International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8167
***Article Received:** 29 June 2023 **Revised:** 26 August 2023 **Accepted:** 08 September 2023*

Comparative Performance of Data Mining Techniques for Cyberbullying Detection of Arabic Social Media Text

Omar Kamal Eldien Hussien[1], Amal Elsayed Aboutabl[2], Riham Mohamed Younis Haggag[3]
[1]Business Information Systems Department, Faculty of commerce,
Helwan university, Cairo, Egypt,
Omar.Kamal-Eldien21@commerce.helwan.edu.eg
[2]Computer Science Department, Faculty of Computers & Artificial Intelligence, Helwan University, Cairo, Egypt
amal.aboutabl@fci.helwan.edu.eg
[3]Business Information Systems Department, Faculty of Commerce and Business Administration,
Helwan university, Cairo, Egypt
reham.mohamed.younis@commerce.helwan.edu.eg

Abstract- Cyberbullying has spread like a virus on social media platforms and is getting out of control. According to psychological studies on the subject, the victims are increasingly suffering, sometimes to the point of committing suicide among the victims. The issue of cyberbullying on social media is spreading around the world. Social media use is growing, and it can have useful and negative implications when you take into account how social media platforms are abused through different forms of cyberbullying. Although there is a lot of cyberbullying detection in English, there are few studies in the Arabic language. Data Mining techniques are often used to solve and detect this problem. In this study, different data mining algorithms were used to detect cyberbullying in Arabic texts.. Our study was conducted The Bullying datasets consisted of 26,000 comments written in Arabic and were collected from kaggle.com, the Cyber_2021 dataset consisted of 13,247 comments collected via github.com, and the Data 2022 dataset consisted of 47,224 comments collected via Instagram. Various extraction features CountVectorizer and Tf-Idf were used Accuracy, precision, recall, and the F1 score were used to evaluate classifier performance. In the study, Bagging Classifier achieve high results of Bullying dataset from Kaggle Accuracy 96.04, F1-Score 95.98, Recall 96.04, Precision 95.95, SVC model gave the highest results of Cyber_2021 dataset from Github an Accuracy 98.49, F1-Score 98.49, Recall 98.49, Precision 98.50, while Data 2022 dataset from (Instagram) achieving an Accuracy of 77.51, F1-Score 76.60, Recall 77.51, and Precision 77.24. Were achieved for Tf-Idf Vectorizer. Tf-Idf Vectorizer the best to all results than count Vectorizer .

Keywords: Cyberbullying detection, Social media, Data mining techniques, CountVectorizer, Tf-Idf..

I. Introduction

Cyberbullying is a type of violence committed by people or organizations using electronic media. Violence can take the form of intimidation, humbling, insulting, or mocking. Email threats, insults in social media comments, and uploading embarrassing images of someone are all instances of cyberbullying behavior.[21]

cyberbullying is harassment carried out using digital technology. The detrimental effects of traditional bullying and cyberbullying are real.

Bullied children, necessitate parental and educational intervention and the adoption of preventative measures. Many nations have developed legal frameworks and policy frameworks to address and curtail cyberbullying behaviors.[19]

Cyberbullying is the practice of humiliating or threatening other people using modern technology, such as cell phones, email, chat rooms, or social networking sites like Twitter or Facebook.[9] Victims of cyberbullying, particularly young individuals, may experience significant effects. Age affects one's capacity to control emotional reactions, according to research.[10] Between 41.7% and 46.7% of young adults use the Internet for social networks.[11]

Cyberbullying is one of the most prevalent types of Internet abuse and a significant social problem, especially for teens. As a result, an increasing number of research are concentrating on ways to spot and eradicate cyberbullying, particularly on social media. Cyberbullying includes assuming a false identity, disseminating a humiliating photograph or video, spreading unfavorable remarks about another person, and even posing a danger. Awful consequences of cyberbullying on social media, including instances when unfortunate victims pass away, are terrifying.[8]

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8167
***Article Received:** 29 June 2023 **Revised:** 26 August 2023 **Accepted:** 08 September 2023*

Arabic Language Over 300 million Arabs worldwide speak Arabic, which is primarily the mother tongue of Muslims. Unlike the English language, the Arabic alphabet is read and written from right to left. Arabic uses 28 alphabet letters and additional unique punctuation known as diacritical marks. Arabic is a challenging and complex language because of its morphological structure. Without a prior understanding of Arabic,

it might be challenging to identify proper nouns in phrases because Arabic does not use upper- and lowercase letters. Additionally, Arabic letters can be written in a variety of ways depending on where they appear in words; there are typically two or three alternative ways to write each letter.[18]

In this paper, models that can be proposed and compared will be presented. Used to detect cyberbullying. In particular, we compare the performance of data mining techniques in analyzing the Arabic language to detect cyberbullying. This study is how to select some 10 classification models. Two features Extract Count Vectorizer and TF_IDF Vectorizer.

II. Related Work

In this section, a literature review in the area of Arabic cyberbullying detection using data mining techniques classifiers is researched.

In [20] have classified cases of cyberbullying on an Arabic comments dataset using convolutional and recurrent neural networks along with Arabic pertained word embedding. Additionally, they have contrasted machine learning models' performance with that of deep learning models and found that the latter exhibit competitive performance with deep learning.

In [22] the authors present the results of machine learning algorithms used to categorize the sentiment of Twitter posts are presented. Regarding certain emoticons used in Twitter communications, they categories tweets as either positive or negative.

In [23] The authors use a variety of Naive Bayes and Maximum Entropy Models, in addition to other well-known machine learning approaches, to tackle the issue of tweet sentiment analysis. Based on error analysis and feelings that are particular to the distinctive rhetoric and language of Twitter, they also performed some optimizations.

III. Data Mining Techniques

Data mining is crucial for a variety of purposes, including pattern recognition, forecasting, learning, and others. Data mining techniques and algorithms like classification, clustering, and others are frequently used to uncover patterns that can be used to predict future business trends. Data mining is recognized as one of the most significant frontiers in database and information systems and one of the most promising multidisciplinary advances in information technology due to the vast variety of application domains it has almost in every industry where the data is created. that approach Several algorithms and methods, such as Classification, Regression, Clustering, Association Rules, Neural Networks, Genetic Algorithms, Nearest Neighbor, Decision Trees, etc.,. [1]

3.1 K- Nearest Neighbor

K-NN is the most basic machine learning algorithm. Using a collection of training samples that are physically close to the new point, the method's basic premise is to anticipate the label. The number of samples may rely on the local point density or it may be a user-defined constant. For length measurements, any metric unit is acceptable. The most common approach for determining the separation between two points is the standard Euclidean distance. Numerous classification and regression issues, like those involving handwritten numbers or the processing of satellite images, have been addressed using the Nearest Neighbors technique.[3]

3.2 A Neural Network

A Neural Network is made up of input/output (I/O) units, each of which connects with a weight. During the learning phase, the net modifies the weights and even forecasts precise row classifications for input groups. ANN is particularly adept in interpreting murky or faulty data. As an example, computer training that speaks English text after rearranging handwritten characters can be used to extract patterns and reveal exceedingly intricate patterns that are unseen to humans or other computer technologies. [1]

3.3 Random Forest

Techniques based on Random Forests are used. Regression and classification are used. While training, it creates numerous decision trees. the new case into a category, it sends it to every tree in the system. After categorization, each tree generates a class. The largest set of classes that are comparable to one another and are generated by several trees is the output class, and it is decided by a majority voting system.

Since minimal programming or study is required, both experts and laypeople can quickly learn how to use Random Forests. [3].

3.4 Support Vector Machine

The SVM algorithm can represent each piece of data as a point in space using an n-dimensional space (where n is the number of properties), using the value of each property that has been assigned a coordinate value. Support vector machines are supervised machine learning techniques that can be used for classification or regression issues. Classification can be done when the hyper-plane that best describes the two classes has been determined. A support vector machine is a directed learning method that may be applied to classification or regression tasks in the context of machine learning.[4]

3.5 Logistic Regression

This model, a linear one that has evolved into a vital tool for multiclass classification and is recognized as a statistical approach utilized in many studies of machine learning and data mining, is one of the most often used models in the field of machine learning.[5]

3.6 Naïve Bayes (NB)

A popular supervised learning method based on statistics and the Bayes theorem is called naive Bayes. By calculating conditional probabilities using the education dataset, text documents are classified using this technique. The ease of usage and effectiveness in resolving categorization issues are the main advantages of naive Bayes.[7]

3.7 Bagging classifier

Each class result is then given to numerous selection procedures, which are thought of as being the same as using multiple classifiers in the estimation process. [15] .

3.8 XGBoost

It is a community learning method built on decision trees, just like traditional gradient enhancement models. It was unveiled in 2016 and is thought to be new. It differs from previous approaches due to its scalability, which enables speedy learning through parallel and distributed computing and provides efficient memory utilization. Both bias and excessive fitting are absent. It is ideal for straightforward implementation because it has decent performance and extensive documentation.[17]

3.9 Classification

The most common data mining technique is classification, which creates a model from a set of previously categorized samples that can categories most data. A classification algorithm analyses the training data when learning. The accuracy of the classification rules is estimated using classification test data. [1]

The two steps in the data classification process are learning, which involves the creation of a classification model, and classification, which involves using the model to predict the classes for a set of data. So, test tuples and the corresponding class labels are combined to form a test set. They were not incorporated into the classifier generation because they are independent of the training tuples. The proportion of test set tuples that a classifier properly recognizes as belonging to a certain test set is used to determine the classifier's accuracy. For each test tuple, the learned classifier's class prediction and the corresponding class label are contrasted. [2]

3.10 Predication

Regression analysis can be used to make predictions. A technique for simulating the relationship between a number of independent factors and dependent variables is regression analysis. While the goal of data mining is to anticipate response variables, independent variables are characteristics that are previously known.[1]

IV. Methodology

This study is referred to as a descriptive-analytical study that focuses on social media commenters. Historical data is acquired, organized, and then presented in an easy-to-understand manner using descriptive analytics. Only historical business events are the focus of descriptive analytics. In contrast to other analysis methodologies, it does not conclude or make predictions from its results. This study compares the effectiveness of data mining methods for identifying cyberbullying messages from social media platforms using Arabic dataset content.

4.1 Proposed Model

Proposed method can be divided into three main phases: input, processing, and output, respectively The initial data pre-processing is cleaned and sorted using tokenization, stemming, and stop words during data collection. The data is then divided into training and testing pools, classification with different algorithms are selected with data mining techniques, Evaluation Measurements by accuracy, precision, recall, and the F1 score were used to evaluate classifier performance

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8167
***Article Received:** 29 June 2023 **Revised:** 26 August 2023 **Accepted:** 08 September 2023*

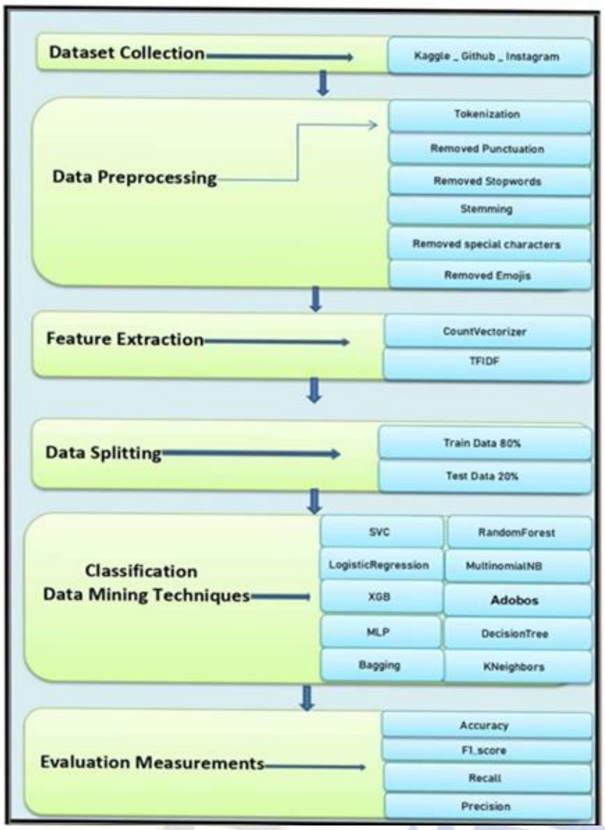


Fig1. The proposed model

4.2 Dataset

The dataset collected by corpus for cyberbullying detection algorithm. For this purpose, initially used to three groups dataset are CSV (comma separated value) the

1- Bullying dataset available at kaggle.com dataset (https://www.kaggle.com/datasets/alanoudaldealij/arabic-cyberbullying-tweets?resource=download) There are total of 26,000 records out of which 2,591 contain Bullying 23,408 Non Bullying.

2- The Cyber_2021 dataset available at (https://github.com/omammar167/Arabic-Abusive-Datasets) There are total of 13,247 records out of which 6,860 contain cyber 6,387 not cyber.

3-The Data 2022 dataset Instagram available at (https://bit.ly/3Md8mj3) There are total of 47,225 records out of which 12,569 contain Bullying, 17,376 Positive, 11,343 Neutral , 5937 Toxic

4.3. Data Preprocessing

The data Preprocessing techniques utilized in this study include eliminating URLs and emoticons. Since reading comments or URLs cannot be used to analyze a tweet. It may result in over fitting as well. Other pre-processing techniques include Stemming, Tokenization, Removing stop words, and Removing punctuation and letters. In data preprocessing, it is aimed to clean the texts from the dataset according to the following procedures :

4.3.1 Removed punctuation

Remove punctuation is used to reduce the quantity of the data, and to delete any unnecessary information such as ('!"#$%&\'()*+,-./: <=>?@[\\]^_`{|}~').

Figure1 Removed punctuation

4.3.2 Tokenization

Tokenization is a data handling technique where a given text is divided into tokens.

Figure2 Tokenization

4.3.3 Removed Stopwords

Stopwords removal is a measure taken to remove words that are unnecessary or meaningless. Whether a word contains the stop word or not, the data dictionary serves as a reference such as (e.g., a, and, the)

Figure 3 Removed Stopwords

4.3.4 Removed emojis

They might also appear in a collection of emojis divided just by Items that could be arranged in a tweet without any interstitial gaps. Emojis were translated to text and implemented as a feature since they are an indispensable part of contemporary text found on social media.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8167
***Article Received:** 29 June 2023 **Revised:** 26 August 2023 **Accepted:** 08 September 2023*

Figure 4 Removed Emojis

4.3.5 Stemming

Even though there are fewer word types and classes in the data, stemming can nonetheless return words to their original form.

Figure 5 Stemming

4.3.6 Cleaned Texts

After removing unnecessary expressions and coordinating the texts, It is obtained cleaned texts.

5. Feature Extraction

Feature extraction is the process of transforming unprocessed data into a useful resource. Clustering is a step in the feature extraction process that includes organizing the obtained data into groups based on their features. By breaking down the text into base characteristics, this technique significantly decreases the quantity of data that needs to be processed while accurately describing the original data. For feature extraction, a variety of methods may be employed.[13]

Feature extraction is used for dimensionality reduction. In this paper, had chosen to use two feature extraction approaches: Count Vectorizer and TF_IDF Vectorizer. It transforms text data into a machine-readable format. Count Vectorizer produces a vector that is encoded and has the same length as the remark and includes an integer count of the number of times each word appears in the comment. TF_IDF Vectorizer which stands for Term Frequency – Inverse Document Frequency (TF) is calculated by the following formulas, for a term T of document D in the document.

V. Classification Method

Arabic-language social media posts were subjected to data mining algorithms in order to identify cyberbullying messages. A binary categorization system has been developed in which any message including cyberbullying is coded as 1 and any message not comprising it as 0. The data set is separated into twelve (10) Data Mining methods algorithms and 80% Training Data and 20% Test Data. (Decision Tree Algorithm, K - Nearest Neighbors Algorithm, Bagging Ensemble Algorithm, Adaptive Boosting Algorithm, XGBoost Algorithm, Logistic Regression Algorithm, Random Forest Algorithm, Multinomial

Naive Bayes Algorithm, Support Vector Classification Algorithm and Neural Network Algorithm) were classified for the purpose of finding cyberbullying. The results of the other 10 algorithms have been processed and used in the Classifier Model algorithm.

VI. Evaluation Measurements

According to the overall classification and estimate,the final result will be produced. This suggestedapproach's effectiveness is assessed by using techniquessuch as Classifier accuracy is a measure of the tool'sThe predictor's accuracy assesses the overallproportion or ratio of classified documents to the totalof cyberbullying that is actually occurring and isaccurately classified, as well as non-cyberbullying.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

By dividing the total number of correct results by the total number of expected results, recall is calculated. Sensitivity in binary classification is referred to as recall. You may think of it as the probability that the query would yield a relevant document.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Precision is calculated by dividing the total number of true positives by the total number of true positives + false positives.

F1_ score is weighted average of accuracy and recall scores is expressed as. F1 scores range from 0 to 1, with 1 being the worst. It is indicated by:

$$\mathrm{F1}\_Score = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

VII. Results and Discussion

As shown in Table 1, used Accuracy, F1-Score, Recall, and Precision to get the results of the performance of ten data mining classifiers on three Arabic datasets. That RandomForestClassifier did better with count Vectorizer feature extraction with the Bullying dataset from (kaggle.com) achieving Accuracy 95.62, F1-Score 95.55, Recall 95.62,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8167
***Article Received:** 29 June 2023 **Revised:** 26 August 2023 **Accepted:** 08 September 2023*

Precision 95.51 While, Multinomial Naive Bayes did the best with count Vectorizer feature extraction with the Cyber_2021 dataset from Github achieved Accuracy 97.55, F1-Score 97.55, Recall 97.55, Precision 97.56. While, Logistic Regression did the best with count Vectorizer feature extraction with the Dataset 2022 from Instagram achieving Accuracy 76.77, F1-Score 76.05, Recall 76.77, and Precision 76.22.

As shown in Table 2 found that Bagging Classifier did better with TF_IDF Vectorizer feature extraction with the Bullying dataset from (kaggle.com) achieving Accuracy 96.04, F1-Score 95.98, Recall 96.04, Precision 95.95 While, whereas Support Vector Classifier did the best with TF_IDF Vectorizer feature extraction with the Cyber_2021 dataset (Github) achieved Accuracy 98.49, F1-Score 98.49, Recall 98.49, Precision 98.50. While, Support Vector Classifier did the best with TF_IDF Vectorizer feature extraction with the Dataset 2022 from (Instagram) achieving an Accuracy of 77.51, F1-Score 76.60, Recall 77.51, and Precision 77.24.

Overall, in regards to feature extraction, the models slightly give the Accuracy, F1-Score, Recall, and Precision better results TF_IDF Vectorizer is used with three datasets to compare count Vectorizer feature extraction.

TABEL 1. Result For Datasets Using Count Vectorizer

No.	Dataset	Algorithms	Accuracy	F1-Score	Recall	Precision
1	The Bullying dataset (Kaggle)	RandomForestClassifier	95.62	95.55	95.62	95.51
		MLPClassifier	95.48	95.44	95.48	95.41
		BaggingClassifier	95.42	95.41	95.42	95.4
		LogisticRegression	95.15	94.75	95.15	94.91
		DecisionTreeClassifier	94.92	94.92	94.92	94.92
		MultinomialNB	94.56	94.04	94.56	94.22
		XGBClassifier	94.52	94.04	94.52	94.15
		AdaBoostClassifier	94.35	93.84	94.35	93.95
		SVC	94	93.14	94	93.77
		KNeighborsClassifier	92.19	90.38	92.19	91.71
2	The Cyber_2021 dataset (github)	MultinomialNB	97.55	97.55	97.55	97.56
		LogisticRegression	97.24	97.24	97.24	97.27
		MLPClassifier	96.97	96.97	96.97	96.97
		SVC	96.08	96.08	96.08	96.21
		XGBClassifier	95.46	95.46	95.46	95.55
		RandomForestClassifier	94.68	94.66	94.68	94.89
		AdaBoostClassifier	93.17	93.17	93.17	93.38
		BaggingClassifier	93.05	93.03	93.05	93.19
		DecisionTreeClassifier	92.58	92.56	92.58	92.77
		KNeighborsClassifier	64.66	58.66	64.66	77.34
3	The Dataset2022 Instagram	LogisticRegression	76.77	76.05	76.77	76.22
		SVC	75.99	74.04	75.99	76.16
		MultinomialNB	75.17	75.47	75.17	76.08
		XGBClassifier	74.73	72.53	74.73	74.74
		MLPClassifier	74.65	74.55	74.65	74.47
		RandomForestClassifier	74.45	72.98	74.45	73.8
		BaggingClassifier	71.96	71.02	71.96	70.99
		DecisionTreeClassifier	70.44	69.84	70.44	69.62
		AdaBoostClassifier	70.11	67.64	70.11	68.74
		KNeighborsClassifier	68.07	62.1	68.07	67.1

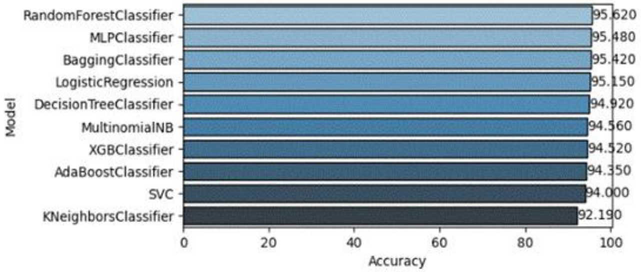
International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8167
***Article Received:** 29 June 2023 **Revised:** 26 August 2023 **Accepted:** 08 September 2023*

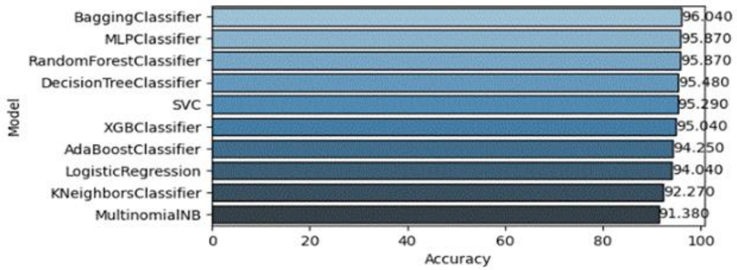
TABEL 2. Result For Datasets Using Tf-Idf Vectorizer

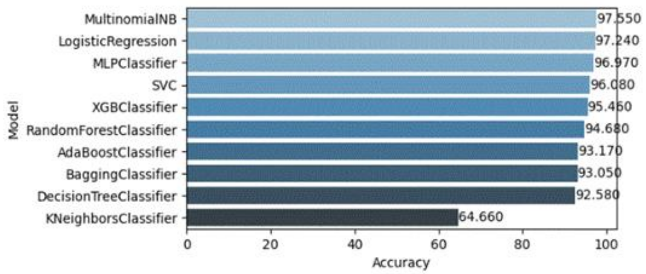
No.	Dataset	Algorithms	Accuracy	F1-Score	Recall	Precision
1	The Bullying dataset (Kaggle)	BaggingClassifier	96.04	95.98	96.04	95.95
		MLPClassifier	95.87	95.83	95.87	95.8
		RandomForestClassifier	95.87	95.73	95.87	95.69
		DecisionTreeClassifier	95.48	95.47	95.48	95.46
		SVC	95.29	94.87	95.29	95.1
		XGBClassifier	95.04	94.65	95.04	94.76
		AdaBoostClassifier	94.25	93.78	94.25	93.82
		LogisticRegression	94.04	93.18	94.04	93.84
		KNeighborsClassifier	92.27	90.5	92.27	91.84
		MultinomialNB	91.38	88.57	91.38	91.6
2	The Cyber_2021 dataset (github)	SVC	98.49	98.49	98.49	98.5
		LogisticRegression	97.79	97.79	97.79	97.82
		MultinomialNB	97.75	97.75	97.75	97.76
		MLPClassifier	97.2	97.2	97.2	97.25
		XGBClassifier	96	96	96	96.07
		RandomForestClassifier	95.73	95.72	95.73	95.88
		BaggingClassifier	94.1	94.09	94.1	94.16
		AdaBoostClassifier	94.06	94.06	94.06	94.14
		DecisionTreeClassifier	93.55	93.55	93.55	93.62
		KNeighborsClassifier	55.22	42.52	55.22	75.97
3	The Dataset2022 Instagram	SVC	77.51	76.6	77.51	77.24
		LogisticRegression	76.52	75.68	76.52	76.09
		MultinomialNB	75.36	73.36	75.36	75.78
		RandomForestClassifier	74.37	72.87	74.37	73.97
		MLPClassifier	74.32	74.12	74.32	74
		XGBClassifier	73.27	70.77	73.27	73.46
		BaggingClassifier	72.28	71.07	72.28	71.43
		AdaBoostClassifier	70.47	67.59	70.47	69.82
		DecisionTreeClassifier	70.03	69.57	70.03	69.38
		KNeighborsClassifier	68.23	62.59	68.23	68.35

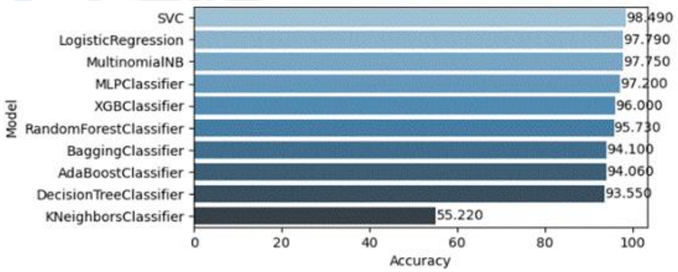
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8167
***Article Received:** 29 June 2023 **Revised:** 26 August 2023 **Accepted:** 08 September 2023*

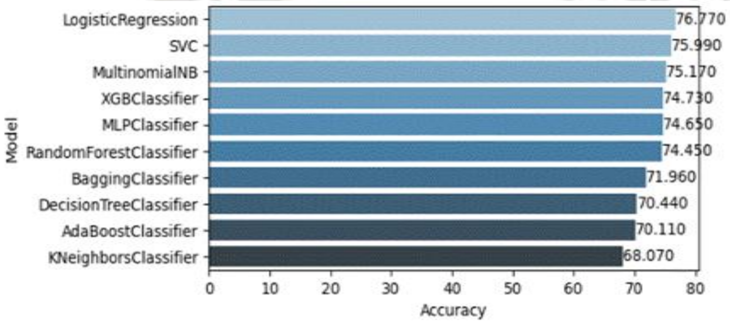
TABEL 3. Comparative For Result Accuracy by Using Count Vectorizer and Tf-Idf Vectorizer

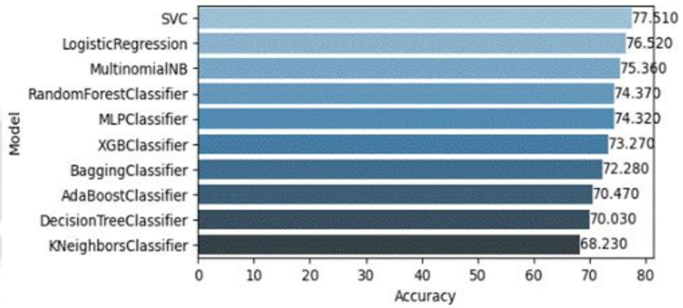
<table>
<tr><th>No</th><th>Dataset</th><th>Count Vectorizer (Accuracy)</th><th>TF_IDF (Accuracy)</th></tr>
<tr><td>1</td><td>The Bullying dataset (Kaggle)</td><td>


</td><td>


</td></tr>
<tr><td>2</td><td>The Cyber_2021 dataset (github)</td><td>


</td><td>


</td></tr>
<tr><td>3</td><td>The Dataset2022 Instagram</td><td>


</td><td>


</td></tr>
</table>

VIII. CONCLUSION

Social media Users have the opportunity to express their sentiments and ideas on a range of issues. Some individuals use social media for malicious purposes, engaging in behaviors such as cyberbullying and expressing hatred, insults, and threats toward other users. Because conducted a cyberbullying study conducted entirely in Arabic on three datasets collected from trusted websites and made use of several data mining techniques, it was determined that it differs from many cyberbullying studies that have been conducted. Mining Algorithms (SVC, LogisticRegression, RandomForestClassifier, MultinomialNB, XGBClassifier, MLP Classifier, BaggingClassifier ,AdaBoostClassifier ,Decision Tree Classifier ,KNeighborsClassifier, Furthermore two feature extraction approaches were used and studied Count Vectorizer and TF_IDF Vectorizer. Results were compared with three datasets in terms using Accuracy, F1-Score, Recall, and Precision. TF_IDF Vectorizer is the best feature extraction approach.

REFERENCES

[1] Bharati, M. & Ramageri. (2010). “data mining technique applications,” Indian Journal of Computer Science and Engineering, Vol. 1 No. 4, pp. 301-305.

[2] Verma, J. Data Mining in Indian Railways (2021). A Survey to Analyze Applications of Data Mining. International Journal of Computer Applications, 975, 8887.

[3] Baid, P., Gupta, A., & Chaplot, N. (2017). Sentiment analysis of movie reviews using machine learning techniques. International Journal of Computer Applications, 179(7), 45-49.

[4] Abdulrahman, S. A., Khalifa, W., Roushdy, M., & Salem, A.-B. M., “Comparative study for computational intelligence algorithms for human identification. Computer Science Review,” 36, 100237, pp.1-11, 2020.

[5] Ghosh, R., Nowal, S., & Manju, G. (2021). Social media cyberbullying detection using machine learning in bengali language. Int J Eng Res Technol.

[6] Ali, R. T., & Kurdy, M. B. Cyberbullying Detection in Syrian Slang on Social Media by using Data Mining. (May 2021).

[7] A. McCallum and K. Nigam, “A comparison of event models for naive bayes text classification”, in Workshop On Learning For Text Categorization, July 1998, pp. 41-48.

[8] Prajakta Ingle, Ramya Joshi, Neha Kaulgud, Aarti Suryawanshi, Meghana Lokhande (2021); Cyber bullying monitoring system for Twitter; International Journal of Scientific and Research Publications (IJSRP) 11(4) (ISSN: 2250-3153), DOchttp://dx.doi.org/10.29322/IJSRP.11.04.2021.p11273

[9] R. M. Kowalski, S. P. Limber and P. W. Agatston, Cyberbullying: Bullying in the Digital Age, West Sussex: Wiley-Blackwell, 2012.

[10] V. Orgeta, “Specificity of age di_erences in emotion regulation,” Aging and Mental Health, 13(6), 818–826, 2009, doi:10.1080/13607860902989661.

[11] A. Weinstein, D. Dorani, R. Elhadif, Y. Bukovza, A. Yarmulnik, P. Dannon, “Internet addiction is associated with social anxiety in young adults,” Annals of Clinical Psychiatry, 27(1), 4–9, 2015.

[12] Cheng, L., Li, J., Silva, Y. N., Hall, D. L., & Liu, H. (2019). XBully: Cyberbullying Detection within a Multi-Modal Context. 19, 339–347. https://doi.org/10.1145/3289600.3291037

[13] A. Rajaraman, J. D. Ullman,(2011) “Data Mining,” Mining of Massive Datasets, 1–17, , doi:doi:10.1017/CBO9781139058452.002

[14] Hadžiosmanović, D., Simionato, L., Bolzoni, D., Zambon, E., & Etalle, S. (2012). N-gram against the machine: On the feasibility of the n-gram network analysis for binary protocols. In Research in Attacks Intrusions, and Defenses: 15th International Symposium, RAID 2012, Amsterdam, The Netherlands, September 12-14, 2012. Proceedings 15 (pp. 354-373). Springer Berlin Heidelberg.

[15] Rajamohana, S. P., Dharani, A., Anushree, P., Santhiya, B., & Umamaheswari, K. (2023). Machine learning techniques for healthcare applications: early autism detection using ensemble approach and breast cancer prediction using SMO and IBK. In Research Anthology on Medical Informatics in Breast and Cervical Cancer (pp. 386-402). IGI Global.

[16] Yang, X. S. (2019). Introduction to algorithms for data mining and machine learning. Academic press.

[17] Rufaida, S. I., Leu, J. S., Su, K. W., Haniz, A., & Takada, J. I. (2020). Construction of an indoor radio environment map using gradient boosting decision tree. Wireless Networks, 26, 6215-6236.

[18] Haidar, B., Chamoun, M., & Serhrouchni, A. (2019, July). Arabic cyberbullying detection: Enhancing performance by using ensemble machine learning. In 2019 international conference on internet of things (ithings) and ieee green computing and communications (greencom) and ieee cyber, physical and social computing (cpscom) and ieee smart data (smartdata) (pp. 323-327). IEEE.

[19] Unicef. (2020). Cyberbullying: What is it and how to stop it. Retrieved from unicef. org: https://www. unicef. org/end-violence/how-to-stop-cyberbullying.

[20] Rachid, B. A., Azza, H., & Ghezala, H. H. B. (2020, July). Classification of cyberbullying text in Arabic. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.

[21] Kurniawanda, M. R., & Tobing, F. A. T. (2022). Analysis Sentiment Cyberbullying In Instagram Comments with XGBoost Method. IJNMT (International Journal of New Media Technology), 9(1), 28-34.

[22] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12), 2009.

[23] Parikh, R., & Movassate, M. (2009). Sentiment analysis of user-generated twitter updates using various classification techniques. CS224N final report, 118, 1-18.