, removing punctuation, and so on. All these steps of pre-processing we performed during this step are to make our data noise-free and capable learning of algorithms efficiently to get more efficient and authentic results in the testing of algorithms. The pre-processing is the most important part of any work, plays a vital role to make the algorithm more efficient and accurate after cleaning the data. The sequence of pre-processing is depicted in Fig. 2.

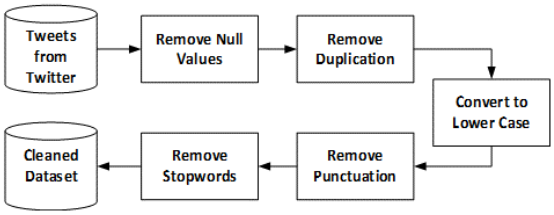

FIGURE 2. Flow of Data Preprocessing

D. REMOVING NULL VALUES

In this step of pre-processing, we analyzed the benchmark dataset and remove all of the null values from the dataset. To make learning classification techniques more appropriate. Several algorithms do not tolerate null values, which might lead to errors when learning and testing methods are used.

E. REMOVING DUPLICATION

Remove all duplicated items from the dataset during the pre-processing step to ensure that all of the dataset's entries are unique. If the majority of entries are repeated, the learning model may be biased towards that item in categorization.

F. CONVERTING TEXT TO LOWERCASE

In this step of processing, we convert the desired text to lowercase, which is more logical for a machine. It's done very easily in python, with a simple line of code. We convert only these features that are required for the training and testing of learning algorithms.

G. REMOVING STOP WORDS

Stop words removal is an important step in making the meaning of the sentence clearer to understand. Stop words include words like he, his, she, her, your, they, which, have, it, me, am, and so on. The removal of stop words improves the performance and meaning of sentences by leaving only meaningful words behind. The Google search engine, on the other hand, uses the stop words removal process to retrieve the search quickly.

H. REMOVING PUNCTUATION

After the removal of stop words meaningful tokens which are left, punctuation being another type of noise, elimination of punctuation cyphers makes the word more meaningful, which are difficult to process them to understand them for the learning algorithms and its leave an impact on the performance of any learning algorithm. So, this step of removing punctuation symbols can improve the performance.

I. FEATURE SELECTION

After the pre-processing of collected data, we can separate the features from our dataset, suitable for our model, which will be used for the training and testing of the suitable machine learning model for our work. A classifier can only work with numbers, not with text-based data, this must transform the textual data format into a vector data representation. There are several techniques for converting text to vector. vectorization using count vectorizer and Tf-Idf vectorizer discussed. We have also used these methods for vector representation of data. Flow is shown in Fig. 3.

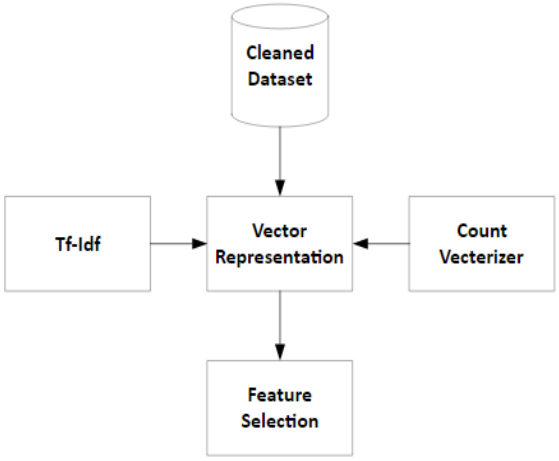

FIGURE 3. Implementation of Tf-Idf & CV Techniques for Feature Selection

IV. RESULTS AND DISCUSSION

Performance metrics are used to measure the performance of algorithms, there are different performance measures, some common (Accuracy, Recall, F1 Score, and Precision), that we used to analyze the performance and founded best algorithm regarding this work. We discuss all these measures based on the following actual and prediction-based metrics Table I.

TABLE I. EVALUATION METRICS

Evaluation Metrics		Actual	
		Positive	Negative
Prediction	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Accuracy is used to identify that how many predictions are completed by a classifier are predicted fittingly. Though, this measure is not much persuaded concerning the dataset which is not balanced due to the issues of biasedness of predictions

towards the category that is high in frequency and predict the other categories wrongly.

$$\boldsymbol{Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (1)$$

Where TP denotes True Positive, TN denotes True Negative, FN denotes False Negative, and TP denotes True Positive. Correspondingly, Precision, Recall, and F1-Score are used to assess classifier performance. These are far superior than accuracy in determining algorithm performance when the dataset is not adequately balanced, and their formulae are shown below.

$$\boldsymbol{Precision} = \frac{(TP)}{(TP + FP)} \quad (2)$$

$$\boldsymbol{Recall} = \frac{TP}{(TP + FN)} \quad (3)$$

$$\boldsymbol{F1\ Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

We discussed all these measures by confusion metric for all best performing algorithms and measured their performance.

A. CLASSIFICATION WITH TF-IDF

Tf-Idf classification is the first combination of a series of classifiers that employed features collected from the Tf-idf approach. This set contains 10 alternative learning algorithms, which are generally employed in different research works indicated in Table II and performed better in their category criteria. In this section, we implement this set of algorithms for Roman Urdu-based cyberbullying in order to discover the best method. In Table II, we reviewed each model using four distinct performance measures, including Precision, F1-Score, Accuracy, and Recall, which are used globally to compare and measure the algorithm's efficacy based on the stated metrics.

TABLE II. CLASSIFIERS' MEASURES WHEN USING TF-IDF FEATURES

Model	Accuracy	F1-score	Precision	Recall
Support Vector Machine	0.978802	0.978795	0.979464	0.978802
MLP	0.971909	0.971904	0.972307	0.971909
Logistic Regression	0.963838	0.963811	0.965343	0.963838
SGD	0.959644	0.959584	0.962521	0.959644
Bagging	0.951096	0.951026	0.953815	0.951096
Decision Tree	0.945156	0.945129	0.946113	0.945156
Extra Tree	0.918788	0.918755	0.919409	0.918788
Random	0.916090	0.916078	0.916299	0.916090
Multinomial Naïve Bayes	0.874578	0.873595	0.886320	0.874578
KNN	0.792595	0.789097	0.813779	0.792595

When we evaluate the classifiers based on their metrics, we can see that SVM is the best-performing algorithm in this combination as well, with an accuracy of 97.8% when we use Tf-Idf features for classification. The second best-performed algorithm is MLP with 97.1% of accuracy. The model which performed worst in this combination is the KNN Classifier with an accuracy of 79.2%. SVM is also the best-performing algorithm in terms of F1-score, precision, and recall, with 97.8 percent, 97.9 percent, and 97.8 percent, respectively.

B. CLASSIFICATION WITH CV

Likewise, this section consists the second blend of algorithms implemented over Count Vectorizer features. The entire bunch of algorithms and their evaluation metrics, which include Precision, Accuracy, F1-score, and Recall. Mentioned in Table III.

TABLE III. CLASSIFIERS' MEASURES WHEN USING CV FEATURES

Model	Accuracy	F1-score	Precision	Recall
Support Vector Machine	0.935563	0.936090	0.937082	0.935563
SGD	0.931008	0.931543	0.932492	0.931008
MLP	0.916676	0.916543	0.916425	0.916676
Logistic Regression	0.900567	0.897825	0.898972	0.900567
Bagging	0.889901	0.890688	0.891802	0.889901
Decision Tree	0.873236	0.874097	0.875241	0.873236
Extra Tree	0.870570	0.864695	0.867865	0.870570
Random	0.870348	0.865332	0.867091	0.870348
Multinomial Naïve-Bayes	0.777469	0.718026	0.807380	0.777469
KNN	0.923461	0.91609	0.907082	0.915563

When we evaluate the classifiers based on their metrics, we can see that SVM is the best-performing algorithm in this combination as well, with an accuracy of 93.5% when we use CV features for classification. The second best-performing algorithm is SGD with 93.1% of accuracy. The model which performed worst in this combination is MNB with an accuracy of 77%. SVM is also the best-performing algorithm in terms of F1-score, precision, and recall, with 93.6 percent, 93.7 percent, and 93.5 percent, respectively.

V. RESULTS COMPARISON

Now, in Fig. 4, we compare the accuracy of both combinations (Tf-Idf and CV features) to find the best suitable algorithm for Roman Urdu based Cyberbullying classification by equating both variations that we have implemented using two different feature extraction approaches. Figure 4.5 depicts a graphical representation of the comparison.

When we compare both results, we found the best algorithm for our work is SVM with the highest accuracy of 97.8% from both of the implemented combinations over the whole set of algorithms we applied.

VI. CONCLUSION AND FUTURE WORK

The main findings of our research are provided below:

- Designed a benchmark dataset in Roman Urdu Language.
- Identified the best classifier for cyberbullying classification in Roman Urdu.

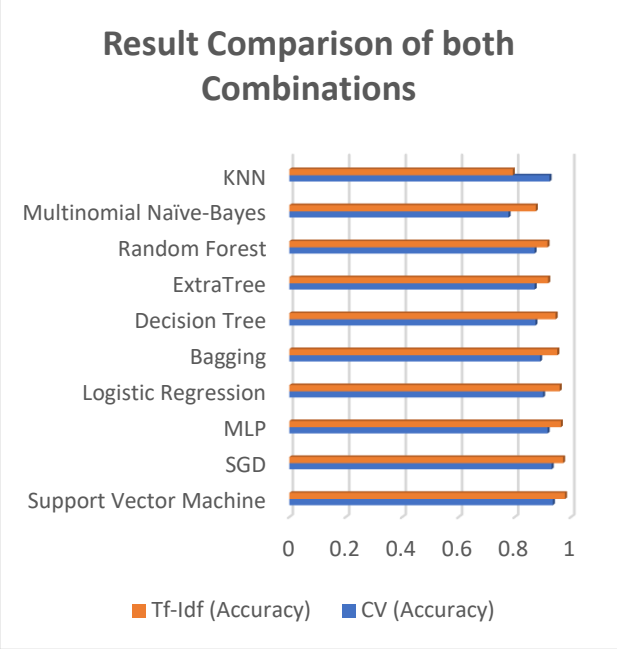

FIGURE 4. Comparison of Both Combinations

A benchmark dataset that will help to concur the need of advance word filtering mechanism for the Roman Urdu based communications and post on social media. Further, the classifiers obtained a high percentage of accuracy with SVM outperforming all with an accuracy of 97.8%.

In future, we can improve our research by incorporating more data into our constructed dataset to obtain more precise and reliable results with high accuracy. Preprocessing can also contribute to enhance accuracy when we use stemming and lemmatization, which can also help to get more relevant features by using various feature extraction strategies and play a critical role in the improved performance of our proposed technique. For enhancements, we can also use convolutional neural networks and deep learning transformer models. Moreover, there are many other future pathways in the domain of cyberbullying identification, with many of these research studies focusing solely on the English language, but there are innumerable regional and international languages, and thus most internet users interact with one another via social media platforms using these languages. Another potential future direction is to enhance features derived from social media sites and subjected to cyberbullying classification models.

Our proposed methodology can help the different online social platforms to better understand and design a filter mechanism for cyberbullying detection in Roman Urdu. As Pakistan alone has over 70 million Online Social Media users who use Roman Urdu as a medium of interaction and communication. This study can prove to be a major filter mechanism to prevent bullying in cyberspace. We believe our model and dataset can prove to be exemplary in assisting automatic cyberbullying detection in any social media interaction platform.

ACKNOWLEDGMENT

I am grateful to everyone at my university, UET Taxila, as well as my friends and family members, for making this work possible and for making my graduating experience one that I will remember for the rest of my life. In addition, I would like to give a special mention to Kazim Raza Talpur and my teachers including Dr. Imran Khan and Mr. Mehmoon Anwar, for their encouragement and assistance throughout the project.

FUNDING STATEMENT

The authors received no specific funding for this study.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] Almutairi, Amjad Rasmi, and Muhammad Abdullah Al-Hagery. "Cyberbullying Detection by Sentiment Analysis of Tweets' Contents Written in Arabic in Saudi Arabia Society." International Journal of Computer Science & Network Security vol. 21, no.3, pp. 112-119, 2021.

[2] Bouma, Henri, et al. "On the early detection of threats in the real world based on open-source information on the internet." Information Technologies and Security. 2012.

[3] K. Dende, "Sentimental Analysis in crime detection: A case study of Kenya law enforcement agencies," University of Nairobi, 2014.

[4] Bolla, Raja Ashok. Crime pattern detection using online social media. Missouri University of Science and Technology, 2014.

[5] William, Andika, and Yunita Sari. "CLICK-ID: A novel dataset for Indonesian clickbait headlines." Data in brief vol. 32, pp. 106231, 2020.

[6] Y. Qian, X. Deng, Q. Ye, B. Ma, and H. Yuan, "On detecting business event from the headlines and leads of massive online news articles," Information Processing & Management, vol. 56, p. 102086, 2019.

[7] Van Hee, Cynthia, et al. "Automatic detection of cyberbullying in social media text." PloS one vol. 13, no.10, pp. e0203794, 2018.

[8] Kareem, Irfan, and Shahid Mahmood Awan. "Pakistani media fake news classification using machine learning classifiers." 2019 International Conference on Innovative Computing (ICIC). IEEE, 2019.

[9] Montufar Mercado, Rolfy Nixon. "Automatic cyberbullying detection in spanish-language social networks using sentiment analysis techniques." 2019.

[10] Gómez-Adorno, Helena, et al. "A Machine Learning Approach for Detecting Aggressive Tweets in Spanish." IberEval@ SEPLN. 2018.

[11] Mouheb, Djedjiga, et al. "Detection of arabic cyberbullying on social networks using machine learning." 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA). IEEE, 2019.

[12] Bharti, Shubham, et al. "Cyberbullying detection from tweets using deep learning." Kybernetes 2021.

[13] Al-Ajlan, Monirah Abdullah, and Mourad Ykhlef. "Deep learning algorithm for cyberbullying detection." International Journal of Advanced Computer Science and Applications vol. 9, no. 9, 2018.