ENSEMBLE LEARNING BASED PREDICTION FOR CYBER HARASSMENT OBSERVATIONS ON TWEETS

ABSTRACT

Now a days social media plays crucial role in allowing individuals to express their views. Based on their views find the information of different keywords/statements like sadness, happiness, teasing, harassment and abuse. Online abuse, a novel form of pestering, have become increasingly predominant in online groups in modern civilization. Detecting harassment is indeed a significant challenge. Several studies provide information on cyber harassment, but none of them offer a solid remedy. Several studies provide information on cyber harassment, but none of them offer a solid remedy. Due to this reason, multiple models can be practiced to recognize and block harassment-related communications. We have utilized ensemble machine learning models to predict accurate results. The twitter dataset used for our research. We observe two models getting accuracy for RF+DT is 92% and SVM+LR is 93%. It is similar accuracy in individual models. So, there is no difference between Ensemble or individual model accuracy rate.

Keywords: Ensemble Machine Learning, Prediction, Cyber Harassment, Tweets.

I. INTRODUCTION

A collection of Web 2.0-based programmers called social media make it possible to create and share user generated content. These are all Internet-based applications. People may take use of social media to gain access to a wealth of knowledge, easy communication, etc. [1]. Cyber bullying is the term used to describe aggressive, deliberate acts committed by a person or group of individuals against a victim using digital communication channels like sending messages and leaving comments online [2].

Speech that is intended to stir up hatred for a specific group—a community, a religion, or a race—is referred to as hate speech. Worldwide increases in violence against minorities, such as lynchings, mass shootings, and ethnic cleansing, have been connected to hate speech on the Internet [3]. Simple word filters do not adequately address this issue, necessitating natural language processing that focuses on this symptom: What constitutes hate speech can be

influenced by factors. The model is trained using the Tweeter dataset from Kaggle. We must initially use a single categorization algorithm to move further with these datasets. We utilized the 0-1 predictor to determine if the text contains cyber bullying material or not [4]. This creates a binary space in which we can train our model and exclude out any grey possibilities. In order to properly classify data, it must first be cleaned of symbols, spacy tokenizer Addresses, mails, line breaks, spaces, digits, commas, separating, and individual characters [5]. Together with an incisive analysis of some published research on methods for detecting cyber bullying, this study offers a thorough and organized overview of robotic incitement identification and examines a few of the existing methodologies [6].

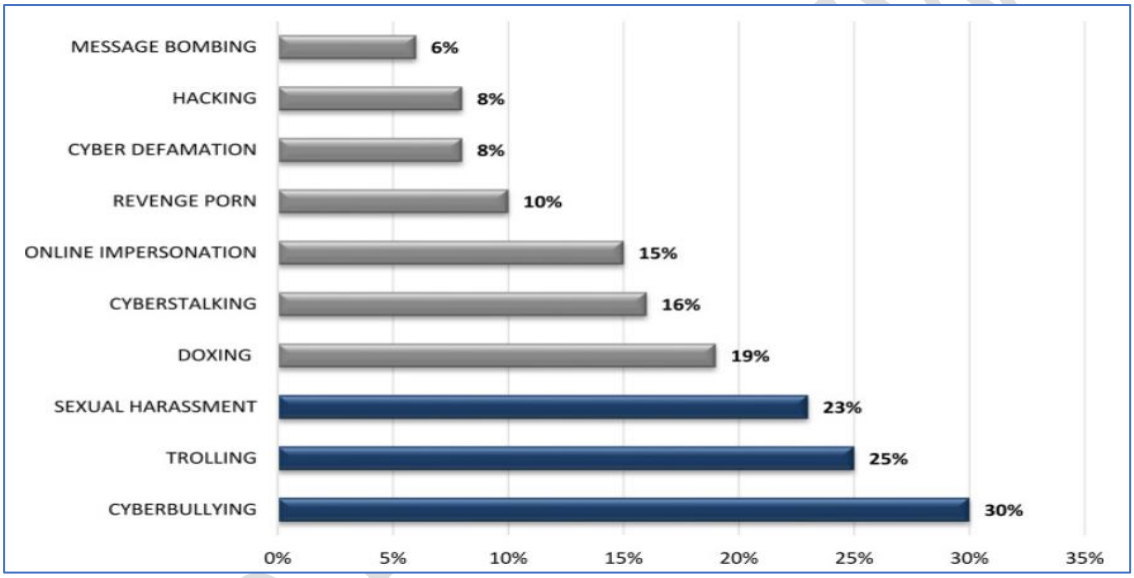

Figure 1: Types of Cyber harassment user experience [7]

From figure 1 shows the types of cyber harassment on user experience and figure 2 shows the impact of cyber harassment on age diversity.

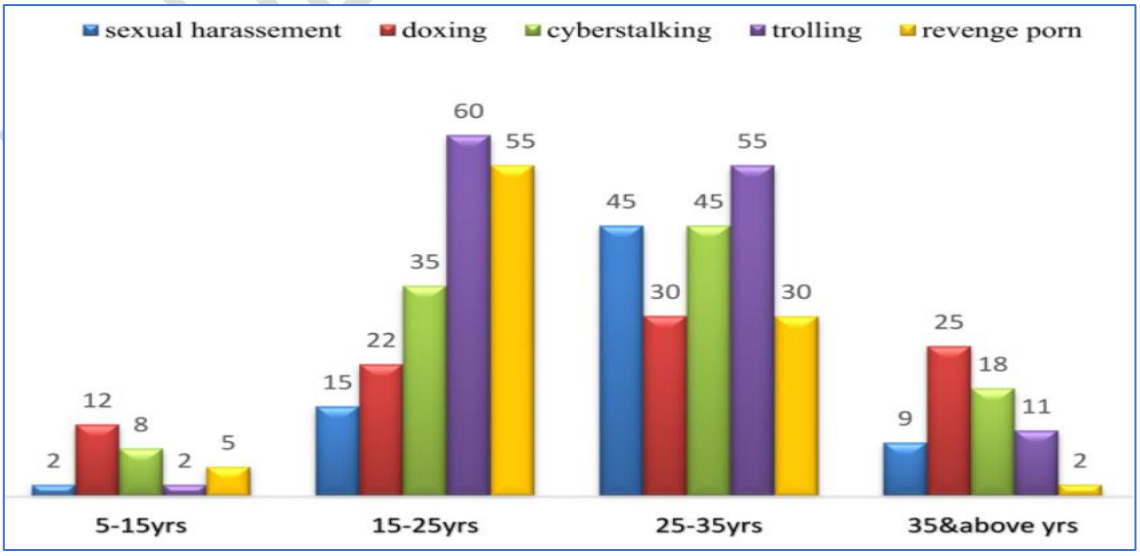

Figure 2: Impact of cyber harassment on age diversity [8]

[3] information on how to post a petition of bigotry on Facebook using an assistance from a deep neural convolutional network. With the help of machine learning algorithms, tweets containing hate speech have been found Utilizing the TensorFlow(tf) procedure, functionalities on Facebook have now been removed. The best ml model is SVM, however in a 4:1 sample used to evaluate trained predictions, it was capable of forecasting 53% of racial hatred messages [9]. [10] comprehensive analysis in newly implemented harassment on Facebook. Moreover, the significance of recognizing the numerous Facebook offenders is discussed. According to the Research report, there are a number of concrete measures that must be taken in order to construct a useful and successful software for detecting Online activity. I use characteristic types, ml models, and knowledge categorization and data logging. [11] proposes the process for achieving them identify & stop digital abuse-controlled ml techniques employing identified on Facebook. Inside this experiment, texts and sample sizes are compiled using the real time [11]. The suggested model evaluates SVM and Bayesian Network on the gathered data sets. Use the TFIDF vectorizer to delete a feature. The findings demonstrate the accuracy of a model for internet abuse constructed using Vector Assist. In comparison to Naïve Bayes classifier, the computer performs around 73.34% superior [12].

The following paper continue next section with proposed architecture. Section three discuss about results and analysis. Last section concludes the paper.

II. PROPOSED ARCHITECTURE

The following diagram provide the process of cyber harassment along with internal model. This is a sample diagram for social media messages how internally reach to society.

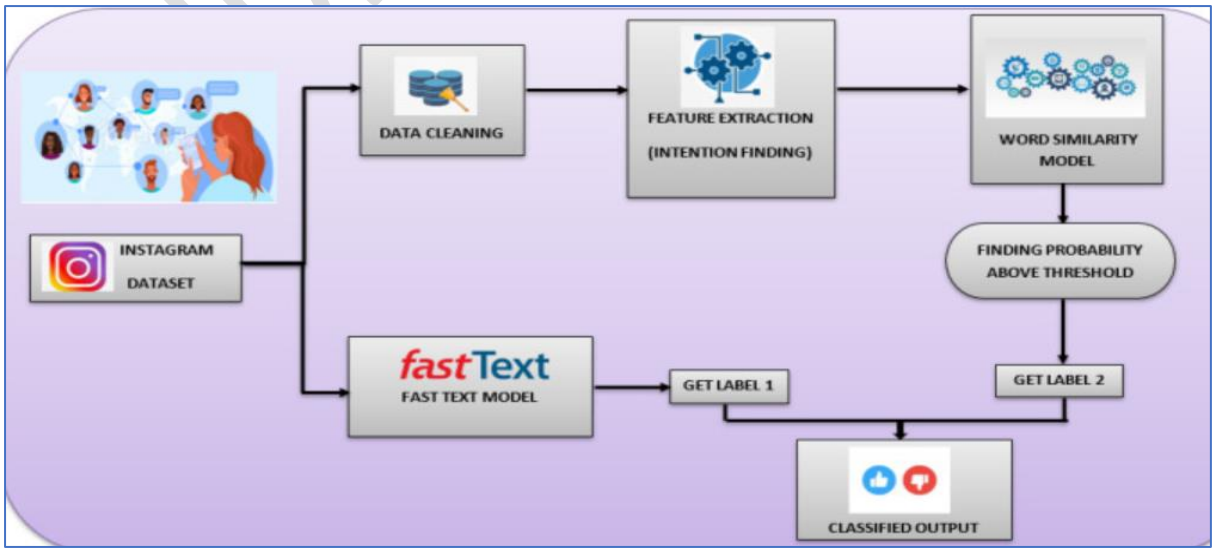

Figure 3: The process for detecting online harassment along with its intention model

Although social networking sites and online chat services give users a place to share their skills and information, they are seldom used to threaten other users with cyber harassment, which makes it difficult to use these services. In this research, we created a strategy for supervised learning to identify cyber harassment. The logistic regression approach It's utilized to test a algorithm for ml on the Random subset and improve it, which has been gathered with features and labels. Online abuse/ bullying detection is a growing area of research that aims to automatically detect instances of cyberbullying in online interactions. These calculations are a widely used Stats model which can be applied to cyberbullying detection. The scope of cyberbullying detection using logistic regression is large and involves several stages of Preparing the data, choosing the features, building the model, testing it, and deploying it: In this work, a message can be detected whether it is hating speech or not.

This proposed scheme is an early version of a cyberbullying detection system that can be attached to social networking sites to clearly detect and keep track of cyberbullying. Data collection: The program would gather information from websites and social networking sites like YouTube, Google, and Pinterest. Text, picture, and video data types are all possible. This collected data would be processed as follows:

1. Pre-processing of data: The collected data will be cleaned, normalized, and pre-processed to remove irrelevant information and ensure that it is in a format suitable for analysis.
2. Feature extraction: relevant features are extracted from the preprocessed data. These features may include linguistic features such as the use of profanity, hate speech, and aggressive language, and behavioral features such as the frequency and timing of online interactions.
3. Feature selection: The most relevant features are selected for training the logistic regression model. In this step, the features that have the greatest impact on the outcome variable, i.e., whether an online interaction is cyberbullying or not, is identified.
4. Model training: the logistic regression gathering all the necessary chosen by parameters is used to train the model. To discover the link between both the attribute values and the output vector, the computer must be trained.
5. Evaluation of this model: the performance of the logistic regression model is evaluated using parameters including highest accuracy, recollection, & accuracy. These metrics

help determine the effectiveness of the model in correctly identifying cases of cyber bullying.

6. Deployment: once trained and evaluated, the logistic regression model can be deployed to identify cases of cyber bullying in real time. The model can be integrated into online platforms and social media networks to identify and flag cases of cyber bullying.

The proposed system for detecting cyber bullying using logistic regression would include collecting and preprocessing data, extracting and selecting relevant features, training and evaluating the model, and using the model to detect cyber bullying cases. The system can be a valuable tool for preventing and mitigating the harmful effects of cyber bullying [9].

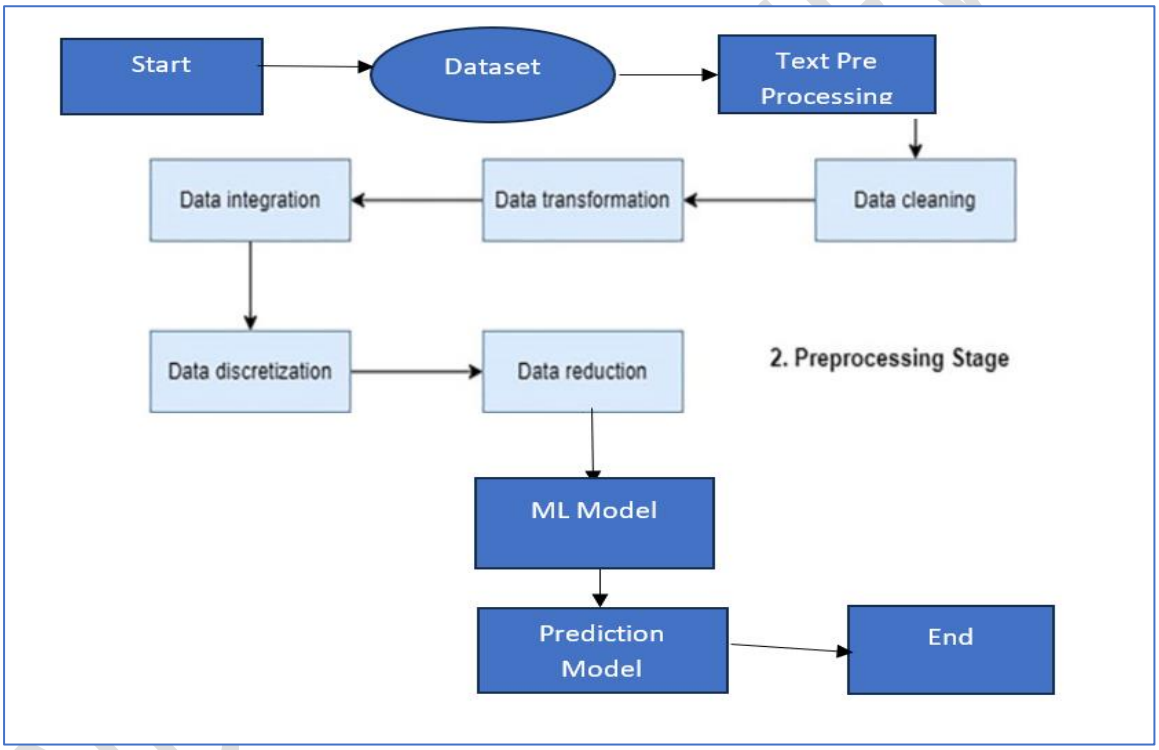

Figure 4: Architecture Model

III. RESULTS AND ANALYSIS

Creating algorithms and statistical models that enable computers to automatically learn from data and improve their performance on a particular activity without being explicitly taught is known as machine learning, and it is a subset of artificial intelligence. The ability to learn from examples and experience is what machine learning enables computers to do, which improves their ability to predict the future or take action in response to that knowledge. This is often done by training a model on a dataset. In this process, the model learns to recognize relationships and patterns in the data and then uses this understanding to predict or make decisions about brand-new, unseen data. Various machine learning applications such as fraud

detection, autonomous driving, language comprehension, speech and image recognition, and recommendation systems utilize a diverse range of algorithms and methodologies to address increasingly intricate challenges in the field.

3.1 Importing necessary libraries and dataset

Importing necessary libraries and upload dataset has been executed in the current project session.

```
Saving cyberbullying_tweets.csv to cyberbullying_tweets.csv
religion               7998
age                    7992
gender                 7973
ethnicity              7961
not_cyberbullying      7945
other_cyberbullying    7823
Name: cyberbullying_type, dtype: int64
```

Figure 5: Types of tweets

There is not much imbalance between different cyberbullying type. Other cyberbullying will be removed since it may cause a confusion for the models with other cyberbullying class.

3.2 Dataset Preprocessing

Data preprocessing is a procedure of making the raw data and construction it appropriate for a machine learning model. It is the primary and vital step while making a machine learning model.

	text	sentiment
9217	fuck no that bitch dont even suck dick &#128...	gender
18954	I apologize Mahmut, but there about 40 million...	religion
42057	Listen here you five handed dumb nigger I'll f...	ethnicity
12335	Things you shouldn't say: 1. Retarded 2. The N...	gender
45409	@tayyoung_: FUCK OBAMA, dumb ass nigger<&lt...	ethnicity
13115	Gay jokes and rape jokes were never funny	gender
37188	those girls bullied me in high school and now ...	age
19862	Christian friends, if you're outraged by this ...	religion
10508	Did you know that it's possible to criticize v...	gender
7032	do ppl really care if they're the first mother...	not_cyberbullying

Figure 6: Sample tweets

3.2.1 Converting categories into numbers

The parameters consider as input, converted into numbers.

```
df["sentiment"].replace({"religion": 1, "age": 2, "gender": 3, "ethnicity": 4, "not_cyberbullying": 5}, inplace=True)

sentiments = ["religion","age","gender","ethnicity","not bullying"]
```

Preprocessing: Tokenize sentences, change to lower case, Correct spelling, remove numbers, remove punctuation, remove stop words, normalize (Lemmatize or Lemmatization)

3.2.3 Predefined functions for text cleaning

In preprocessing context drop the few lines based on redundancy, missing values and abnormal values.

	text	sentiment	text_clean
0	In other words #katandandre, your food was cra...	5	word katandandr food crapilici mkr
1	Why is #aussietv so white? #MKR #theblock #ImA...	5	aussietv white mkr theblock today sunris studi...
2	@XochitlSuckkks a classy whore? Or more red ve...	5	classi whore red velvet cupcak
3	@Jason_Gio meh. :P thanks for the heads up, b...	5	meh p thank head concern anoth angri dude twitter
4	@RudhoeEnglish This is an ISIS account pretend...	5	isi account pretend kurdish account like islam...

3.2.4 Checking tweet duplicates

There are around 1000 duplicates. We will remove them at the next cell. After removing duplicates, the value counts per sentiment is shown above. There are only a few differences (350) on the sentiment with most text and the sentiment with least text so there is no major imbalance on the data.

3.2.5 Checking tweet length

Checking tweet length also play key role in this pre-processing for quick generating the results. The following figure 7 shows the tweets count.

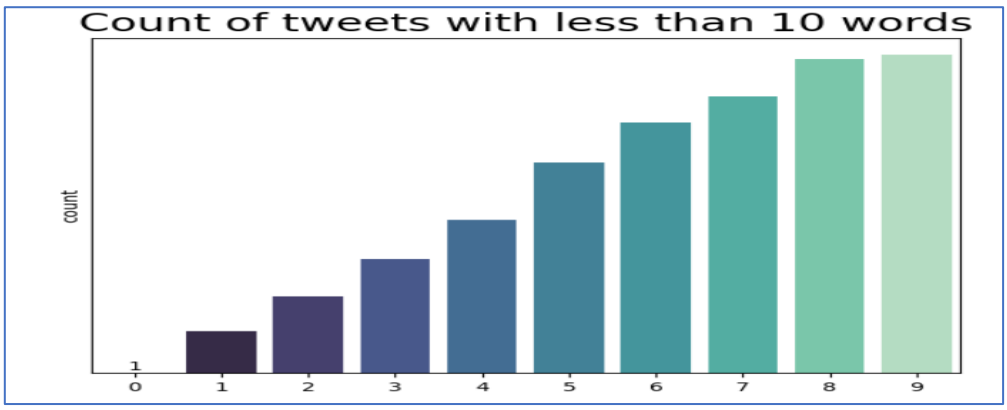

Figure 7: Graph for tweets length checking

	text	sentiment	text_clean	text_len
44035	You so black and white trying to live like a n...	4	black white tri live like nigger pahahahaha co...	187
45165	@hermdiggz: “@tayyoung_: FUCK OBAMA, dumb ass ...	4	fuck obama dumb ass nigger bitch ltthi whore s...	162
33724	... I don't feel guilty for killing him, I jus...	2	feel guilti kill feel guilti enjoy torment sin...	137
1317	@EurekAlertAAAS: Researchers push to import to...	5	research push import top antibulli program us ...	137
47037	@Purely_Ambition: Sooo mad. RT @TracePeterson ...	4	sooo mad rt fuck obama dumb nigger go switzerl...	125
...	...	...	...	...
1607	@harmlesstree2 Here7 https://t.co/xWJzpSodGj	5	here7	1
6696	@LiamTighe Rebecca who?	5	rebecca	1
558	@root_tim this is my work :)	5	work	1
3462	@jaredchase killing you how?	5	kill	1
10	@Jord_ls_Dead http://t.co/UsQlnYW5Gn	5		0

38820 rows × 4 columns

Figure 8: Tweets length count based on user id (Before)

We should place a condition for removing tweets with less than or equal 4 words and more than 100 words as they can be considered as outliers. The following figures 9 before applying condition for remove tweets and figure 10 after removing tweets based on condition.

	text	sentiment	text_clean	text_len
0	In other words #katandandre, your food was cra...	5	word katandandr food crapilici mkr	5
1	Why is #aussietv so white? #MKR #theblock #ImA...	5	aussietv white mkr theblock today sunris studi...	10
2	@XochitlSuckkks a classy whore? Or more red ve...	5	classi whore red velvet cupcak	5
3	@Jason_Gio meh. :P thanks for the heads up, b...	5	meh p thank head concern anoth angri dude twitter	9
4	@RudhoeEnglish This is an ISIS account pretend...	5	isi account pretend kurdish account like islam...	8
...	...	...	...	...
47687	Black ppl aren't expected to do anything, depe...	4	black ppl expect anyth depend anyth yet free p...	21
47688	Turner did not withhold his disappointment. Tu...	4	turner withhold turner call court abomin concl...	28
47689	I swear to God. This dumb nigger bitch. I have...	4	swear god dumb nigger bitch got bleach hair re...	13
47690	Yea fuck you RT @therealexel: IF YOURE A NIGGE...	4	yea fuck rt your nigger fuck unfollow fuck dum...	10
47691	Bro. U gotta chill RT @CHILLShrammy: Dog FUCK ...	4	bro u got ta chill rt dog fuck kp dumb nigger ...	13

37114 rows × 4 columns

Figure 9: Tweets length count based on user id(After)

5.3 Creating a word cloud

Set the numbers based on user parameters.

0- Religion

1- Age

2- Gender

3- Ethnicity

4- Cyberbullying

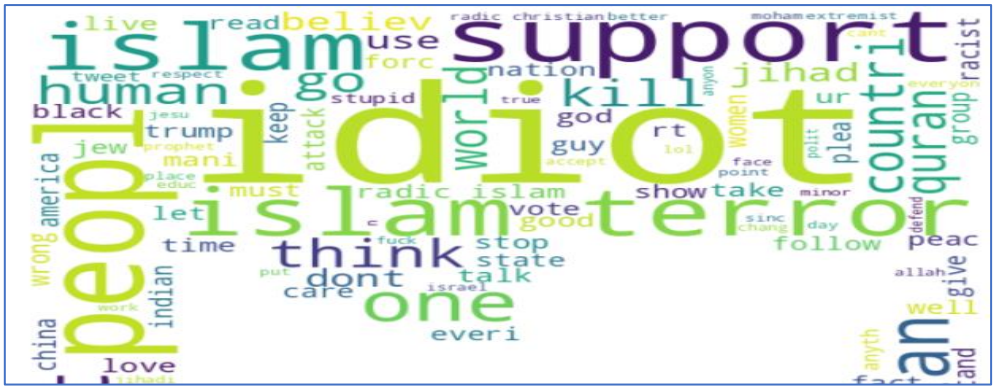

Figure 10**:** Different types of Tweets words

3.3.1 Turning text into numbers

Train test split of the data, after checking the training set, there are imbalance of data and it may cause a bias in training the model. Oversampling using SMOTE (synthetic minority oversampling**)** will be used to balance the train dataset. Now the training set is balanced and ready to be feed into the ml model

3.4 Model Training

After data splitting into train and test, train the model based on data.

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import accuracy_score, confusion_matrix, f1_score
from sklearn.metrics import classification_report
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_validate, cross_val_score
from sklearn.svm import SVC
from sklearn import metrics
```

3.5 ML Model Selecting and Prediction

3.5.1 Model Building

Now, we are ready to build our model for prediction, for that I made function for model building and preforming prediction and measure its prediction and accuracy score. The following Arguments are

1. model => ML Model Object

2. Feature Training Set data
3. Feature Testing Set data
4. Targeted Training Set data
5. Targeted Testing Set data

Let's make a dictionary for multiple models for bulk predictions

Before, sending it to the prediction check the key and values to store its values in Data Frame below.

3.5.2 Model Implementing

Now, Train the model one by one and show the classification report of particular models wise.

3.5.2.1 Logistic Regression

Logistic regression predicts the 93 percent accuracy rate.

Classification Report of 'LogisticRegression '

	precision	recall	f1-score	support
0	0.97	0.94	0.95	1579
1	0.96	0.97	0.96	1566
2	0.95	0.86	0.91	1462
3	0.98	0.97	0.98	1542
4	0.76	0.87	0.81	1274
accuracy			0.93	7423
macro avg	0.92	0.92	0.92	7423
weighted avg	0.93	0.93	0.93	7423

3.5.2.2 Random Forest Tree

Random Forest Tree predicts the 92 percent accuracy rate.

Classification Report of 'RandomForestClassifier '

	precision	recall	f1-score	support
0	0.94	0.94	0.94	1579
1	0.95	0.98	0.96	1566
2	0.93	0.86	0.89	1462
3	0.98	0.96	0.97	1542
4	0.79	0.83	0.81	1274
accuracy			0.92	7423
macro avg	0.92	0.92	0.92	7423
weighted avg	0.92	0.92	0.92	7423

3.5.2.3 Decision Tree

Decision tree classifier predicts the 92 percent accuracy rate.

Classification Report of 'DecisionTreeClassifier '

	precision	recall	f1-score	support
0	0.95	0.93	0.94	1579
1	0.98	0.97	0.97	1566
2	0.90	0.87	0.89	1462
3	0.98	0.98	0.98	1542
4	0.77	0.82	0.80	1274
accuracy			0.92	7423
macro avg	0.92	0.92	0.92	7423
weighted avg	0.92	0.92	0.92	7423

3.5.2.4 Support Vector Machine

Decision tree classifier predicts the 93 percent accuracy rate.

3.6 Comparative Study

In our research, we have to practice four machine learning algorithms. Among these algorithms focus on accuracy classification of data. SVM and Logistic Regression generate 93 percentage and Decision Tree and random forest tree generate 92 percentage.

Table 1: Comparison of accuracy

S, No.	**Name of the Classifier**	**Accuracy Rate (%)**
1	Logistic regression	93
2	Random Forest Tree	92
3	Decision Tree	92
4	Support Vector Machine	92

A confusion matrix is a table used in machine learning and statistics to assess the presentation of a classification model. It showing the counts of true positive, true negative, false positive, and false negative predictions. The following confusion metrics graphs for different metrics.

Figure 11: Confusion Matrix for multiple metrics

6.7 Ensemble models

Ensemble learning is a technique of hybrid machine learning that improves accuracy and resilience in forecasting by merging predictions from different models. In ensemble learning two types of models practice for getting better accuracy. The two combinations are

1. Decision Tree + Random Forest
2. SVM + Logistic Regression

6.7.1 Decision Tree + Random Forest

6.7.2 SVM + Logistic Regression

If you observe these two models getting accuracy for RF+DT is 92% and SVM+LR is 93%. It is similar accuracy in individual models. So, there is no difference between Ensemble or individual model accuracy rate.

IV. CONCLUSION

Online abuse is a novel kind of pestering, that latterly become more predominant as online groups. In modern civilization detect harassment as it true. Several studies provide information on cyber harassment, but none of them prevent for solid remedy. Due to this reason multiple models can be practice for recognize and block harassment related communications. We have used ensemble machine learning models to predict accurate result. The twitter dataset used for our research. We observe two models getting accuracy for RF+DT is 92% and SVM+LR is 93%. It is similar accuracy in individual models. So, there is no difference between Ensemble or individual model accuracy rate.

REFERENCES

1. P. Pranathi, V. Revathi, P. Varshitha, Subhani Shaik and Sunil Bhutada," Logistic Regression Based Cyber Harassment Identification", Journal of Advances in Mathematics and Computer Science, Volume 38, Issue 8, Page 76-85, June-2023.
2. Patchinand S. Hinduja JW. Bullies move Beyond the School yard; a Preliminary Look at Cyberbullying. Youth Violence and Juvenile Justice. 2006;4(2):148–169.
3. https://www.researchgate.net/publication/351131976_Cyberbullying_Detection_on_Social_Ne tworks_Using_Machine_Learning_Approaches.
4. Dinakar, Karthik, Roi Reichart, Henry Lieberman. Modeling the detection of Textual Cyberbullying. The Social Mobile Web. 2011;11(02):11-17.
5. Willard NE. Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Research Press; 2007.
6. https://ieeexplore.ieee.org/document/8980379
7. https://www.ijraset.com/best-journal/cyberbullying-detection-using-natural-language processing.
8. Ch. Shravya, Pravallika and Subhani Shaik," Heart disease prediction using Machine learning Techniques", International Journal of Innovative Technology and Exploring Engineering, Vol. 8, Issue 6, 2019.
9. Shiva Keertan J and Subhani Shaik," Machine Learning Algorithms for Oil Price Prediction", International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-8, 2019.

10. KP Surya Teja, Vigneshwara Reddy and Subhani Shaik," Flight Delay Prediction Using Machine Learning Algorithm XGBoost", Jour of Adv Research in Dynamical & Control Systems, Vol. 11, No. 5, 2019.
11. Dr. R. Vijaya Kumar Reddy, Dr. Shaik Subhani, Dr. G. Rajesh Chandra, Dr. B. Srinivasa Rao," Breast Cancer Prediction using Classification Techniques", International Journal of Emerging Trends in Engineering Research, Vol. 8, No.9,2020.
12. Mr. Sujan Reddy, Ms. Renu Sri and Subhani Shaik," Sentimental Analysis using Logistic Regression", International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.
13. Ms. Mamatha, Srinivasa Datta and Subhani Shaik," Fake Profile Identification using Machine Learning Algorithms", International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.
14. R. Vijaya Kumar Reddy, Subhani Shaik, B. Srinivasa Rao, "Machine learning based outlier detection for medical data" Indonesian Journal of Electrical Engineering and Computer Science, Vol. 24, No. 1, October 2021.