A MAJOR PROJECT II REPORT

ON

Offensive Language Detection from Social Media Text Using Machine Learning Classification Methods

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted By:

Ms. TANAYA PATRA
Roll No-2K22/CSE/25

Under the Supervision of

Dr. MINNI JAIN
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY,

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-110042

MAY, 2024

ACKNOWLEDGEMENT

I would like to express my gratitude to **Dr. Minni Jain**, Assistant Professor at the Department of Computer Science and Engineering, Delhi Technological University, from the bottom of my heart for all of her support and assistance during this project. Her extensive knowledge, drive, skill, and perceptive criticism have been invaluable in helping me with every step of creating this study strategy.

I am also grateful to **Prof. Dr. Vinod Kumar**, Head of the Department, for his insightful remarks, advice, and careful assessment of my study. His knowledge and intellectual leadership have greatly improved the quality of this thesis.

My heartfelt thanks go out to the esteemed faculty members of the Department of Computer Science & Engineering at Delhi Technological University. I want to express my sincere appreciation to my friends and coworkers for their consistent encouragement and support throughout this difficult road. Their friendly banter, insightful criticism, and intellectual interactions have enhanced and genuinely fulfilled my research experience.

Even while it is impossible to thank each and every person who has helped along the way, I still want to recognize their efforts and accomplishments as a group. Their unwavering love, support, and encouragement have been crucial to finishing my MTech thesis.

Tanaya Patra
(2K22/CSE/25)

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE DECLARATION

I **TANAYA PATRA** hereby certify that the work which is being presented in the thesis entitled **Offensive Language Detection from Social Media Text Using Machine Learning Classification Methods** in partial fulfillment of the requirements for the award of the Degree of Master of Technology submitted in the Department of Computer Science and Engineering, Delhi Technological University in an authentic record of my own work carried out during the period from August, 2022 to May, 2024 under the supervision of Dr. Minni Jain.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Tanaya Patra

This is to certify that the student has incorporated all the corrections suggested by the examiner in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s) Signature of External Examiner

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that Tanaya Patra (Roll no 2K22/CSE/25) has carried out their research work presented in this thesis entitled "**Offensive Language Detection from Social Media Text Using Machine Learning Classification Methods"** for the award of **Master of Technology** from Department of Computer Science and Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Dr. Minni Jain
(Supervisor)
Assistant Professor,
Department of CSE,
DTU-Delhi, India

ABSTRACT

The utilization of social networking platforms has significantly surged in recent years, leading to a substantial rise in user-generated content across the web. This information predominantly appears in unorganized and somewhat organized forms. Numerous social media platforms face the issue of hate speech, which takes on different forms including aggressive language and the development of visual content like memes.

This research focuses on employing Twitter data to identify offensive speech online. Nowadays, techniques for ML which is machine learning and NLP which is natural language processing) have been increasingly utilized for detecting hateful content on the internet. This study specifically addresses the issue of offensive speech detection in textual data by applying machine learning techniques. Prior to utilizing the dataset with machine learning models, feature selection was conducted. Various machine learning algorithms were applied to an openly accessible Twitter dataset.

Offensive speech can be defined as, use of such text or words which are aggressive, violent, or abusive in nature and directed towards a certain group or individual who shares a gender, ethnicity, set of beliefs, or place of residence. The suggested model can automatically identify hateful content on Twitter. This method relies on the TF-IDF where TF is known as term frequency and IDF is known as inverse document frequency methodology and a bag of words. Machine learning classifiers are trained using these features. Thorough tests are carried out on the available Twitter dataset, and by comparing 5 different models based on their performance we can conclude that Random Forest Classifier algorithms works best with highest accuracy of 95.22%.

LIST OF PUBLICATIONS

1. Tanaya Patra "Offensive Language Detection from Social Media text using Machine Learning Classification Methods: A Review", Accepted at "**International Conference on Artificial Intelligence, Machine Learning and Big Data Engineering (ICAIMLBDE) organized by ISETE**" on 22nd April 2024 at Kolkata, India.

   Paper Id: IST-BDE-KLKT-220424-5607

   Indexed by Scopus.

2. Tanaya Patra "Offensive Speech Identification from Social Media Text by Applying Machine Learning Classification Algorithm: Logistic Regression", Accepted at "**International Conference on Intelligent Computing and Communication Techniques at JNU New Delhi, India**" on 28th & 29th June 2024 at New Delhi, India.

   Paper Id: 916

   Indexed by Scopus.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ML	Machine Learning
NLP	Natural Language Processing
UN	United Nations
LR	Logistic Regression
ANN	Artificial Neural Network
SVM	Support Vector Machine
CNN	Convolutional Neural Networks
LSTM	Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
PCA	Principal Component Analysis
TF-IDF	Term Frequency – Inverse Document Frequency
AUC-ROC	Area Under the ROC Curve
UNR-IDD	University of Nevada – Reno Intrusion Detection Dataset

CHAPTER 1

INTRODUCTION

1.1 A Brief Overview

In the current time of expanding online connectivity, propelled by various social media platforms like Instagram, Facebook, Twitter, among others, striving to unite individuals worldwide through smartphones, computers, and the internet has become achievable. The increasing number of internet users signifies the substantial growth in this domain. The increase in using the internet and social media has made life easier, it has also caused new problems. One of the big problems is dealing with mean or hurtful stuff that gets posted online. Because there are so many people using these websites and a lot of things are getting posted all the time, it's really hard to control or stop all the bad stuff that's being shared.

To help with this problem, in the last ten years, people have started using special computer programs that are really good at learning and understanding language. These programs help find and stop harmful things that people post online, even though there's so much information being shared every second. The Plan of Action on Offensive Speech by the UN (United Nations) characterizes offensive language as, any type of communication through text or speech, written expression, or conduct that target or employs offensive language towards a person or a group based on their identity traits, such as religion of a person, ethnicity, nationality of any individual, race, colour, lineage, gender, or other factors related to identity. Considering the complexity of defining hate speech, international law doesn't primarily aim to ban hate speech but rather focuses on prohibiting actions that incite discrimination, hostility, and violence. Incitement holds peril as it has the potential to prompt actual 'action,' turning hate speech into hate crimes. Hate speech that doesn't escalate to incitement isn't mandatory for states to address. However, this doesn't diminish the harm caused by any form of hate speech, as its unrestricted presence can lead to societal division.

The term 'hate' implies an intense dislike towards a person, community, entity, or idea.

Hate is a very strong feeling of disliking someone or something a lot. It's one of many emotions people can feel, like happiness, sadness, anger, and love. Nowadays, using the help of web and many social media platforms, all individuals talk to each other mainly by writing messages and sharing pictures or videos. But online, it's hard to understand how someone feels because we can't see their face or hear their voice. This makes it tough to know if someone is feeling hate towards something.

On social media, hate often comes through mean or aggressive words that people use. These words can make others feel bad or hurt their feelings. So, even though we can't see or hear each other online, sometimes the way people write or speak can show strong dislike or hate towards something or someone.

Expressing freedom of speech sometimes results in a lot of hurtful talk on the internet. Both hate speech and online bullying are similar and fit into one category. Hate speech involves saying or doing things that harm an individual person or any group because of person's race, culture, religious belief, sexuality, or other parts of who they are. This kind of speech makes people feel less confident, more anxious, and very upset, affecting their mental health a lot. When hate speech is aimed at certain groups, it causes big problems in society and can even lead to terrible things like mass violence against those groups.

Fig 1.1: 4 Forms of Hate Speech

Identifying abusive, explicit, and hateful content on media platforms that includes

Facebook, Instagram, and Twitter has become an overwhelming task due to rapid proliferation of user generated information on these networks. The way hate is shown mostly relies on words that might not have clear meaning if we don't understand the situation or context.

This research mainly target on the identifying offensive words or text through machine learning and NLP methods. It is specifically designed to identify offensive words on many social networking platforms (Facebook, Instagram, and Twitter etc.). Main target is to evaluating the performance of several classification methods on social media datasets for identifying the best method for actual offensive speech identification. As a result, various pre-selected techniques for resampling, feature development, and combined classification algorithms are used with a dataset taken from social media.

1.2 Objective

This research focuses on employing Twitter data to identify hate speech online. In current years, ML and natural language processing methods have been increasingly utilized for detecting hateful content on the internet. This study specifically focuses on the issue of offensive speech detection in textual data with the use of machine learning techniques. Prior to utilizing the dataset with machine learning(ML) models, feature selection was conducted. Various machine learning algorithms were applied to an openly accessible Twitter dataset.

1.3 Problem Statement

The target of this is to categorize textual content as offensive or non-offensive. The proposed approaches categorized data as offensive speech by using various methods of feature engineering and ML algorithms. Main focus of this work is to design an automated ML based method for identifying objectionable language and hate speech. Automated detection is synonymous with automated learning includes supervised machine-learning technique. To identify offensive language, we employ supervised learning techniques. Organize texts into three groups according to their sentiment and other observable characteristics.

It is technically challenging to recognize hate speech automatically in social media since the way that users express their hate is through words, and words by themselves have no meaning until we attempt to understand them in the context of anything else. Addressing these problems will result in practical and effective solution. The solutions achieved thus far still face certain challenges and without the appropriate context the system's word detection is insufficient.

CHAPTER 2
REVIEW OF LITERATURE

2.1 A Brief Overview

A subfield of computer science known as machine learning where computers can learn without explicit training. As the name implies, the computer does human-like data processing; but, in contrast to traditional method of programming, the machine may or may not be given access to the results for analysis. It is the term for a group of algorithms that, given a set of data, provide intelligent predictions. There are millions of unique data points in these enormous data sets. Recently, ML has proven to be able to extract information and understand semantics at a level comparable to that of a person, also can recognize general patterns more prominently than human experts. Because of the web sectors' demands, vast data quantities, exponential rise in processing power, and advancement in algorithm design have made machine learning (ML) a powerful tool in the current day. There are several different machine learning techniques, or models, in use nowadays [13]. The model to be chosen for a particular problem on the basis of various characteristics of the data as well as the kind of expected result.

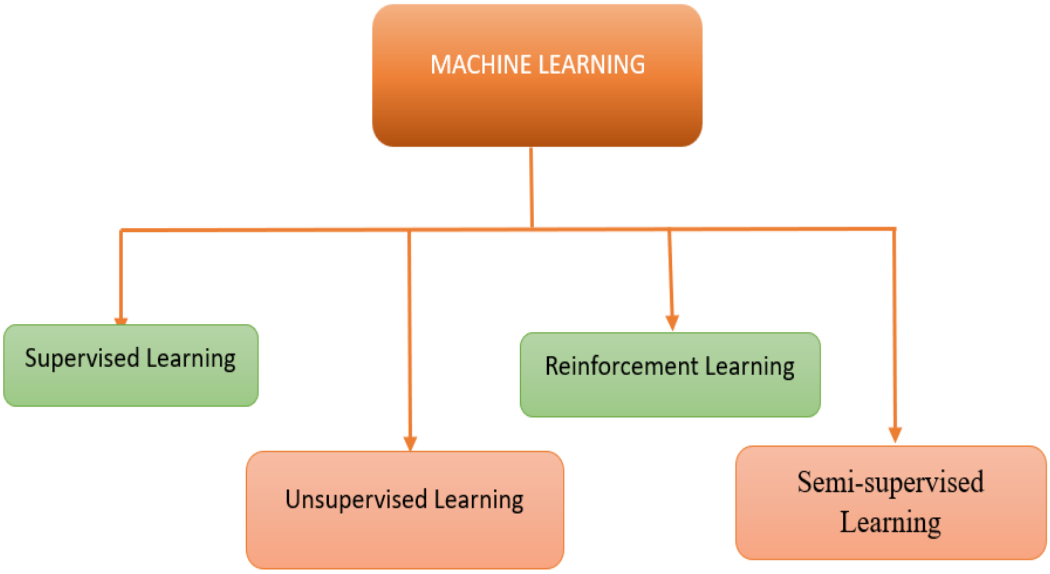


Fig 2.1: Categories of ML

Four categories of ML algorithms exist:

2.1.1 Supervised Learning: This is where the machine learns from labelled data, meaning the data is already indexed with the defined value. The mapping function from the given input parameters to the output variable is then learned by the algorithm using this labelled data. Logistic regression (LR) is one of the common methods in supervised learning.

2.1.2 Unsupervised Learning: The learning type, where the algorithm will be provided with a dataset without any specific instructions on what to do with it. It analyses patterns and structures from the dataset given without any specific feedback. Tasks in unsupervised learning approach include clustering and reduction of dimensionality. Popular techniques in this category are PCA (principal component analysis) and clustering method known as K-means.

2.1.3 Reinforcement Learning: In this learning method, an agent gain's ability to make decisions by operating in a way that maximizes a concept of cumulative reward in its environment. Instead of learning from labelled data, the agent trains from outcomes of its activities.

2.1.4 Semi supervised Learning: The method is defined as the hybrid combination of supervised methods and unsupervised methods. It makes extensive use of data while which are unlabelled using a limited amount of labelled data. The algorithm attempts to leverage both in order to maximize the accuracy of learning.

These are just some fundamental concepts in machine learning theory, and there's much more depth to explore within each topic. In order to categorize tweets as either hateful or not, we take into consideration a logistic regression classifier. Using an array of words and TF-IDF features taken out of the pre-processed dataset, and the system model training is done by using training dataset. The model basically classifies the test dataset into different categories after it has been trained on the training dataset. The logistic regression (LR) method is known as a binary (0,1) classification machine learning technique that determine the value of a dependent variable or outcome variable. Logistic regression uses a function called sigmoid function to place a dependent variable's value inside the interval [0,1].

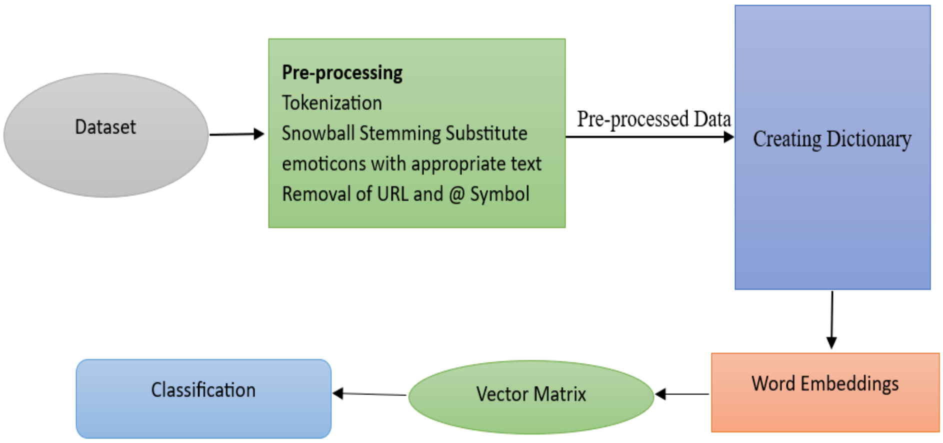


Fig 2.2: Machine Learning steps applied

2.2 MACHINE LEARNING STEPS:

The pre-processing procedures were executed in the following sequence:

2.2.1 Word Tokenization: Tokens, defined as the smallest individual elements or fundamental components within a sentence, were derived. These tokens serve as basic units within a paragraph. Word Tokenization facilitated the conversion of our text into distinct words from list.

2.2.2 Stop words filtering: A list of stopwords in the English lexicon was produced and then removed using nltk.corpus.stopwords.words('english'). Stop words in the interpretative data, including "the," "a," "an," and "in," are deemed non-significant and have no effect on its interpretation.

2.2.3 Punctuation removal: The process involved preserving only non-punctuation characters, identified through string punctuation.

2.2.4 Stemming: To obtain stemmed tokens, nltk.stem.porter.PorterStemmer was utilized. Stemming is a linguistic normalization procedure that reduces words to their base form. For example, the common root word "extract" was used to replace words like "extraction," "extracted," and "extracting."

2.2.5 Elimination of digits: Numeric content was filtered out as it holds no relevance to cyberbullying analysis.

2.2.6 Feature extraction: TF-IDF Transformation using Python's sklearn module was used to convert the data so that machine learning algorithms could use it. TF-IDF mainly reduces the importance of any words that occur often in several texts, making them insufficient for document classification when compared to a word frequency count (as performed by CountVectorizer). Each document as row, each word as column, and their corresponding relevance as weight, determined by tf * idf where values within the matrix are contained in the resulting matrix.

Pre-processing, feature-generation, over-sampling, and classification are the four main phases that are covered. Pre-processing involves cleaning raw data to remove extraneous components and provide better characteristics for next stages. The refined data taken from the previous stage is then transformed into several feature vectors during the feature creation phase. An oversampling strategy was used to address the dataset's extremely unbalanced character in this study. Due to the small amount of data in the hate speech category which, if improperly managed, could result in inaccurate results in the model this strategy involves augmenting the minority classes. For more balanced training, the minority class's up sampling was done in an effort to equalize the volume of data with that of the majority class.

2.3 TYPES OF MACHINE LEARNING MODEL

2.3.1 LR(Logistic Regression)

Linear Regression: Linear regression is a technique which is used to depict the linear relationship between two variables while in which one variable is dependent and the other is independent. In simple terms, the aim of linear regression is to identify and find the right line that seems to describe the interaction _between the variables and this line could be used for further analysis or for making predictions. The procedure entails employing the least squares technique particularly for finding the linear regression line.

This approach produces the best fit because it strives to reduce the value added to the differences between the square of the actual values and the _predicted values. The linear regression model is used significantly for forecasting trends, establishing relations between variables, and for predicting in various fields such as finance, economics, social sciences, and engineering and so on. This one is relatively simple, yet effective, method which provides valuable information about the basic properties of the data. Some distinguished application of linear regression include forecasting of the share prices, sales estimates, climatic conditions, and the impact of a marketing promotion campaign on the sales of a product.

2.3.2 Random Forest

Linear regression is a technique which is used to depict the linear relationship between two variables while in which one variable is dependent and the other is independent. In simple terms, the aim of linear regression is to identify and find the right line that seems to describe the interaction _between the variables and this line could be used for further analysis or for making predictions. The procedure entails employing the least squares technique particularly for finding the linear regression line.

This approach produces the best fit because it strives to reduce the value added to the differences between the square of the actual values and the _predicted values. The linear regression model is used significantly for forecasting trends, establishing relations between variables, and for predicting in various fields such as finance, economics, social sciences, and engineering and so on. This one is relatively simple, yet effective, method which provides valuable information about the basic properties of the data. Some distinguished application of linear regression include forecasting of the share prices, sales estimates, climatic conditions, and the impact of a marketing promotion campaign on the sales of a product.

2.3.3 Decision Tree

Decision trees are one of the primary approaches used in machine learning that deals with both regression and classification problems. Indeed, it is a kind of Supervised Learning technique in which the data partition is done in a tree like structure. Until a decision in made at the terminal or _leaf node decisions are made at each and every node depending on an aspect or attribute value. This means that the decision tree model works in a manner that builds derived models that split the dataset into more specifically defined subsets according to the attributes that yield the largest amount of information gain. This is done until some termination criteria are satisfied such as the achievement of a specified minimum sample size or a certain depth of the tree.

One of the most significant advantages of decision trees is their ability to interpret the model and the relation between them and their predictions. Decision tree is capable to handle both classified and numerical data also it can differentiate between the two When it comes to missing data, decision trees will also be suitable. Nonetheless, decision trees are faster in execution and frequently overfit the training dataset which results in unsatisfactory performance on _new data.

2.3.4 Support Vector Machine

SVM, or support vector machine, is a potent supervised learning method that is applied to regression and classification problems. Although it is also used for tasks related to regression, SVM is mainly employed for classification difficulties. SVM determines the optimal boundary (hyperplane) for classifying data that has the greatest margin of separation between classes. The distance calculated between the closest data points from various classes and the hyperplane is called the margin, and SVM seeks to maximize this distance. SVM looks for a strong decision boundary that adapts well to new data by optimizing the margin.

2.3.5 ANN (Artificial Neural Network)

A computational model known as ANN is based on the neural architecture of human brain. An ANN consist of nodes arranged into several layers with the ability to connect between them in some manners. These layers deliver info, and each node in the network computes. During training, the connections between nodes are changed because data is fed into a neural network to minimize the prediction error. Since ANNs aim at detecting intricate patterns and relationships in data, these machine learning models find application in image and audio recognition, natural language processing and others. ANNs that include feed forward, recurrent, and convolutional ANN options have emerged as cutting-edge solutions in a multitude of applications in fields of machine learning for very complex problems.

2.3.6 Naïve Bayes Classifier

This Classifier[6] is a probabilistic machine learning algorithm suitable for classification problems. That all traits are uncorrelated, and the origins of naïve in the Bayes' theorem are what make anything naïve. For large datasets, this assumption make and enhance the processing when computing probabilities. The Naive Bayes Classifier model predicts the probability of any given class occurring given the input features. These probabilities are then multiplied to arrive at the posterior probability of the given input belonging to any of the class. Another strength of the Naive Bayes Classifier is that it is capable of handling data with high dimensionality which makes it ideal for text categorization and spam detection. It works fine when it is used with a less number of dataset for training. Nevertheless, the Naive Bayes Classifier has an assumption of independent features, which might lead to suboptimal performance in some situations. It also depends heavily on the quality of the input data and it can be very sensitive to outliers.

2.3.7 Gradient Boosting Classifier

With the help of the well-known machine learning algorithm gradient boosting classifier, numerous weak learners can be combined to produce a powerful prediction model. It functions by repeatedly adding new decision trees to the system model, each of which tries to fix the flaws of the preceding trees. Based on the gradient descent optimisation process, which updates the performance metrics of the system model in the direction of the loss function's negative gradient, is the gradient boosting classifier model. This approach allows the system model to learn from its mistakes and gradually improve the accuracy. One of the main benefits of gradient boosting classifiers is its functionality to handle complex behaviour, high dimensional data sets with a large no. of features. They also tend to perform well on imbalanced datasets, where the number of samples in each class varies widely. However, gradient boosting classifiers can be sensitive to overfitting if it is not properly regularized, and they can be expensive to train computation model, especially when dealing with large no. of datasets.

2.4 PERFORMACE METRICS

To compare different models used in this project we have used the following performance metrics to evaluate models:

2.4.1 Accuracy: This is commonly used performance metric in machine learning and data analytics that calculates the proportion of correct predictions made by a model. It is a simple, straight forward measure that is easy to interpret, making it a popular choice for evaluating the effectiveness of different models. We divide the number of correctly predicted cases by the total number of instances in the dataset to determine accuracy. For example, if a system model correctly predicts 92 out of 100 instances, its accuracy is 92%.

$$Accuracy = \frac{CorrectOutputs}{TotalDataItems} \tag{1.1}$$

While accuracy is a useful measure of overall performance, it can be misleading in certain situations. For example, if the data is imbalanced with one class vastly out-numbering the others, a model that simply predicts the major part of the class for every instance will have high accuracy even though it is not actually making useful predictions. In addition, accuracy does not take into account the cost or benefit of different types of errors. In some cases, false negatives which is predicting a negative output when it's actual output is positive may be more costly than false positives which is predicting a positive output when it's actual output negative, or vice versa. Despite these limitations, accuracy remains a valuable performance metric in many applications, particularly when the dataset is balanced and the cost of different types of errors is roughly equal.

2.4.2 Precision: In machine learning and data science precision plays a role, as a performance metric. It measures the accuracy of classifications made by a model. Precision is calculated as the ratio of outcomes (TP) to the sum of true positive results and false positive results (FP).

$$Precision = \frac{TP}{TP + FP} \tag{1.2}$$

In other words, simply precision indicates the accuracy of predictions. It is crucial, in scenarios where false positives can lead to consequences or harm. Consider the case of medical diagnosis: if the model identifies healthy patients as having the disease, these patients will undergo treatment unnecessarily, which may be detrimental to their health. In such a scenario, having high precision is more important than having high recall, where recall is the proportion of actual positives that were identified by the model. Yet, it is important to stress that precision should not be assessed alone: it should be assessed jointly with other model performance metrics such as recall, accuracy, F1 score, and so on. A high precision model can have low recall, in such a case it will miss a large number of positive instances. Precision is an essential performance metric to determine how good the model is at identifying positive instances. On its own it does not give a complete picture. Be sure to consider other performance metrics as well.

2.4.3 Recall: Recall is a metric in machine learning for evaluating a classification model. It measures a model's ability to label every positive instance in a dataset correctly. In simpler terms, recall finds the ratio of true positives (TP) over actual positive examples, that is TP + false negatives. Mathematically, recall can be expressed as:

$$Recall = \frac{TP}{TP + FN} \tag{1.3}$$

A high recall value means that the model is good at selecting most of the positive cases in the dataset. A low recall value 18 may also mean the model is not good for the positive instances. Let's consider a marketing case. Companies prefer to be informed about all customers who are interested in their products or services, even if some are not interested. High recall rates help companies understand all customers and let them compare their products and services, even if they are not interested in some customers. If the recall of the previous model is 99%, do you consider it as a good model? Do we have to think about these values? If we look at the model problem from this perspective, the answer to the question "which is the best PPC model" will be redefined.

2.4.4 F-Score: F-score is a performance measure in machine learning to quantify the effectiveness of a classification model. It weighs precision and recall against each other by aggregating them into a weighted harmonic mean. The F-score is calculated as follows:

$$F - Score = \frac{2 * (Precision * Recall)}{Precision + Recall} \qquad (1.4)$$

where recall is the fraction of actual positive behaviour that you've correctly identified and precision is the fraction of predicted positive behaviour that actually occurred (TP / (TP + FP)). The F-score is between 0 and 1, with 1 being perfect recall and precision. In a real-world context, a high Fscore tells you that you have managed to balance recall and precision: while maximizing true positive predictions, you are simultaneously minimizing both false positives and false negatives. When a model is dealing with imbalanced data, which is when there is disproportionate frequency in models between classes (for example, 95 percent of examples are of class 0, and the other 5 percent are of class 1), the measure helps. In an imbalanced data set, an algorithm might predict a single class and attain high accuracy. The model could only be predicting the majority class, though, leading to a high number of false negatives. This is why accuracy is not a good measure for such scenarios, while the F-score is, since it correctly measures the performance of a model on a positive class against a negative class. The F-score is an important metric to consider when evaluating the performance of classification models, especially when you are dealing with an imbalanced data set.

2.5 TOOL USED

2.5.1 Python: Guido van Rossum designed python a well-liked top-level programs language in 1991. In a variety of markets consisting of internet growth, clinical computer data evaluation, expert system plus automation it has actually ending up being commonly made use of. It is renowned for having basic understandable phrase structure that is simple to find out. The Python neighbourhood has actually made a considerable payment to the advancement of many collections, components together with plans that allow designers to rapidly as well as effectively produce reputable services. These consist of Flask along with Django for internet growth, TensorFlow coupled with pyTorch for artificial intelligence, Matplotlib for information visualization along with numpy for mathematical calculations, as well as pandas for information control. Numerous shows versions, consisting of step-by-step, object-oriented, as well as useful shows, are sustained by the versatile language Python. Its code is succinct as well as can attain intricate jobs with marginal lines of code. Its solid neighbourhood assistance makes it a superb selection for both newbies as well as skilled programmers wanting to create durable applications for varied use situations, varying from straightforward manuscripts to complicated internet applications as well as artificial intelligence versions.

2.5.2 Jupyter Notebook: Utilizing the open-source internet application Jupyter Note pad, customers might develop as well as share records with formulas, live code, message and also visualizations. Countless programs languages, such as Python, R, Julia as well as Scala, work with it. In information scientific research as well as artificial intelligence, Jupyter note pads are rapidly coming to be increasingly more prominent because of their ability to boost testing, team effort as well as result repeatability. Due to its interactive particular, it's a fantastic training device for interactive tutorials or shows. For trainees, academics or designers alike that require to share ideas making use of code Jupyter Note pad is an efficient as well as flexible device.

2.5.3 Scikit-learn: An open-source artificial intelligence collection called Scikit-learn has actually come to be preferred amongst Python developers together with information researchers. It supplies a series of finding out formulas both managed and also without supervision such as dimensionality decrease, gathering, regression as well as category. Individuals can additionally pick as well as review designs, preprocess information together with perform assessments with scikit-learn. As a result of its detailed documents, instinctive user interface and also exceptional efficiency Scikit-learn is currently utilized by both amateur coupled with experienced artificial intelligence experts. Scikit-learn is made use of by scientists firms plus programmers for a range of applications consisting of financing, health care, picture as well as message evaluation and also a lot more. Due to the fact that scikit-learn is reputable and also effective it is still among one of the most prominent collections for artificial intelligence remedies.

2.5.4 NumPy: NumPy is a prominent Python collection utilized for clinical computer particularly in information evaluation, artificial intelligence and also clinical research study. It provides a selection item plus features for dealing with selections consisting of mathematical procedures that can be done on whole selections of information promptly as well as effectively. Among NumPy's crucial functions is its capability to manage direct algebra, Fourier changes as well as arbitrary number generation, making it optimal for mathematical computation. Contrasted to typical Python listings, NumPy ranges are a lot more effective because of their consistent information kinds as well as memory use.

CHAPTER 3

LITERATURE SURVEY

3.1 Overview

Recently, sentiment analysis and polarity identification in Twitter data have attracted a lot of study attention. Researchers primarily concentrate on discerning users' sentiments concerning products, services, or events. To achieve the intended results, they mostly collect information from social media platforms and evaluate it. Additional research endeavours, prompted by the vast volume of social network data, aim to categorize users with akin interests in a product or service and to identify potentially harmful content circulating within social network data.

The information gathered from social networks is largely unstructured, posing challenges for existing methodologies to effectively analyse such data. Researchers have conducted numerous studies that have significantly improved these analytical approaches. Instances have frequently been reported where conflicts or clashes between communities or groups have erupted, with social media posts often identified as a primary catalyst. These occurrences have spurred researchers to actively seek ways to identify and address violent content on social networking platforms.

Scholars in the domain that includes natural language which mainly focusing on developing different models that can identify hate speech in statements. Compared to their intuitively superior machine learning and soft computing counterparts, these types of models are computationally slower. As supervised learning models are required to do these tasks, one requirement placed on the machine learning models has been the necessity for labelled datasets. Scholars have consistently invested their time and energy in this field, with numerous researchers creating original datasets. Such studies typically take the form of data extraction from social networking sites followed by message classification according to the existence of hateful material.

For example, Ahmed et al. [7] carefully assigned binary labels to a sample that included both English and Bengali texts. Similar to this, Sahi et al. designed a model of supervised learning using a Turkish language dataset that was specifically intended to identify hate speech targeting women. Then gathered the information for this study that included women's fashion decisions, and then used that dataset to create a machine learning model. An important part of creating the data set is the work of annotators, and Waseem investigated how annotators affected various categorization models. In a different study, Waseem et al. constructed a dataset with 16,000 tweets and looked into identifying the key elements that can enhance model performance. The use of freely available data sets to develop models using machine learning for the aim of detecting hate speech has been the subject of extensive research.

3.1.1 “Twitter as a corpus for sentiment analysis and opinion mining[8].”:

Suggested a method that ranks the N-gram features based on their TFIDF values once they are extracted from the tweets. The machine learning algorithm uses these features to classify tweets into categories like hateful, non-hateful or clean etc. Tweets are first translated into lower case before any extraneous information, such as stop words, URLs, space patterns, and Twitter mentions, is eliminated. Following data processing, a grid search is used to train some machine learning classification algorithms like Logistic Regression (LR), SVM and Naïve Bayes. These classifiers are used to identify all possible combinations of the feature parameter. Test data were subjected to cross-validation, and the output indicate that the algorithm logistic regression works best with an accuracy rate of 95.6%.

3.1.2 “A Lexicon Based Approach for Hate Speech Detection[2].”:

A method that can gather Twitter corpus of objective, negative, and positive attitudes automatically was shown. After extracting n-grams and PoS tags from these corpora, stop words are eliminated by tokenization. Two ways have been presented for extracting the features of n-grams. The first strategy computes the probability distribution's entropy for every instance of data. High entropy n-grams were thrown away, while low entropy n-grams were retained. We refer to the second tactic as salience. High salience negatives were preserved. Salience outperforms bigrams in terms of model building accuracy between these two methodologies. After that, tweets are categorized into one of three groups using a multinomial naïve Bayes classifier.

3.1.3 "Hate speech detection on twitter using multi-nomial logistic regression classification method[3].":

An approach where logistic regression method is applied to identify hate speech on Twitter was presented by P. Sari et al. in 2019. They used Tokenizing, Case Folding, Stemming, and Filtering methods in the initial processing phase after gathering Twitter data. After pre-processing, vectorization is carried out by applying TF-IDF approach. Following feature engineering, 84% accuracy was found using the Logistic Regression technique.

3.1.4 "Deep learning for hate speech detection: a comparative study[4].":

Through the use of the three most commonly used datasets, Malik et al. did a statistical analysis of hate-speech identification algorithms. They contrasted the effectiveness of various deep learning models with conventional classifiers, evaluating the performance of CNN (Convolutional Neural Networks), LSTM (Long Short-Term Memory), and Bidirectional Encoder Representations from Transformers (BERT). In addition, they evaluated these models' practical performance in terms of detection accuracy, computational economy, domain generalization, and ability to use pretrained models.

3.1.5 "Application of machine learning techniques for hate speech detection in mobile applications[5].":

Raufi and Xhaferri[5] used machine learning(ML) methods for detecting offensive text in mobile apps. They examined the performances obtained for various classifiers, such as Decision Trees, Random Forests, and Naive Bayes, using a dataset of user comments.

3.1.6 “Detecting Offensive Language in Social Media to Protect Adolescent Online Safety[6].”:

With these methods, one can identify fake news in along with tweets. In addition to hate speech and false information, we can use machine learning to identify clickbait. Having an attention-grabbing headline that makes the reader want to click on the link is known as "clickbait." Character-level embedded words can be employed as a feature engineering strategy for this distributed word embedding. Recurrent neural networks are used in the implementation. One way to filter offensive content on social media is to employ a "Moderation Blacklist." A tweet is automatically labelled as spam and restricted when someone enters terms that are on the blacklist.

3.1.7 “Detecting nastiness in social media[8].”:

Sambhagadi et al. aim to use natural language processing (NLP) techniques to identify negative content on social media in order to finally identify and classify cyberbullying. Because NLP strategies are implemented in this way, they can even differentiate between neutral and harmful uses of language in data. Crowdsourcing and in-lab annotations are utilized to iteratively modify the annotations used in the paper. Data set was gathered from various posts written in English-language on online sites, including ask.fm and other somewhat anonymous platforms. Using NLP in conjunction with a ranked list of expletives made crawling more efficient.

3.1.8 “Cyberbullying ends here: Towards robust detection of cyberbullying in social media[9].”:

Parallel testing of hypothesis formulation is utilized by Yao et al. to significantly minimize the number of features needed for classification while retaining excellent accuracy. The main goals of this strategy are scalability, timeliness, and high accuracy. Models are developed using semi-supervised machine learning techniques on an Instagram dataset that was gathered using snowball sampling and partially manually categorized by a team of specialists. The usage of one set of data that was exclusive to Instagram, the inability to verify the accuracy of the classifications, and the amount of overhead brought on by the challenge of gathering comment-based labels were the approach's limitations.

3.1.9 "A lexicon-based approach for hate speech detection[11].":

Njagi Dennis et al. classified offending speech in blog sites and also on-line discussion forums using a maker learning-based. The writers created a master vector of functions utilizing a dictionary-based approach. Semantic as well as subjective attributes with an emphasis on disapproval speech were utilized to develop characteristics based upon view expressions. The resulting attribute vector was after that provided to a rule-driven by the writers. The writers made use of a precision efficiency standard to evaluate their in the speculative situations and also they had the ability to accomplish 73% accuracy.

3.2 Datasets used for analysis

A dataset in artificial intelligence is a collection of information factors or instances that are made use of to establish, examination together with validate artificial intelligence versions. A selection of i/p attributes or variables as well as the going along with outcome tags or target worths comprise a dataset. The modification of artificial intelligence designs call for datasets. They offer the design the information it requires to determine patterns and also links in between the i/p attributes along with the result tags. The efficiency coupled with generalizability of the educated versions are substantially affected by the dataset's high quality together with representativeness. Depending upon the specific problem available along with the sort of artificial intelligence task available, datasets could vary in dimension, ins and out, as well as organisation. They could be unstructured, like message, images or songs or arranged, like table with rows plus columns. Artificial intelligence datasets are generally separated right into 3 subsets:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31962 entries, 0 to 31961
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      31962 non-null  int64
 1   label   31962 non-null  int64
 2   tweet   31962 non-null  object
dtypes: int64(2), object(1)
memory usage: 749.2+ KB
```

Fig 3.1: Dataset Count

1. Training data set: The ML model is trained using this subset. The model uses the many tagged examples in it to figure out the underlying relationships and patterns.
2. Validation data set: This part is utilized to alter the version's hyperparameters plus examine its efficiency throughout training. It aids in stopping overfitting coupled with overviews the choice of the most effective version design along with specification setups.
3. Test data: This data set is utilized to assess the trained model's ultimate performance and generalizability. It contains unseen examples that were not used during training and validation. The test set offers a neutral evaluation of how nicely the model predicts on fresh, new data.
Information Gathering: To arrive at the final results, we used Twitter Dataset for offensive text data Detection on social media, which we collected from Kaggle. Given the gravity of the problem we hope to tackle, selecting a dataset that was broad, reliable, relevant, and brief was essential. Although we also took into account a large number of other datasets, a manual review revealed that a large number of them contained irrelevant data, lacked sufficient quality, or had missing properties. As a result, after experimenting with numerous other publicly available datasets, we settled on [10].

3.3 Data Cleaning

```
 @user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction.   #run

@user @user @happy thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx.    #disapointed #getthank
ed

  happy birthday love

#model   i love u take with u all the time in urð±!!! ðððððð¦ð¦ð¦   #disrespect

 factsguide: society now    #motivation #cunning #despise #hatred

[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo

 @user camping tomorrow @user @user @user @user @user @user @user dannyâ¦

the next school year is the year for exams.ð¯ can't think about that ð #school #exams  #imagine #actorslife #revolutionscho
ol #girl
```

Fig 3.2: Dataset before preprocessing

The data pre-processing procedures were executed in the following sequence:

```
def data_processing(tweet):
    tweet = tweet.lower()
    tweet = re.sub(r"https\S+|www\S+http\S+", '', tweet, flags = re.MULTILINE)
    tweet = re.sub(r'\@w+|\#','', tweet)
    tweet = re.sub(r'[^\w\s]','',tweet)
    tweet = re.sub(r'ð','',tweet)
    tweet_tokens = word_tokenize(tweet)
    filtered_tweets = [w for w in tweet_tokens if not w in stop_words]
    return " ".join(filtered_tweets)
```

Fig 3.3: Tokenizing the data

1) Word Tokenization: Tokens, defined as the smallest individual elements or fundamental components within a sentence, were derived. These tokens serve as basic units within a paragraph. Word Tokenization facilitated the conversion of our text into distinct words from list.

2) Stop words filtering: A list of stopwords in the English lexicon was produced and then removed using nltk.corpus.stopwords.words('english'). Stop words in the interpretative data, including "the," "a," "an," and "in," are deemed non-significant and have no effect on its interpretation.

3) Punctuation removal: The process involves removal of punctuation characters, identified through string library function.

4) Stemming: To obtain stemmed tokens, nltk.stem.porter.PorterStemmer was utilized. Stemming is a linguistic normalization procedure that reduces words to their base form. For example, the common root word "extract" was used to replace words like "extraction," "extracted," and "extracting."

5) Elimination of digits: Numeric content was filtered out as it holds no relevance to cyberbullying analysis.

6) Feature extraction: TF-IDF Transformation using Python's sklearn module was used to convert the data so that machine learning algorithms could use it. TFIDF reduces the importance of words that occur often in several texts, making them insufficient for document classification when compared to a word frequency count (as performed by CountVectorizer). Each document as row, each word in column, and their corresponding relevance which is weight is determined by tf * idf, are contained in the resulting matrix.

```
user father dysfunctional selfish drags kids dysfunction run

user user happy thanks lyft credit cant use cause dont offer wheelchair vans pdx disapointed getthanked

happy birthday love

model love u take u time ur disrespect

factsguide society motivation cunning despise hatred

22 huge fan fare big talking leave chaos pay disputes get allshowandnogo

user camping tomorrow user user user user user user user dannyâ

next school year year exams cant think school exams imagine actorslife revolutionschool girl
```

Fig 3.4: Displaying tweets after processing the data

CHAPTER 4

METHODOLOGY

4.1 Overview

Hate speech currently threatens society greatly, harming people's self-respect, cohesiveness, and the country as a whole. Various derogatory terms are interchangeably used. Figure 2 displays a word cloud explicitly illustrating hate speech derived from. Hence, the elimination and categorization of hate content across social platforms are imperative and demand immediate attention.

Fig 4.1: Word cloud for hate speech

Since this study involves supervised classification, to train on the dataset, it used SVM, ANN, LR, and naïve Bayes classifiers. These methods are commonly utilized in machine learning for supervised learning tasks.

4.1.1 Logistic Regression Method

Machine learning uses the classification technique known as logistic regression to forecast categorical outcomes, particularly in binary classification issues. It collects and preprocesses data and is used when the dependent variable (also known as the target variable) is categorical. Make sure the variables are appropriately encoded for analysis, and that they are all numeric.

The model modifies its coefficients throughout training for reducing the discrepancy between expected and actual results. The best-fitting parameters are found through the application of optimization methods.

Use the trained model to ascertain the probability that an observation will fall into a certain class. By using a threshold, these probabilities are converted into class labels (such class 0 or class 1). Generally speaking, if the expected probability is more than 0.5, it is categorized as belonging to one class; if not, it belongs to the other.

Using a different validation or test dataset, evaluate the model's performance using evaluation measures such as accuracy, precision, recall, F1-score, ROC curve, and AUC-ROC (Area Under the ROC Curve). These indicators aid in determining whether the model needs to be improved or adjusted based on how well it is generating predictions. Analyse the logistic regression model's coefficients to determine how each predictor affects the result. Where negative coefficients suggesting a negative link with the outcome, positive coefficients show a positive relationship with it.

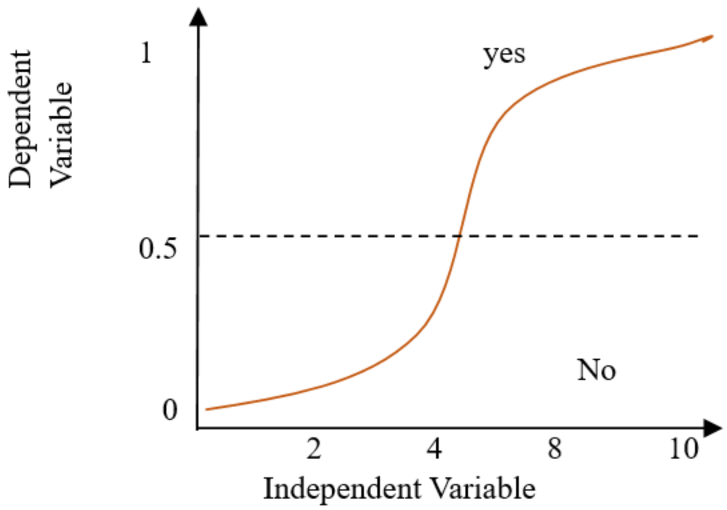


Fig 4.2: Logistic Regression Model

4.1.2 Naïve Bayes Classifier

Based upon the self-reliance of forecasters (functions) thought by Bayes' theorem the Naïve Bayes Classifier is a category formula. It often functions properly in a selection of real-world situations regardless of its simple presumption of function self-reliance, especially in message category plus spam filtering system.

The Bayes theory which establishes the chance of a theory (course tag) provided the observable proof (functions) is the structure of the approach. P(programs onto course) = P(programs|training course) * P(training course)]/ P(programs) is the mathematical expression for it where P(class|functions) is the chance of the course provided the observed attributes. The property of Naïve Bayes is that the presence of a particular within a course is un-associated from the presence of various other functions. By determining the possibility of each course provided the observed features the classifier projections the course tag for an offered circumstance. As the predicted course tag it picks the course with the highest possible possibility.

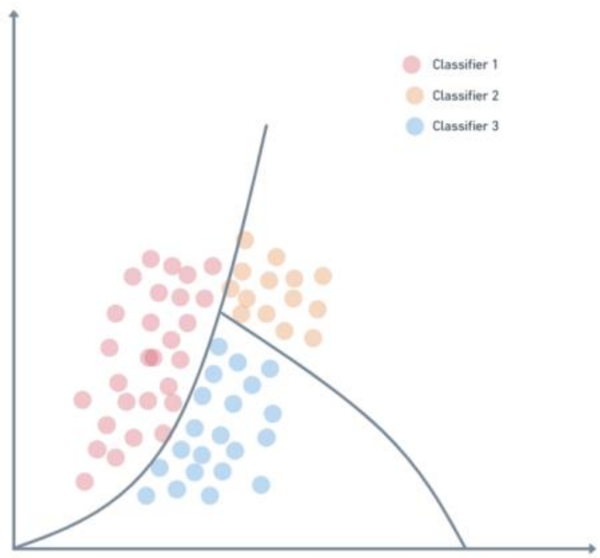


Fig 4.3: Naïve Bayes Classifier Model

4.1.3 SVM (Support Vector Machine)

SVM is a potent supervised learning method which is applied for regression and classification problems. Although it can also be applied for regression related tasks, SVM is mainly employed for classification difficulties. SVM determines the optimal boundary (hyperplane) for classifying data that has the greatest margin of separation between classes. The difference between the closest data points from various classes (referred as support vectors) and the hyperplane is called the margin, and SVM seeks to maximize this distance. SVM looks for a strong decision boundary that adapts well to new data by optimizing the margin.

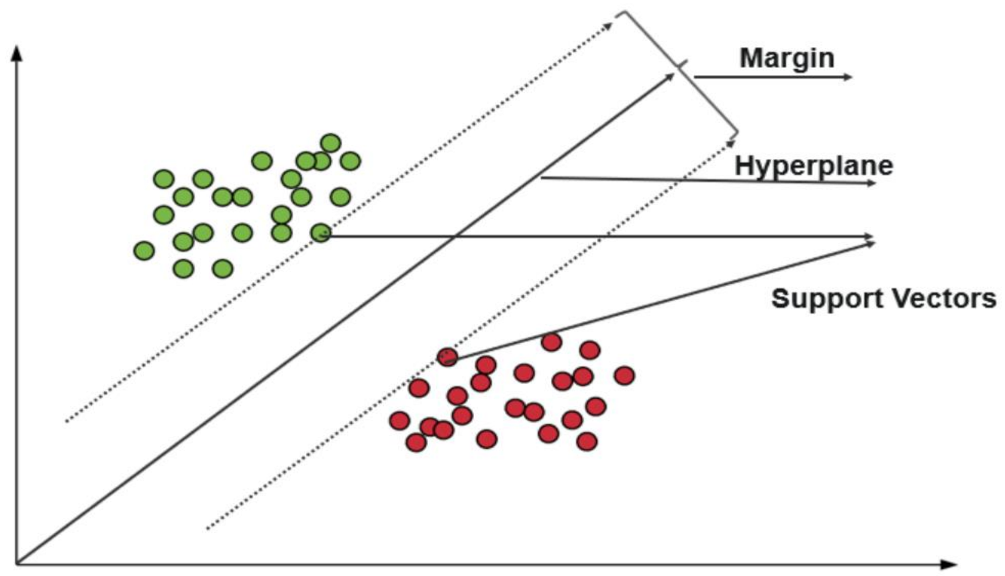


Fig 4.4: Support Vector Machine Model

4.1.4 Artificial Neural Network

An Artificial neuronal Network (ANN) is a computational version that simulates the neural framework of the mind of a human. The elements of an ANN or expert system are connected nodes prepared in layers. Details propagates with these layers, with each node executing estimations. In order to decrease forecast mistakes, node links are changed throughout training as information moves throughout the network. Expert system (ANNs) are durable devices for image and also speech acknowledgment, all-natural language handling as well as several various other applications due to the fact that they are exceptional at finding detailed patterns as well as partnerships in information. ANNs which are available in a range of kinds such as feedforward, reoccurring plus semantic networks have actually totally altered the area of artificial intelligence by offering reliable options for complicated concerns throughout a vast array of applications.

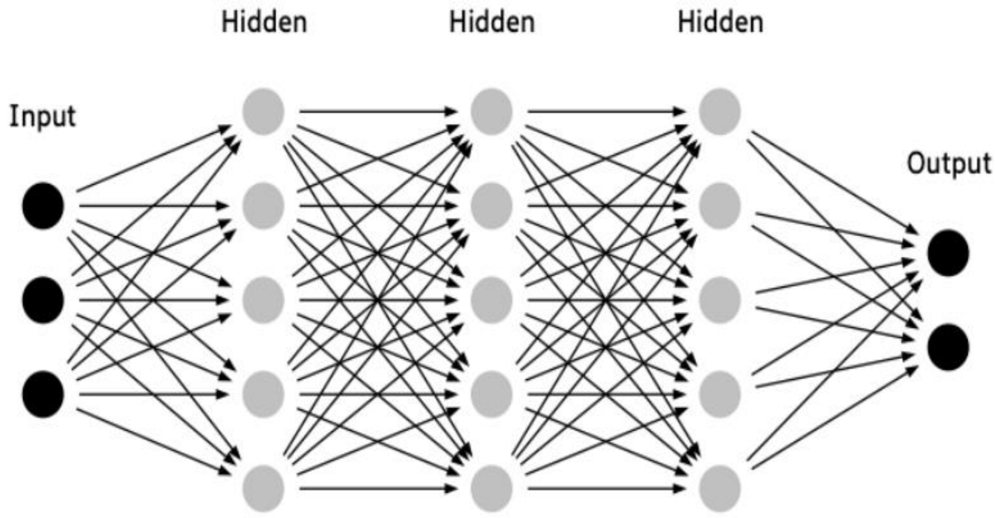


Fig 4.5: Artificial Neural Network Model

4.1.5 Random Forest:

Specifically, one of the most efficient and accurate techniques of ensemble learning – Random Forest – is effective and correct while solving problems in the sphere of regression and classification. That is, in the training phase, it develops a number of decision trees and the outcomes of these trees are then used in forming the predictions. Every decision tree is designed on the basis of the maximum number of features that are available, and in addition, only a random subset of the training sample data is used; that is why overfitting is minimized, and, therefore, the capacity for generalization of the algorithm is enhanced. In the training phase, bootstrap is performed on the training data in order to construct each tree, which belongs to this forest. By using replacements instead of random samples taken from the first training set, this approach creates different training sets for every tree. Every tree in the forest produces a result on its own when it comes to forecasting new data. Next, the average of these individual tree forecasts used for regression or majority vote used for classification is used to get the final prediction. When it comes to managing noisy data, identifying complex data linkages, and preventing overfitting, Random Forest shines.

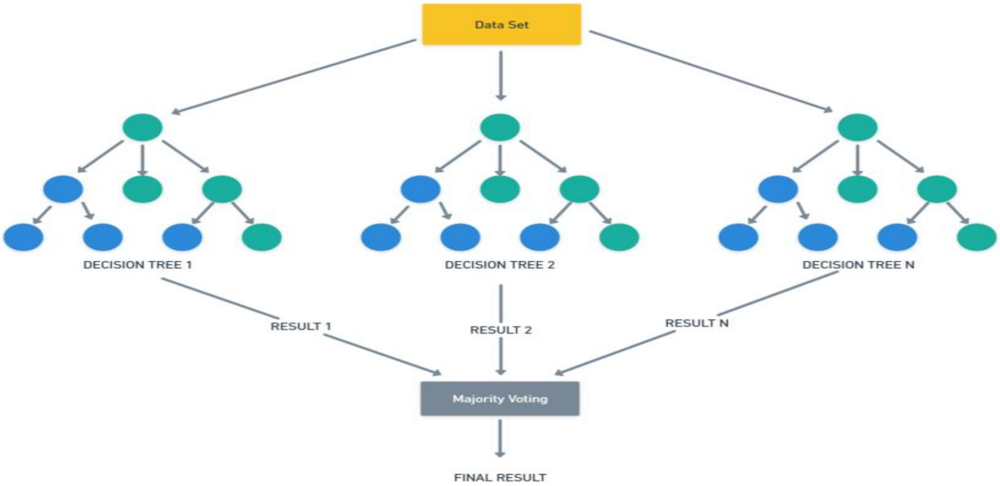


Fig 4.6: Random Forest Classifier

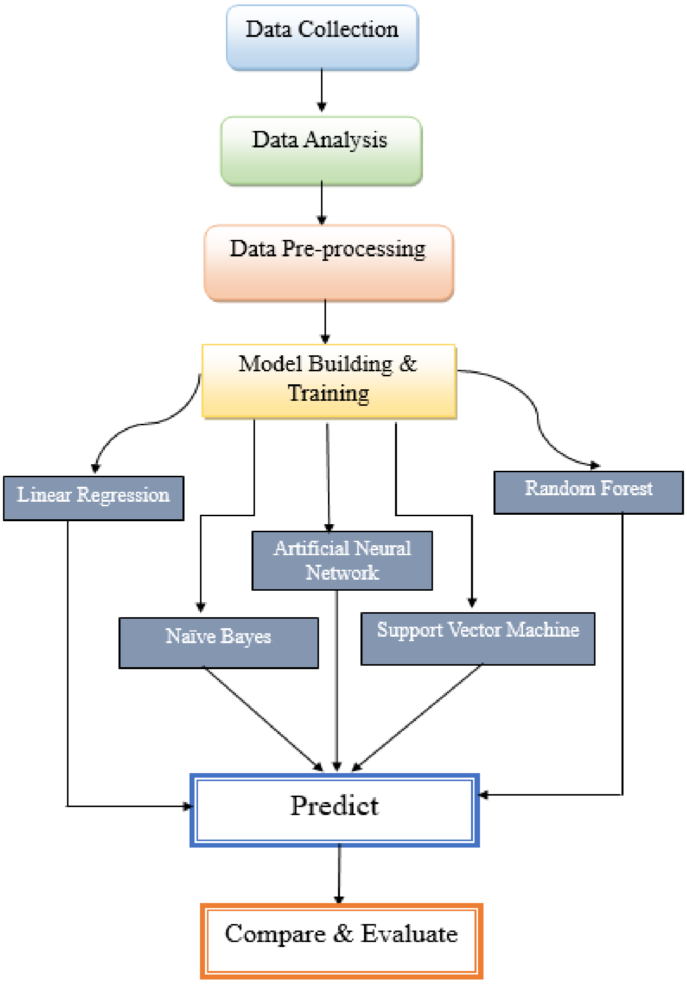


Fig 4.7: Flowchart of the model

4.2 Working of the Model

The following steps are taken in this project:

1. Data Collection: In this particular step, data collection is done in the form of csv file. In this project UNR-IDD dataset is used.

2. Data Analysis: In this particular step, the data is properly analysed on different parameters.

3. Data Pre-Processing: In this step, the data is divided into two categories, i.e., Training set and Testing data set.

4. Model Building and Training: In this step, Machine Learning (ML) models are created and trained on the dataset processed in the above stage.

5. Prediction: The sample dataset is mainly used to predict the outcomes of the training model. 6. Compare and Evaluate: In this step, performance of different models is evaluated and compared by different parameters like Precision, Recall, Accuracy, F-Score etc.

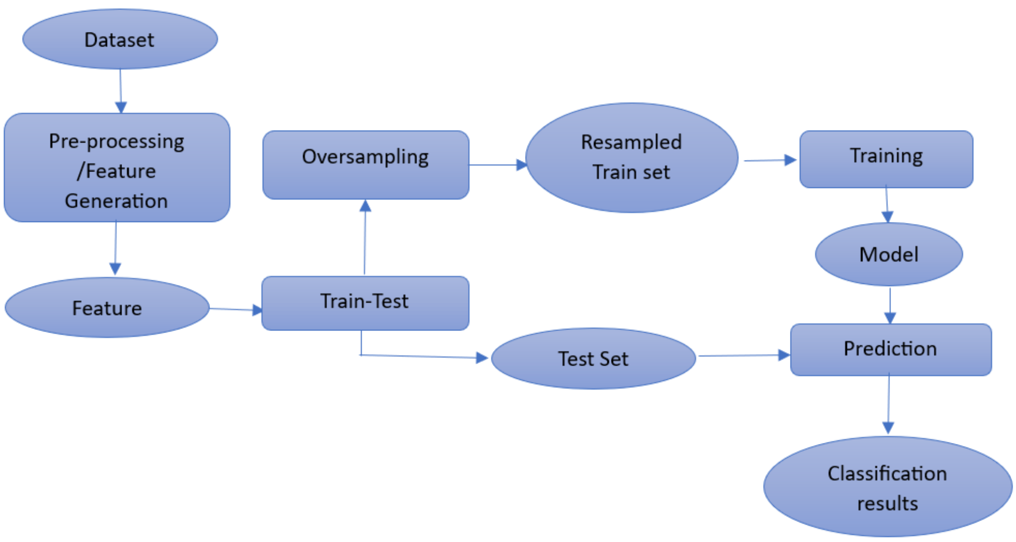


Fig 4.8: Various Phases of methodology

1) Dataset: A structured collection of data used to train, validate, or test machine learning models. It consists of individual data points, each containing features

(attributes or variables) and corresponding labels or target values. Datasets serve as the foundation for teaching machine learning algorithms by providing examples to learn patterns, relationships, or correlations between features and outcomes. They are usually separated into test sets, validation, and training sets. The test set calculates the performance parameter of the model on unlabeled data, while the validation set aids in model parameter optimization.

2) Pre-processing: Before raw data is put into a machine learning system, it must first undergo a number of data preparation stages known as preprocessing. Its main objective is to increase the quality and appropriateness of the data for modelling by cleaning, transforming, and organizing it.

3) Feature: A specific measurable property or characteristic of the data that is used as machine learning model input is referred to as a feature in the context of machine learning and data analysis. Predictors, characteristics, and independent variables are other terms for features. Features are essential for training machine learning models because they facilitate the recognition of patterns, connections, or structures in the data by algorithms. The effectiveness of a model and its capacity to produce precise predictions or classifications can be strongly impacted by the Caliber and applicability of its features. Refining the input features through feature selection, extraction, and engineering is a popular way to increase the efficacy of machine learning models.

4) Train – Test: In the context of machine learning, training and testing are two fundamental phases involved in developing and evaluating a predictive model:

Training Phase:

- A machine learning model is exposed to labelled data (training dataset) comprising input features and the labels that correspond to their goal outputs during the training phase.
- From the training data, the model discovers patterns, correlations, and dependencies among the target labels and the input features.

- The target is to minimize the difference between the target values and the expected outputs by fine-tuning the model's parameters, or coefficients.

Testing Phase:

- After the training is done, it needs to be evaluated for its performance and generalization ability on new, unseen data.
- The testing dataset is a different dataset that is given to the trained model; it was not used in the training process.
- Based on the input properties in the testing data, the model predicts the results; these predictions are then contrasted with the actual, known target values.
- Evaluation measures: The model's performance is evaluated by using some metrics like accuracy of the model, precision value, recall value, F1-score, and others. These measurements show how well the model can forecast fresh, unobserved data.

5) Training Model: Training a model refers to the method of training an algorithm on a dataset to learn patterns, relationships, or representations within the data. This process involves several steps:

- Selecting a Model: In this paper linear regression model is used based on the nature of the problem and the dataset characteristics.
- Preparing the Data: Dataset should be pre-processed by cleaning, transforming, and organizing the data. And this includes removing duplicate data, scaling features, tokenizing data, and dividing the dataset into training dataset and testing dataset.
- Training the Model: The prepared training dataset needed to feed into the algorithm is selected. The model studies from the given input features and corresponding target values by adjusting its values or weights during an iterative process to reduce the difference between the obtained outcomes and expected ones.
- Model Fitting: The model iterates through the training dataset multiple times (epochs or iterations) to optimize its parameters and improve its predictive performance.
- Evaluation: To determine if the trained model can generalize and produce correct predictions on fresh, new data, calculate its performance on a different test dataset.

- Hyperparameter Tuning: This process is done to improve its performance further. This can done with cross validation or random search, and so on.

- Deployment: When the model demonstrates satisfactory performance on the testing data, it can be implemented to make classifications on new, real-world data.

6) Classification Results: Classification results in machine learning refer to the outcomes obtained from a classifier model when it's applied to predict or classify instances into predefined categories or classes. The output of a classification task are usually the various measures that are used in measuring effectiveness of the model in the investment it has taken in order to make forecasts as compared to the actual results.

CHAPTER 5

RESULT & ANALYSIS

The paper evaluating performance of various algorithms in-depth and offered significant details about the way they work. The analysis focused on using machine learning strategies to differentiate among two separate classes: the relationship between the speech features under consideration and the presence of the human voice that was positive and the negative correlation. This was made clear by the measurements like F1 score, Precision, and Recall that show the increased level of difficulty in this binary categorization. These metrics enabled the evaluation of the models up to the best level of detail. While Precision and Recall decided a precise perception of the model and how resourceful it is in underlining significant cases on the other hand, F1 score gave a composite randomized conclusion of the model's accuracy. This strict assessment made it possible for the author to have a clear insight in the effectiveness of an algorithm and shortcomings which in return was useful in the development of the detection systems aimed to prevent hate speech.

The Graph below showing the performance measures for the algorithm considered:

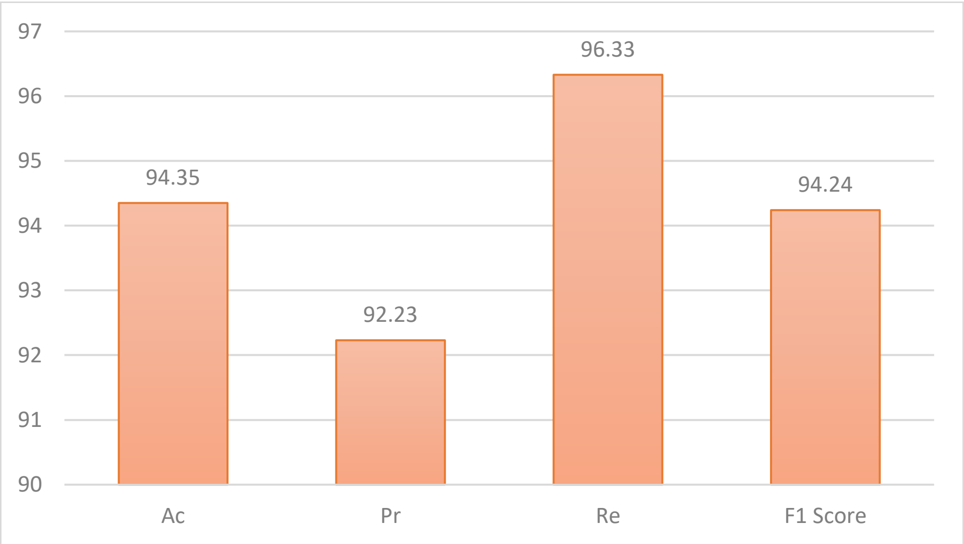


Fig 5.1: Logistic Regression Performance Metrics

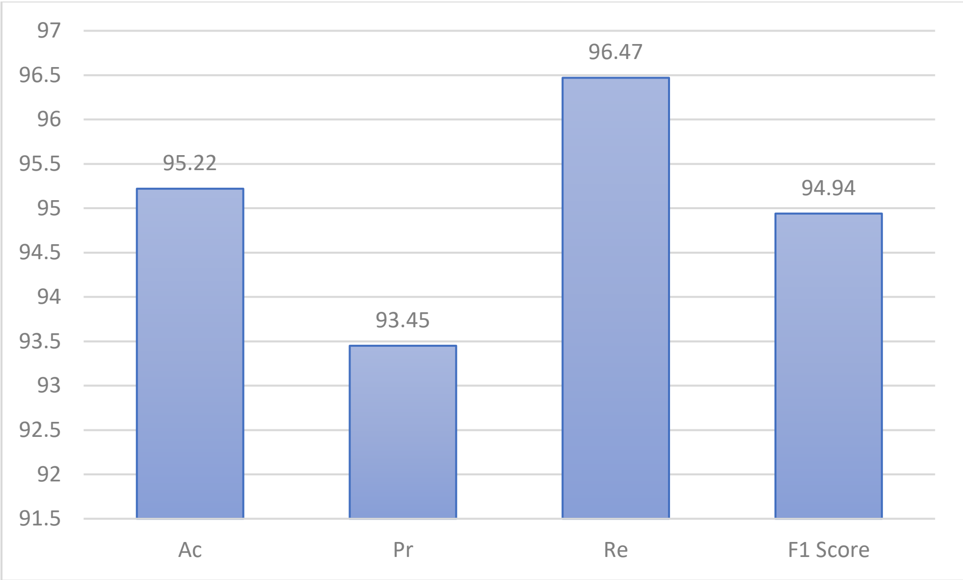


Fig 5.2: Random Forest Performance Metrics

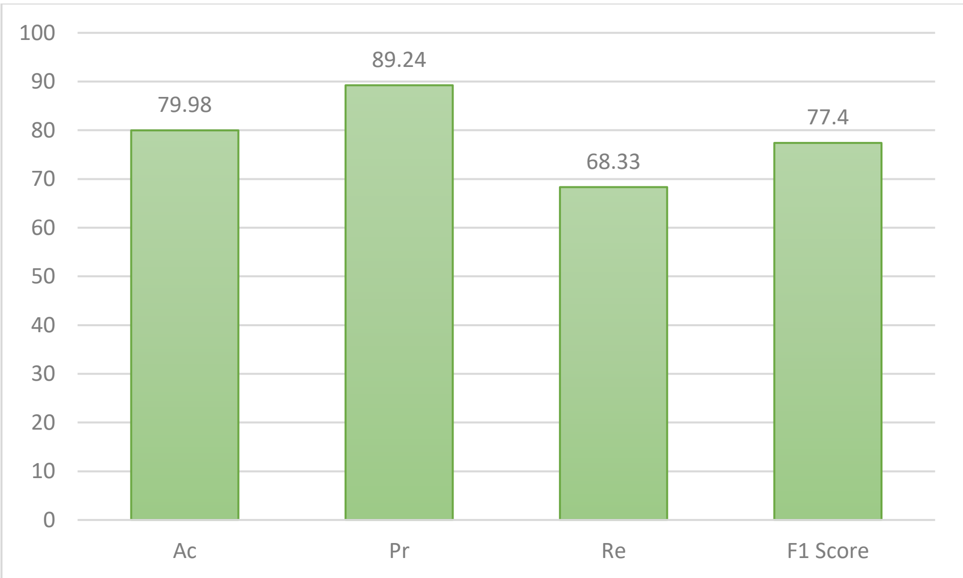


Fig 5.3: Naïve Bayes Classifier Performance Metrics

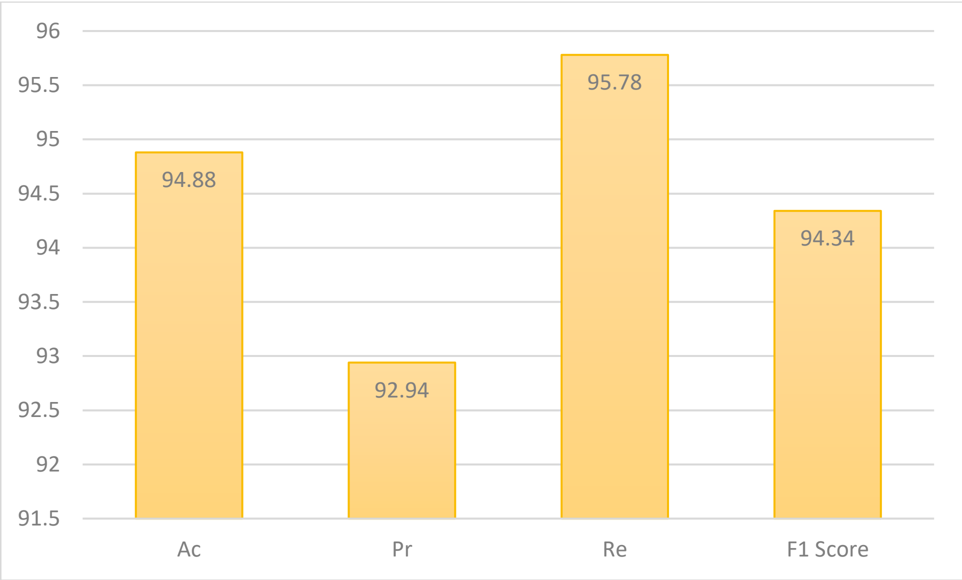


Fig 5.4: Support Vector Machine Performance Metrics

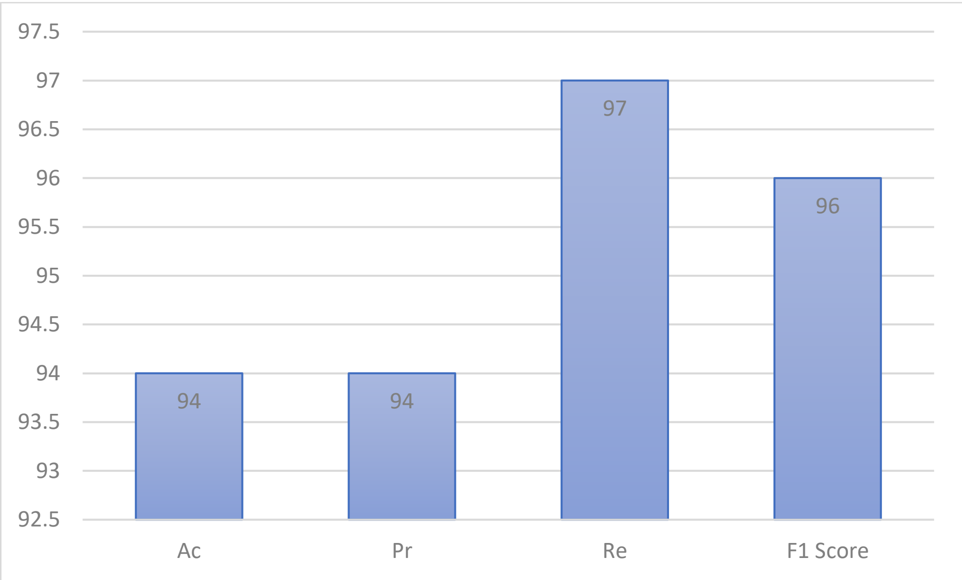


Fig 5.5: Artificial Neural Network Performance Metrics

After preprocessing the test dataset by applying tokenizing, stop word removal, duplicate data removal step by step, the data was labelled and categorized into two sets: label 0 is for the tweets that are not considered as obscene or abusive, and label 1 for the tweets that contain abusive content.

The below graph shows the data labelled using color along with 0 and 1 value.

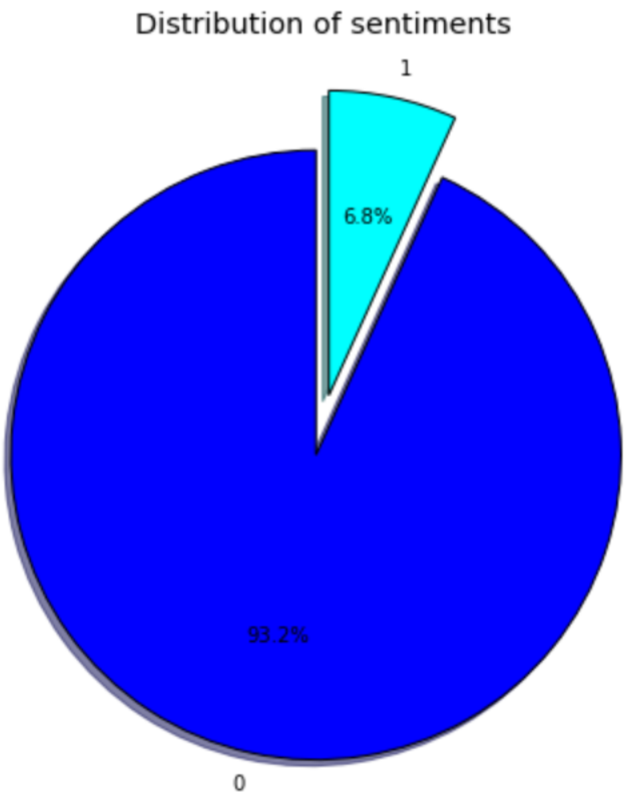


Fig 5.6: Graph represents labelled data

	id	label	tweet
0	1	0	user father dysfunctional selfish drags kids d...
1	2	0	user user happy thanks lyft credit cant use ca...
2	3	0	happy birthday love
3	4	0	model love u take u time ur disrespect
4	5	0	factsguide society motivation cunning despise ...

Fig 5.7: 0 - non-offensive tweets

	id	label	tweet
13	14	1	user cnn calls michigan middle school build wa...
14	15	1	comment australia opkillingbay seashepherd hel...
17	18	1	retweet agree
23	24	1	user user lumpy says prove lumpy
34	35	1	unbelievable 21st century wed need something l...

Fig 5.8: 1 – offensive tweets

TABLE 5.1: Comparative Analysis of the Proposed System

Model	**Ac**	**Pr**	**Re**	**F1 Score**
Logistic Regression	94.35	92.23	96.33	94.24
Random Forest	95.22	93.45	96.47	94.94
Naïve Bayes	79.98	89.24	68.33	77.40
Support Vector Machine	94.88	92.94	95.78	94.34
Artificial Neural Network	94	94	97	96

By comparing 5 different models based on their performance we can conclude that Random Forest Classifier algorithms works best with highest accuracy of 95.22%.

CHAPTER 6

CONCLUSION & FUTURE SCOPE

6.1 CONCLUSION:

The study displays the various uses of machine learning (ML) and data analysis techniques. It pulls valuable information from raw data and illustrates how a model used in this study may be used to identify hate speech online. The accuracy and efficiency of ML algorithms increase with experience. They are able to make better judgments as a result. As the volume of data increases, these algorithms get faster and more accurate in making predictions. Thousands of tweets make up the sample Twitter data, and an ongoing examination of it reveals that people frequently vent on the internet about things that could infuriate them. we have included some of the supervised learning approaches, like Naïve Bayes, also Random Forest, also Support Vector Machine and Logistic Regression. Along with the comparative analysis it is also included a performance measure data table according to referred paper mention. As offensive language continues to pose societal challenges, future research in this domain should aim to enhance existing methodologies, explore diverse deep learning methods, and extend analysis to real-world datasets. Such endeavours can substantially contribute to developing more robust and efficient models capable of combatting the proliferation of offensive language in online platforms, promoting a digital environment that is more inclusive and safer.

6.2 FUTURE SCOPE

Given the continuous progress in research and technology, the field of hate speech identification appears to have bright future prospects. By creating increasingly complex algorithms and utilizing cutting-edge natural language processing techniques, researchers are attempting to increase the accuracy of hate speech detection models. This entails taking into account cultural allusions, sarcasm, and contextual data to better comprehend the subtleties of offensive speech. Offensive speech can appear in a variety of media formats, including photos and videos, and can be expressed in a number of languages. Future advancements seek to identify offensive information across many media by incorporating multimodal analysis and extending hate speech detection to multiple languages. Prompt intervention and efficient online platform moderation depend on the prompt detection of hate speech. The development of models and systems that can instantly handle and analyse user-generated information and streaming text may be the main emphasis of future improvements.

REFERENCES

[1] Pak Alexender, et al. "Twitter as a corpus for sentiment analysis and opinion mining." Proceedings of the Seventh conference on International Language Resources and Evaluation, vol 10,2010, pp. 1320-1326.

[2] Gitari, Njagi Dennis, et al. "A Lexicon-Based Approach for Hate Speech Detection." International Journal of Multimedia and Ubiquitous Engineering, vol. 10, no. 4, 2015, pp. 215–230.

[3] Sari & B. Ginting. (2019). Hate speech detection on twitter using multinomial logistic regression classification method, pp. 105–111

[4] Malik, J. S., Pang, G., & Hengel, A. V. D. (2022). Deep learning for hate speech detection: a comparative study. arXiv preprint arXiv:2202.09517

[5] Raufi, B., & Xhaferri, I. (2018, September). Application of machine learning techniques for hate speech detection in mobile applications. In 2018 International Conference on Information Technologies (InfoTech) (pp. 1-4). IEEE.

[6] Chen, Y., Zhu, S., Zhou, Y., & Xu, H. (n.d.). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety.

[7] S. Ahammed, M. Rahman, H. M. Niloy and S. M. H. Chowdhury, "Implementation of Machine Learning to Detect Hate Speech in Bangla Language," in International Conference on System Modeling & Advancement in Research Trends, Moradabad, India, 2019

[8] Samghabadi, Niloofar Safi, et al. "Detecting nastiness in social media." Proceedings of the First Workshop on Abusive Language Online. 2017

[9] Yao, Mengfan, Charalampos Chelmis, and Daphney? Stavroula Zois. "Cyberbullying ends here: Towards robust detection of cyberbullying in social media." The World Wide Web Conference. 2019.

[10] DataTurks. (2018, July 12). Tweets Dataset for Detection of Cyber-Trolls. Retrieved November 07, 2020, from https://www.kaggle.com/dataturks/dataset-for-detection-of-cybertrolls?select=Dataset+for+Detection+of+Cyber-Trolls.json

[11] Gitari, N.D., et al., A lexicon-based approach for hate speech detection.International Journal of Multimedia and Ubiquitous Engineering, 2015.10(4): p. 215-230

[12] James A. Nichols, Hsien W. Herbert Chan and Matthew A. B. Baker- Machine learning: applications of artificial intelligence to imaging and diagnosis (11 Feb, 2019).

[13] Kaggle, https://www.kaggle.com/eldrich/hate-speechoffensive-tweets-by davidson-et-al.

[14] Hodeghatta, Umesh Rao. "Sentiment Analysis of Hollywood Movies on Twitter." Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM 13, 2013.

[15] Hinduja, Sameer, and Justin W. Patchin. "Bullying, Cyberbullying and Suicide." Archives of Suicide Research Cyberbullying and Suicide." Archives of Suicide Research, vol. 14, no. 3, 2010, pp.206- 221.

[16] Nobata, Chikashi, et al." Abusive Language Detection in Online User Content." Proceedings of the 25th International Conference on World Wide Web - WWW 16, 2016. [1] J clement, "https://www.statista.com/statistics/282087/number-ofmonthly-active-twitter-users/," Number of monthly active Twitter users worldwide, 2021.

[17] S. A. Devi, P. Sapkota and M. Obulesh, "Sentiment analysis on products using social media," Journal of Advanced Research in Dynamical and Control Systems, pp. 137-141, 2017. [3] M. Bhargava and D. Rao , "Sentimental analysis on social media data using R programming," International Journal of Engineering and Technology(UAE), vol. 7, no. 2, pp. 80-84, 2018.

[18] C. G. Krishna , D. R. Meka, V. S. Vamsi and K. M. V. S. Ravi , "A survey on twitter sentimental analysis with machine learning techniques," International Journal of Engineering and Technology(UAE), vol. 7, no. 2.32, pp. 462-465, 2018.

[19] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," pp. 88–93, 2016, DOI: 10.18653/v1/n16-2013.

[20] P. Burnap and M. L. Williams, "Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making," Policy and Internet, vol. 7, no. 2, pp. 223–242, 2015, DOI: 10.1002/poi3.85.

[21] P. Burnap and M. Williams, "Hate Speech, Machine Classification and Statistical Modelling of Information Flows on Twitter: Interpretation and Communication for Policy Decision Making," in Internet, Policy & Politics, 2014, pp. 1–18, [Online]. Available: http://orca.cf.ac.uk/id/eprint/65227%0A.

[22] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, Shivakant Mishra. (2015). Detection of Cyberbullying Incidents on the Instagram Social Network."

[23] Dadvar, Maral Eckert, Kai. (2018). Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study. 10.13140/RG.2.2.16187.87846.

[24] Nandhini, B. Sri, and J. I. Sheeba. "Cyberbullying detection and classification using information retrieval algorithm." Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering Technology (ICARCSET 2015). 2015.

[25] Chen, Y., Zhu, S., Zhou, Y., & Xu, H. (n.d.). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety.

[26] Bhandary, U. (2019). Detection of Hate Speech in Videos Using Machine Learning. https://doi.org/10.31979/etd.5efp-73s4

[27] McAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PLoS ONE, 14(8). https://doi.org/10.1371/JOURNAL.PONE.0221152

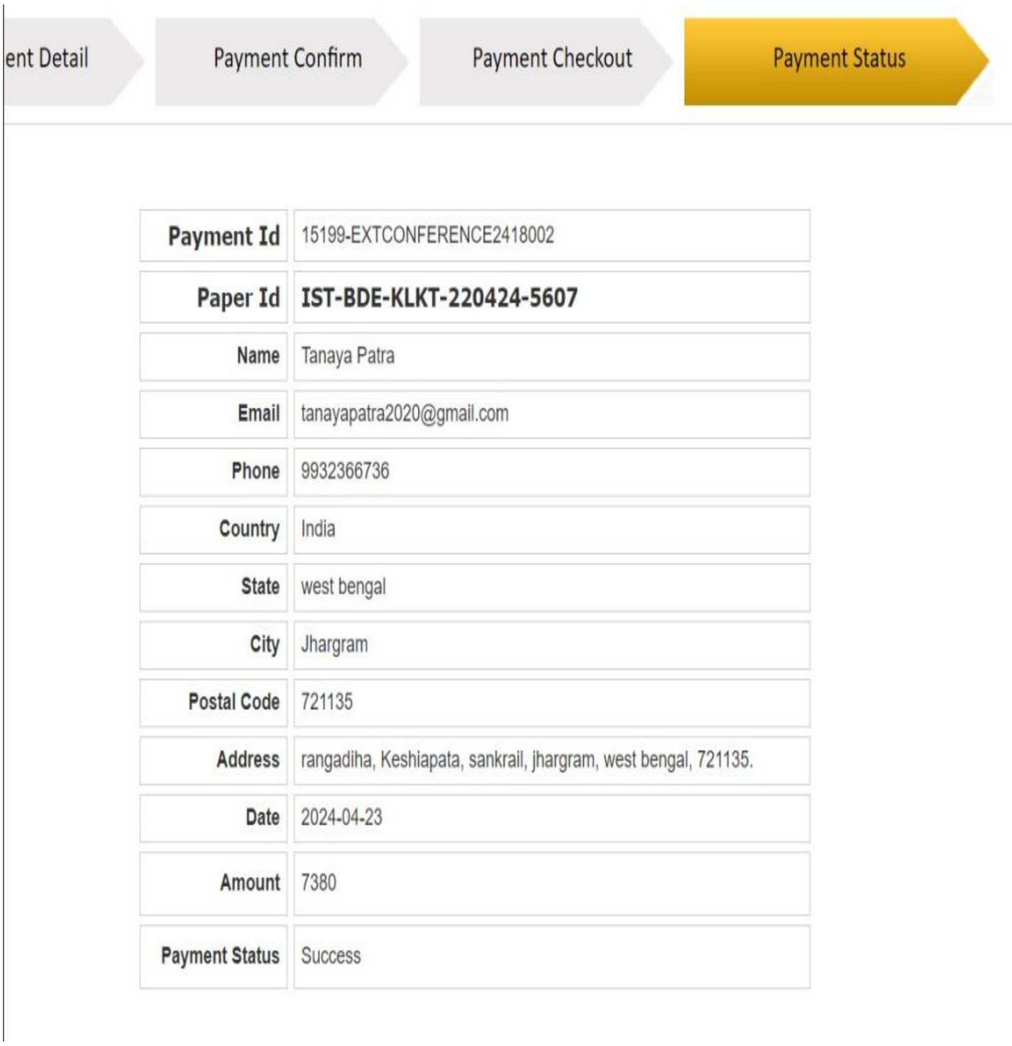
[28] IEEE Xplore Full-Text PDF: (n.d.). Retrieved March 24, 2023, from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9455353

[29] Hasan, A., Sharma, T., Khan, A., & Hasan Ali Al-Abyadh, M. (2022). Analysing Hate Speech against Migrants and Women through Tweets Using Ensembled Deep Learning Model. Computational Intelligence and Neuroscience, 2022.https://doi.org/10.1155/2022/8153791

LIST OF PUBLICATIONS AND THEIR PROOFS

1. Tanaya Patra "Offensive Language Detection from Social Media text using Machine Learning Classification Methods: A Review", Accepted at "**International Conference on Artificial Intelligence, Machine Learning and Big Data Engineering (ICAIMLBDE) organized by ISETE**" on 22nd April 2024 at Kolkata, India.

   Paper Id: IST-BDE-KLKT-220424-5607



ent Detail | Payment Confirm | Payment Checkout | Payment Status

Payment Id	15199-EXTCONFERENCE2418002
Paper Id	**IST-BDE-KLKT-220424-5607**
Name	Tanaya Patra
Email	tanayapatra2020@gmail.com
Phone	9932366736
Country	India
State	west bengal
City	Jhargram
Postal Code	721135
Address	rangadiha, Keshiapata, sankrail, jhargram, west bengal, 721135.
Date	2024-04-23
Amount	7380
Payment Status	Success

Gmail Tanaya Patra <tanayapatra2020@gmail.com>

PAPER ACCEPTANCE CONFIRMATION

1 message

ISETE <info.iseteconference@gmail.com> 23 April 2024 at 14:58
To: "tanayapatra2020@gmail.com" <tanayapatra2020@gmail.com>

Dear Tanaya,
Greetings and best wishes of the day !!

Our International Conference has accepted your Paper entitled **"Offensive Language Detection from Social Media text using Machine Learning Classification Methods: A Review"** with Paper ID **IST-BDE-KLKT-220424-5607**

Mode : Online/Offline (Hybrid)

Note : Today is the last day of Registration.
There are no additional charges for Additional Certificates for Additional Authors and Publication.
To avoid GST you can make a Bank Transfer through any UPI Application by using our Account Number and IFSC Code.

Registration link http://paymentnow.in/

OR

Bank Details
Name: Institute of Research and Journals
A/c No. 33547315754
IFSC CODE: SBIN0010927
SWIFT CODE: SBININBB270 (For International users)
Bank Address - SBI Khandagiri, BBSR

Only after your registration has been confirmed will you receive a Formal acceptance letter and Conference Schedule.

Categories	Registration Fee For Author outside of India	Registration Fee For Author of India
Authors (Academician/Practitioner)	300 USD	8800 INR
Authors (Student M.tech/PhD)	250 USD	7200 INR
Authors (B.Tech)	200 USD	6200 INR
Listeners:	150 USD	3000 INR
Additional Paper (s)**	100 USD	3000 INR

Thank You
Regards,
Deepak Swain

Conference Coordinator
Asst. Organizing Secretary, International Division
ISETE
Call/Whatsapp: +91-8895188931
E-mail: info.iseteconference@gmail.com
Web: www.isete.org
Follow us @ Facebook and Twitter

IST-BDE-KLKT-220424-5607

INTERNATIONAL SOCIETY FOR ENGINEERING AND TECHNICAL EDUCATION

International Conference on

Artificial Intelligence, Machine Learning and Big Data Engineering

Organized by: ISETE I Kolkata, India I 22nd April 2024

Certificate
of Presentation

This is to certify that Tanaya Patra has presented a paper entitled "Offensive Language Detection from Social Media Text using Machine Learning Classification Methods: A Review" at the International Conference on Artificial Intelligence, Machine Learning and Big Data Engineering (ICAIMLBDE) held in Kolkata, India on 22nd April, 2024.

Conference Coordinator
International Society for Engineering and Technical Education

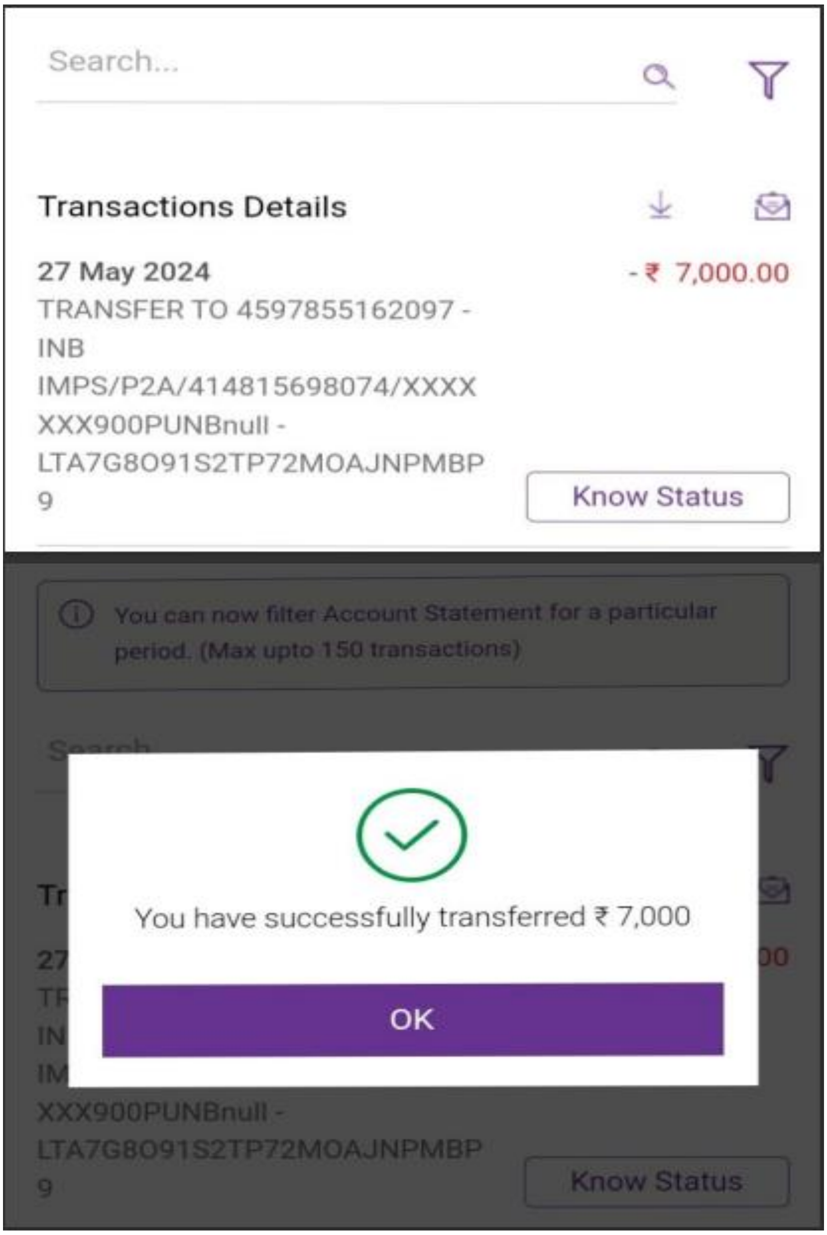
Chairman
International Society for Engineering and Technical Education

www.isete.org

info.iseteconference@gmail.com

2. Tanaya Patra “Offensive Speech Identification from Social Media Text by Applying Machine Learning Classification Algorithm: Logistic Regression”, Accepted at “**International Conference on Intelligent Computing and Communication Techniques at JNU New Delhi, India**” on 28th & 29th June 2024 at New Delhi, India.

   Paper Id: 916



Gmail

Tanaya Patra <tanayapatra2020@gmail.com>

Notification of acceptance of paper id 916

1 message

Microsoft CMT <email@msr-cmt.org> 27 May 2024 at 10:36
Reply-To: ICICCT 2024 <icicctcon@gmail.com>
To: Tanaya Patra <tanayapatra2020@gmail.com>

Dear Dr./ Prof. Tanaya Patra,

Congratulations...

Your paper / article paper id 916: Offensive Speech Identification from Social Media Text by Applying Machine Learning Classification Algorithm: Logistic Regression has been accepted for publication in International Conference on Intelligent Computing and Communication Techniques at JNU New Delhi, India.

Kindly save your paper by given paper id only (eg. 346.docx, 346.pdf, 346_copyright.pdf)

Registration Link:

https://forms.gle/mSsHa8GMLtMkWuaq8

Please ensure the following before registration and uploading camera ready paper.
1. Paper must be in Taylor and Frances Format.
Template and copyright with author instruction are given in below link: https://icicct.in/author_inst.html

2. Minimum 12 references should be cited in the paper and all references must be cited in the body. Please follow the template.
3. The typographical and grammatical errors must be carefully looked at your end.
4. Complete the copyright form (available at template folder).
5. The regular fee (Available in registration section) will be charged up to 6 pages and after that additional Rs.1000 for Indian authors / 10 USD for foreign authors per additional page will be charged.
6. Reduce the Plagiarism below 10% excluding references and AI Plagiarism 0%. The Authors are solely responsible for any exclusion of publication if any.
7. Certificate will be issued by the name of registered author (Single author only).
8. Certificates may be issued to all other authors on the extra payment of 1000/- INR per author.
9. Last Date of registration and uploading copyright and camera-ready copy: 31/05/2024.
10. Make a single payment which includes registration fee + Extra certificates fee + Extra page fees.
11. Permissions: Kindly make sure the permissions for each copyrighted artwork file have been cleared ahead of the submission, with the details listed in the Permission Verification form (attached). All permission grants must be submitted along with your final manuscript.
12. Each Illustration must include a caption and an alternative text description to assist print impaired readers ('Alt Text').
(Alt Text is mandatory for each Illustrations)
Figures: Please make sure no figures are missing, and all figures are high resolution and alt text is included
Tables: Please ensure that there are no missing tables, and the tables in your manuscript are not pasted as figures.
Citation: Kindly ensure there are no missing citations in your manuscript

Registration Link: https://forms.gle/mSsHa8GMLtMkWuaq8

Registration Fee to be deposited in below account

Bank Account Details :

Indian Account Details:

Account Holder Name: EVEDANT Foundation

Account Number: 0674002190422900

IFSC Code: PUNB0067400

Gmail Tanaya Patra <tanayapatra2020@gmail.com>

ICICCTJNU- Submission Sheet Taylor and Frances
1 message

Google Forms <forms-receipts-noreply@google.com> 27 May 2024 at 16:08
To: tanayapatra2020@gmail.com

Thanks for filling in ICICCTJNU- Submission Sheet Taylor and Frances

Here's what was received.

ICICCTJNU- Submission Sheet Taylor and Frances

Kindly fill the data as per given in research paper
kindly follow the same serial order of the authors as given in research paper.

It is again advised to all the registered authors to check all the prescribed details and submit the paper in **CRC template only.**
As per Instructions Received from Taylor and Frances, you need to properly check the Grammar, Language and plagiarism in your paper.
Plagiarism should be less than 10 %

Kindly save your paper by given paper id only (eg. 346.docx, 346.pdf, 346_copyright.pdf)

Email *

tanayapatra2020@gmail.com

Paper ID (Digits Only e.g. 101, 45,) *

916

Similarity Report

PAPER NAME

main_content_new (3).pdf

WORD COUNT	CHARACTER COUNT
10398 Words	**57474 Characters**
PAGE COUNT	FILE SIZE
46 Pages	**983.1KB**
SUBMISSION DATE	REPORT DATE
May 24, 2024 5:14 PM GMT+5:30	**May 24, 2024 5:15 PM GMT+5:30**

● 7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 4% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database
- 5% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material
- Cited material
- Small Matches (Less then 8 words)

Summary

main_content_new (3).pdf

My Files

My Files

Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:59926845

Submission Date

May 24, 2024, 5:14 PM GMT+5:30

Download Date

May 24, 2024, 5:17 PM GMT+5:30

File Name

main_content_new (3).pdf

File Size

983.1 KB

46 Pages

10,398 Words

57,474 Characters

turnitin Page 1 of 48 - Cover Page Submission ID trn:oid:::27535:59926845

How much of this submission has been generated by AI?

0%

of qualifying text in this submission has been determined to be generated by AI.

Caution: Percentage may not indicate academic misconduct. Review required.

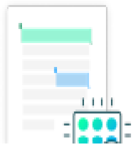
It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

What does the percentage mean?

The percentage shown in the AI writing detection indicator and in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was generated by AI.

Our testing has found that there is a higher incidence of false positives when the percentage is less than 20. In order to reduce the likelihood of misinterpretation, the AI indicator will display an asterisk for percentages less than 20 to call attention to the fact that the score is less reliable.



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis ____

____

Total Pages ______________ Name of the Scholar ______________________________

Supervisor (s)

(1) ______________________________

(2) ______________________________

(3) ______________________________

Department ____

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: ____________________ Similarity Index: ______, Total Word Count: ________

Date: ___________

Candidate's Signature **Signature of Supervisor(s)**

BRIEF PROFILE

I am Tanaya Patra, pursuing my MTech in Computer Science and Engineering from Delhi Technological University. Currently, I am in the final semester of my degree and I scored 8.23 CGPA in the first three semesters of my MTech.

I completed my BTech in Computer Science and Engineering from Raja Reddy Institute of Technology Bangalore in 2019 which comes under Visvesvaraya Technological University (Belagavi, Karnataka). After that I prepared for Gate exam through Unacademy for a year and worked as an Assistant System Engineer Trainee in TCS where I worked as java backend developer.

I have developed many skills in our computer science domain. Mainly learned many languages like java, C++, Python and many scripting languages like HTML, CSS, JavaScript, NodeJS. While working with TCS I got familiar with may tools like MongoDB, PosmanAPI etc. Also solved many problems over GFG and leetcode platform for my placement preparation.