Uncovering Cyberbullying using Machine Learning

|  |  |  |
| --- | --- | --- |
| Dr. M. Sakthivel,  Professor, Department of CSE  Sree Vidyanikethan Engineering College, Tirupati, Andhra Pradesh, India [sakthivel.m@vidyanikethan.edu](mailto:sakthivel.m@vidyanikethan.edu) | V. Nagesh,  UG Scholar, Department of CSE  Sree Vidyanikethan Engineering College,  Tirupati, Andhra Pradesh, India  [vaddiralanagesh0191@gmail.com](mailto:samanumanvithareddy@gmail.com) | V. Ganga Pradeep,  UG Scholar, Department of CSE Sree Vidyanikethan Engineering College,  Tirupati, Andhra Pradesh, India  [vallapupradeep@gmail.com](mailto:samanumanvithareddy@gmail.com) |
| V. Hemanth Kumar Reddy,  UG Scholar, Department of CSE Sree Vidyanikethan Engineering College,  Tirupati, Andhra Pradesh, India  [hemhemanth5239@gmail.com](mailto:samanumanvithareddy@gmail.com) | V. Sindhura  UG Scholar, Department of CSE Sree Vidyanikethan Engineering College,  Tirupati, Andhra Pradesh, India  [sindhuv1807@gmail.com](mailto:samanumanvithareddy@gmail.com) | T.Chirudeep  *UG Scholar, Department of CSE Sree Vidyanikethan Engineering College,*  Tirupati, Andhra Pradesh, India  [chirudeeptupakula@gmail.com](mailto:%20chirudeeptupakula@gmail.com) |

*Abstract:* cyberbullying has become increasingly common, there has been a lot of focus on cyberbullying detection because cyberbullying has a fatal effect on both users of social media platforms and society. Social media has become a channel for many people to communicate their opinions, knowledge, and so on. It would have been one of the best things if it had not become a tool for many people to exact revenge, manipulate others, or harass and humiliate others. As a result, we do not require a monitoring system to control the misbehavior and bullying that spreads through social networks, which has led to the development of this model to automate the identification of cyberbullying. “Our main goal is to create a model that categorizes comments as positive or negative.

Keywords: Cyber Bullying, Sentiment Analysis, Personality Analysis

1. INTRODUCTION

Cyberbullying is just bullying or harassing others via various Social Media platforms. As the technological era has progressed, the use of social media platforms has skyrocketed, with more than half of the world's population now utilizing them. As its popularity has grown, a piece of news or other information can now be shared with others in seconds. This is now one of its most essential applications. However, some people took advantage of this and began to seek enjoyment from it. Cyberbullying has since begun and has grown significantly. Cyberbullying is known as an unseen crime, and many victims have been tormented online through toxic comments. Research suggests that approximately 50% of children have been affected by cyberbullying, resulting in a range of negative consequences such as mental health issues, academic struggles, and even suicidal thoughts. Unfortunately, many victims do not receive the help and support they need due to factors such as social status. In addition to the emotional toll, cyberbullying can also damage a person's reputation through the spread of false rumors and harmful content. Approximately 8 out of 10 children are victims of various forms of cyberbullying. Machine learning is used to detect cyberbullying by identifying and classifying instances of online harassment, abuse, and bullying. This technique often entails training a model on a large dataset of cyberbullying cases as well as non-cyberbullying text examples. The model can then be used to predict whether future occurrences of online text are likely to be cyberbullying. Personality Analysis, Sentiment Analysis, and User Traits are some of the factors that the algorithm takes into account while making these predictions. The purpose of this technique is to create a rapid, scalable, and automated mechanism to detect and solve cyberbullying, so that online communities can be safer and more supportive for everyone.

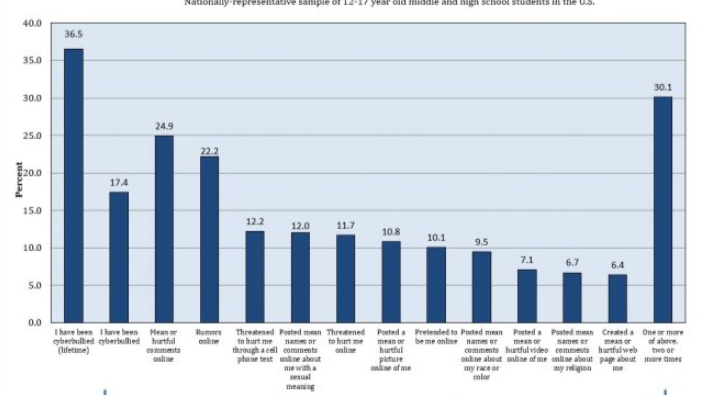


Fig.1.CyberBullying Victimization and Reasons



Fig.2. Types of Cyber Bullying

1. *RESEARCH OBJECTIVES*
2. We would be using various machine Learning Techniques to build a model that accurately identifies instances of cyberbullying.
3. To develop a model that Classifies the type of cyberbullying being used .Various types of cyberbullying being harassment, threats, spreading rumors, Flaming etc.
4. To train and test the model using the twitter dataset that is available online.
5. The dataset consists of both positive and negative contexts of cyberbullying.
6. The primary objective is to reduce the rate of cyberbullying.
7. Optimize the performance of model using ensemble algorithms.
8. *SCOPE OF THE WORK*

The proposed work will utilize the supervised learning approach to detect and address cyberbullying. The dataset will consist of two categories of data, positive and negative, related to cyberbullying. The goal is to identify instances of cyberbullying and transform negative comments into positive forms. The work will involve implementing multiple machine learning algorithms.

1. *PROBLEM STATEMENT*

Cyberbullying is the use of numerous social media platforms to manipulate, torment, or harass another person, and it is a developing concern in today's technology day. With the increased use of online platforms, cyberbullying has increased significantly, causing a slew of serious issues in victims such as depression and inferiority complexes. Machine Learning has various sets of algorithms that are used to develop various types of predictive models. This has simplified many challenges, one of which is cyberbullying detection. Cyberbullying has become more common as the use of social media platforms has grown.

1. LITERATURE SURVEY

The field of Machine Learning offers diverse algorithms that aid in the creation of predictive models for various applications, including the detection of cyberbullying. With the increasing use of social media platforms, cyberbullying has become prevalent, necessitating the need for models that can recognize such instances in real-time, given the constant growth of data. This can be achieved by utilizing machine learning techniques such as Bayes, Decision Tree, and Foresting, as well as ensemble algorithms that optimize the model.

A. RELATED WORK:

Andreas Weiler et al investigated the run-time and task-based performance of numerous Twitter event recognition algorithms. The writers took a two-pronged approach. The scholars gathered a dataset of cyberbullying-related tweets and developed a ML classifier to identify cyberbullying tweets, the classifier used a SVM classifier algorithm and a collection of variables such as sentiment analysis, frequency of abusive terms, and user mentions.

Chen et al. (2012) conducted a study that developed a method for detecting offensive language using a syntactic element which outperformed the standard learning-based method. In a similar study, Dadvar et al. (2013) utilized SVM to detect cyberbullying in YouTube data and found that integrating user-generated content improved the results of SVM detection. Dadvar et al. used MySpace data sets to reach their conclusions.

The research conducted by Cynthia Van Hee and Gilles Jacobs focused on automatic cyberbullying identification in social media text. They gathered data from social networking sites and used online bullying standards to create corpora for modeling postings made by bullies and victims. Common natural language processing pre-processing steps such as tokenization and lemmatization were utilized to convert unstructured text into a structured text that can be used for Classification and Regression algorithm. Although the accuracy results of 64% for English and 61% for Dutch were modest, they serve as a promising starting point for further research and improvements to the SVM ML algorithm.

Hannah L. Schacter et al investigated how cyberbullying victims' disclosures on Facebook influence bystanders' features of guilt, empathy, and ideas of intervening on behalf of a casualty following the completion of a cyberbullying incident. The study's findings revealed that participants who examined the high personal disclosure profile, regardless of valence, blamed the victim more and felt less empathy for the victim. As a result, the likelihood of bystander intervention in the bullying incidence was anticipated to be lower.

The study by S.E. Vishwa Priya, Ajay Gour, and their team on uncovering hate speech and objection language on Social media tweet is noteworthy. The researchers utilized labelled datasets from Crowd flower and GitHub, which had categories such as Hateful, Offensive, Sexism, and Racism. The team employed N-gram features with TF-IDF weighting, which is a popular approach for text classification. These findings suggest that their approach has great potential for automated detection and prevention of hate speech and objectionable language on Twitter.

Matthew Pittman and Brandon Reich performed a mixed-design survey to investigate whether text- and image-based social media sites like Twitter and Yik Yak, as well as platforms like Instagram and Snapchat, can help reduce loneliness by promoting more intimacy. Use of text-based media, in contrast, seemed ineffectual in this regard. The qualitative findings revealed that the increased intimacy provided by image-based social media use was what caused the observed impacts. Using image-based social media platforms reportedly increased participants' sense of community and sense of shared experience.

College students' experiences with various types of social media cyberbullying on social networking sites were the subject of a study by Kassandra Gahagan. (SNS). The study's findings revealed that 46% of college students had seen cyberbullying on SNS and that 19% of them had experienced bullying on the platform. In addition, 61% of college students who saw cyberbullying on SNS took no action.

H. Watanbe, M. Bouaizi, and T. Ohtsui's research on detecting hate speech on Twitter using unigrams and automatically gathered patterns. They included tweets categorized as clean, offensive, hateful, sexist, racist, or neither. Pooling these datasets together is a common approach to create a larger and more diverse dataset for machine learning Using unigrams and automatically gathered patterns from the dataset is a popular approach for text classification tasks, especially for detecting hate speech and offensive language. However, it's important to note that unigrams may not capture the full context and meaning of the text, so it may be necessary to consider other features such as bigrams, trigrams, or even semantic features.

In their research, Lida Ketsbaia, Biju Issac, and colleagues focused on identifying harmful texts using ML and Deep learning. One notable aspect of their approach was the use of datasets from two different universities, with one labeled as "Hate" and the other labeled as "Non-Hate," to train their model. This allowed them to create a more diverse and comprehensive dataset for their research.

1. METHODOLOGY
2. *MODEL DESCRIPTION:*

# Data collection:

To build a machine learning model for identifying cyberbullying, the first step is to gather a comprehensive and varied dataset of text that includes examples of both offensive and non-offensive tweets. This dataset can be obtained from social media platforms, online forums, and comments sections. The aim is to ensure that the dataset is large and diverse, with a wide range of texts representing different contexts and scenarios.

# Data Pre-processing:

In the data preprocessing step, the collected data is cleaned and standardized to prepare it for machine learning algorithms. Additionally, text normalization techniques like converting all text to lowercase may be applied to ensure consistency in the data. The goal of data preprocessing is to obtain a clean dataset that can be used for feature engineering and model training.

**Feature Extraction:**

Feature extraction is a crucial step in building a machine learning model for text classification. It involves selecting the features that the model will use to make its predictions. In the case of text, this typically involves creating numerical representations of the text that capture the most important information. This can be achieved through word embeddings, which convert the text into numerical vectors. These features are then used as input to the classifier algorithms for training, testing and prediction.

**Model Selection:**

Choose an appropriate algorithm for the task of text classification, such as a SVM Classifier or a foresting algorithm.

**Model Training:**

Train the model on the preprocessed and feature-engineered data. This involves providing the model with the features and corresponding labels (e.g., cyber bullying or non-cyber bullying) for each instance in the training data.

**Model Evaluation:**

The evaluation step to test the model on a separate test dataset and computing metrics such as accuracy, precision, confusion matrix. These metrics help to assess the model's effectiveness in detecting cyberbullying. By evaluating the model's performance, we can identify potential areas for improvement and fine-tune the model further to enhance its accuracy and effectiveness in detecting cyberbullying.

**Model Fine-tuning:**

Fine-tuning the model involves adjusting its parameters or testing different algorithms to improve its performance, and repeating the evaluation step if necessary. This step is crucial as it helps to optimize the model and improve its accuracy. Contingent upon the consequences of the assessment, the model might should be tweaked by changing its boundaries or attempting an alternate calculation to accomplish improved results. By repeating the evaluation step, we can assess whether the changes made to the model have improved its performance. Overall, fine-tuning is a necessary step in the model development process to ensure that the model is as accurate and effective as possible.

# *B CYBER BULLYING DETECTION MODEL :*

The cyberbullying detection framework consists of two key components: N L Processing (NLP) and Artificial Intelligence. To train our model, we collected real-time tweets from various internet sources and platforms. Pre-processing text data is an essential step in NLP to standardize the data and remove unwanted characters such as emojis, special characters, and punctuation that may interfere with the analysis and modeling process. Therefore, we cleaned and prepared the data for detection by removing unwanted patterns such as numerical characters, hashtags, and special symbols using NumPy and vectorize routines.

We then utilized NLP techniques such as tokenization, lemmatization, and vectorization to transform raw text into numerical vectors. After completing pre-processing, we divide the dataset into test data and train data. We used two significant text selection functions, Inverse Frequency and Count Vectorizer, to train our model. We also employed a variety of machine learning algorithms such as SVM, decision tree, naive bayes, Logistic Regression, Random Forest, and K Neighbors Classifier to evaluate the accuracy for each model in the second phase.

To find the best suitable combination of dimension selection methods such as term frequency-inverse document frequency and count vectorizers and machine learning models, we compared both TF-IDF and count vectorizer. Using these techniques, our model can predict whether the text belongs to the context of bullying or not in English.

**i. Data collection**

We collected a dataset consisting of both English and Hinglish text, which contained tweets from various social media platforms and networking websites. The dataset consisted of approximately 15,307 entries in English, and for the Hinglish dataset, we extracted tweets and WhatsApp messages in real-time, resulting in approximately 9,482 entries in total. Human annotators manually labeled all entries as toxic or non-toxic. In addition, we extracted tweets and comments from YouTube and Twitter chats, which were combined to create a larger dataset of approximately 3000 entries

**ii. Data Preparation**

In order to prepare the data for machine learning models, it is crucial to clean the data by removing unnecessary characters, symbols, and stop words. This step can significantly enhance the accuracy of the model and prevent overfitting. It is commendable that you have already taken the necessary steps to clean the data and remove unwanted words and symbols. Additionally, it is worth noting that while a manual list of stop words can be helpful, there are also pre-built libraries and packages that can automatically remove stop words based on the language.

**iii. Preprocessing Methodologies**

Because the algorithm cannot operate directly with unordered and uncleaned text that is, it cannot understand the sentences provided to it we used natural language processing techniques after cleaning the data. To do this, we transformed the sentences into an understandable format using some preprocessing techniques. We use the following

• Tokenization - Breaking down a text sequence into smaller units, or tokens, like words, phrases, ideas, and symbols, is referred to as tokenization. It is a critical stage in natural language processing and is frequently carried out as a pre-processing stage before text data is fed into machine learning models. Top of Form

• Lemmatization - Lemmatization is a crucial process in natural language processing that involves transforming a word into its base. This process is important because it helps to decrease the number of distinct words and variations present in a text document, thus improving the accuracy of text analysis.

**iv. Data segmentation:**

In the process of building a machine learning model for text analysis, namely training data and testing data. The testing data is obtained by extracting text data from various sources using techniques such as text mining. Both the training and testing data go through preprocessing steps and are then used for training and evaluating multiple machine learning models.

**v. Feature choice:**

After separating the data, an essential NLP step called feature selection entails extracting pertinent text features. This technique aids in evaluating the precision of the final vector representations. It works by locating words with comparable meanings that frequently occur close to terms. The text features are then vectorized into numerical vectors before the data is fed into the machine learning models for training and assessment.

**1) Count Vectorization:** A common Natural Language Processing (NLP) method called count vectorization turns a group of written documents into a matrix of token counts. This method entails breaking the text down into individual words or tokens, noting the frequency of each token within the document, and then converting the counts into numerical values. The outcome is a sparse matrix of integers, where each row denotes a document and each column denotes a distinct token within the collection. Numerous NLP applications, including document classification, sentiment analysis, and topic modelling, depend on count vectorization as a basic step.

2) **TF-IDF**: TF-IDF, an acronym for Term Frequency-Inverse Document Frequency, is a numerical metric that measures the relevance of a word in a document within a corpus. This metric is frequently employed in natural language processing and information retrieval to conduct text analysis and ranking in search engines. The importance of a word is determined by both its frequency in the document and the frequency of the word in the corpus. Top of Form Bottom of Form

**1) SVM (Support Vector Machine):** Support Vector Machines (SVMs) are a class of machine learning algorithm that have been extensively used in many fields, including image classification, text classification, and bioinformatics. SVMs are an example of supervised learning algorithms, which means that in order to teach them, labelled data is necessary. The objective of SVMs is to identify a hyperplane that has a maximum distance from the closest data points in each class and separates the various classes of data with the greatest possible margin.

**2) KNN (K Neighbors):** Machine learning techniques like the K-Nearest Neighbors (KNN) method are frequently employed in classification and regression analysis. KNN is a non-parametric method, which means it doesn't assume anything about how the data are distributed. Instead, it locates the K data points in the training set that are closest to a new data point, and then labels the new data point with the plurality of the K nearest neighbors.

**3) Logistic Regression:** The primary objective of logistic regression is to determine the most appropriate S-shaped curve for the target variable The logistic regression model uses the maximum likelihood estimation approach to compute the sigmoid function's parameters. The model aims to identify the coefficient values that maximize the probability of the input features given the observed data. This optimization problem is resolved utilizing various optimization algorithms such as gradient descent. However, logistic regression presupposes that the input features are independent and linearly associated with the target variable, which may not always be valid in real-world applications.

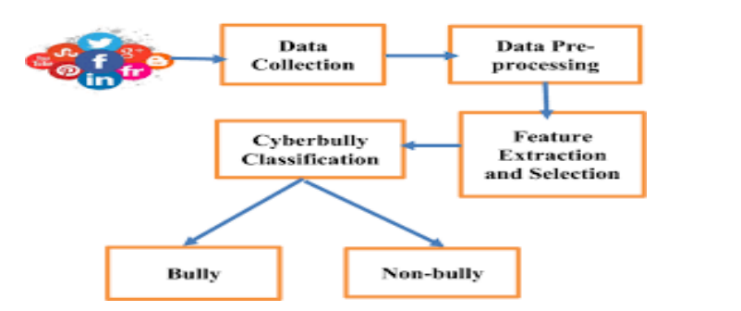
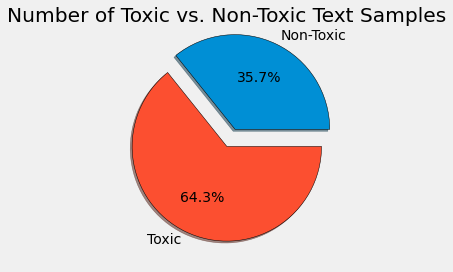
**4) Decision Tree Learning:** In the field of research, decision trees have proven to be a useful method for detecting instances of cyberbullying. This supervised learning technique involves categorizing data into multiple groups based on certain criteria. Decision trees can be applied in cyberbullying detection to analyze various forms of online content, including text, photos, and videos, and determine whether they constitute cyberbullying or not. In case the model's performance is not satisfactory, one can adjust its settings or redefine its characteristics to enhance its accuracy. 

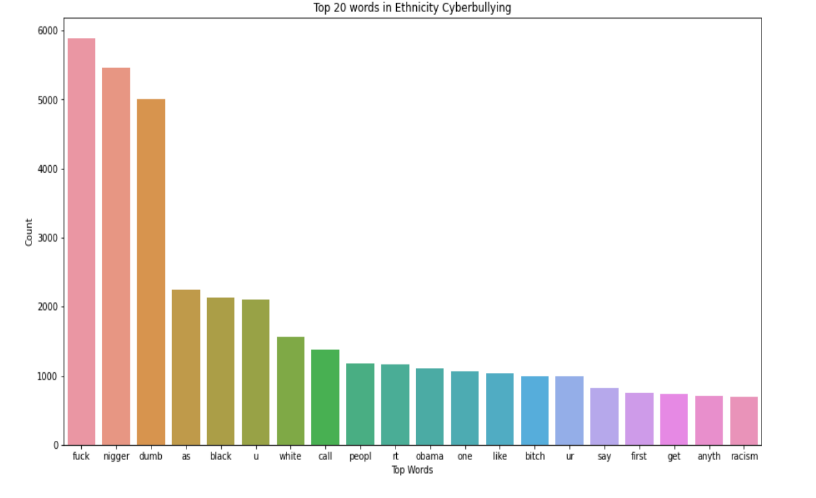
Fig.3.Model

1. *DATASET DESCRIPTION*

The Dataset is the snapshot of Twitter information during a particular period. It consists of tweets from all kinds of users. It has seven columns namely ‘Unnamed, ‘count', 'hate speech', 'offensive language ', 'neither', 'class', 'tweet '. The class is our dependent variable and the remaining are the independent variables. The class consists of numeric values, so we would convert them into discrete values by labeling them. The dataset has nearly 24000 rows. we would be partitioning it for testing and training. It does not have any missing values or inconsistent values. The tweets are the most important variable as we are going to use them for classifying the comments. The class as our target variable which will be predicted.



**:**

****

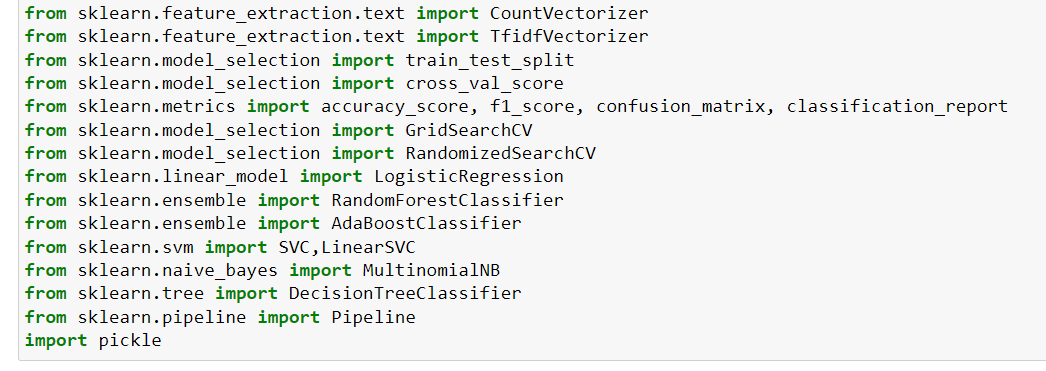
**Fig 4.Top 20 words used in cyberbullying tweets**

# 

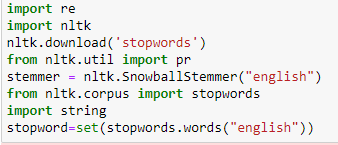
# Fig 5 Top 20 words used according to age

1. EXPERIMENTS AND RESULTS

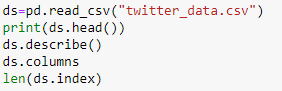
**IMPORTING LIBRARIES:**

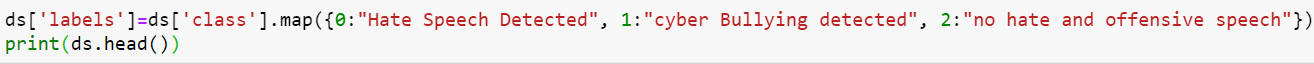
****

**IMPORTING AND DOWNLOADING NLP TOOLS :**

****

**LOADING DATASET :**

****

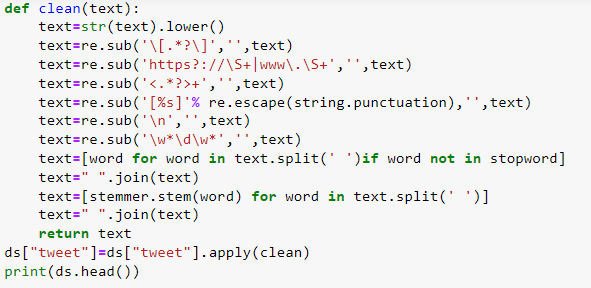
****

Here, we would be converting the target variable into a discrete variable as we would be working with decision tree classifier.



Here, we would be visualizing the changes we have made.

**USER DEFINED FUNCTIONS:**

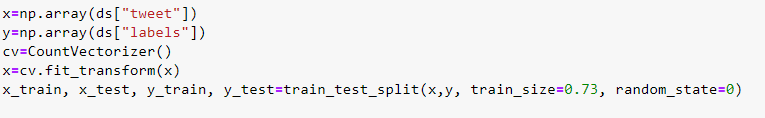
****

The inbuilt function clean is used to remove the unwanted gibberish or symbols from the tweets that are in the dataset.

**Function to remove emoji’s in text:**

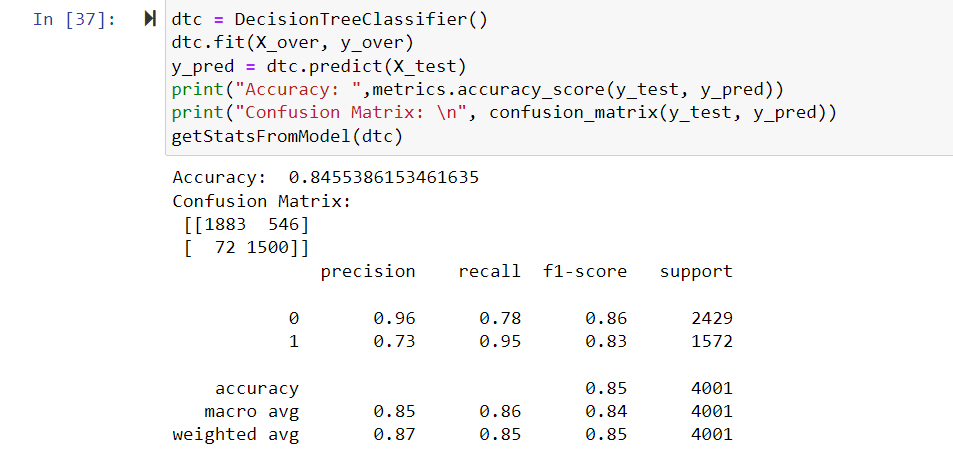
****

**SPLITTING THE DATASET**:

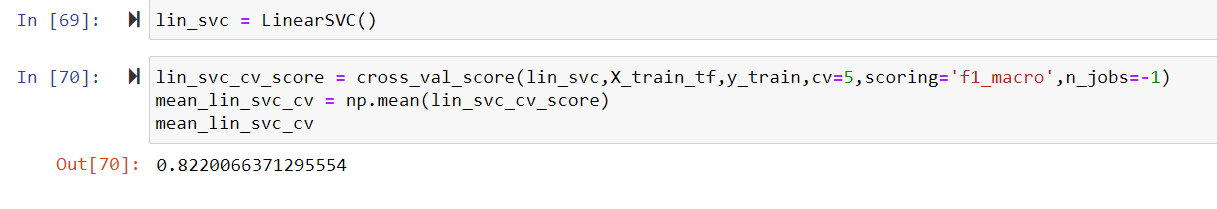


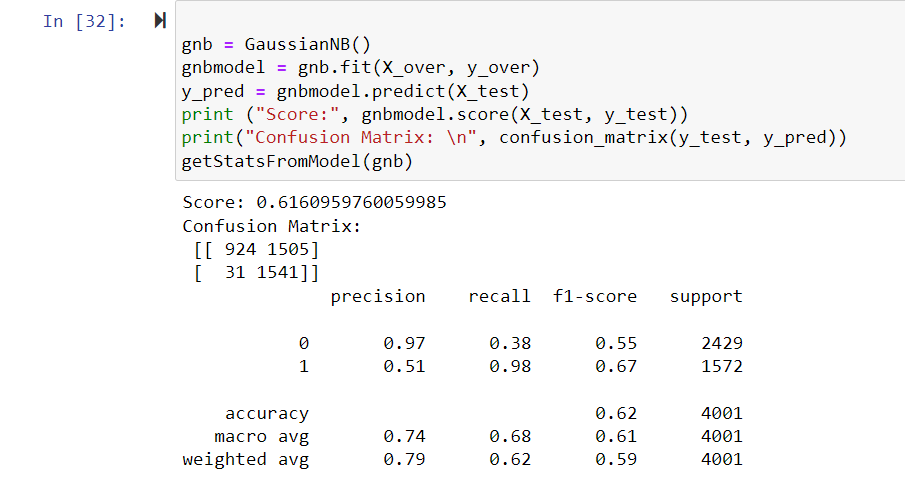
**CREATION AND TRAINING OF DIFFERENT MODELS:**

1. **Decision Tree Classifier**

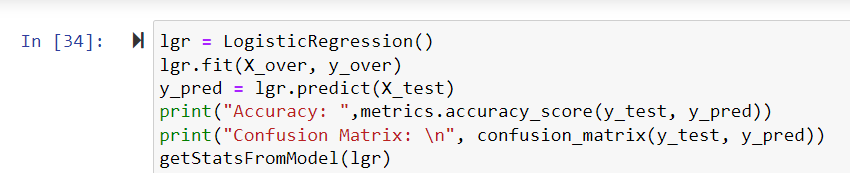
****

1. **Support Vector Machine**

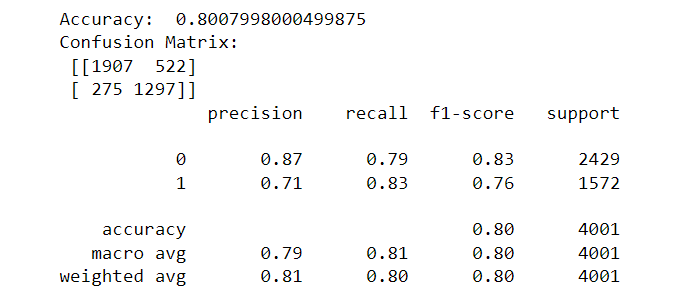
****

****

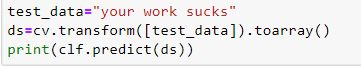
1. **Logistic Regression:**

****

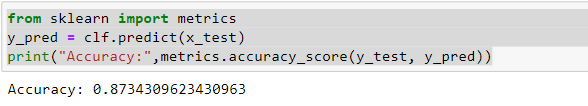
**Logistic Regression Confusion matrix:**

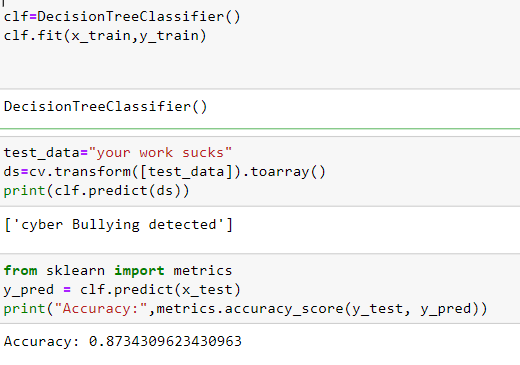
****

**TESTING THE MODEL :**

****

1. RESULTS





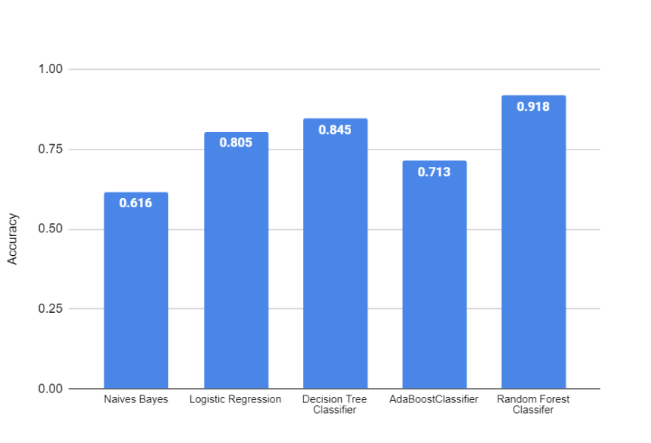
The four machine learning methods Support Vector Machines (SVM), Random Forest (RF), KNearest Neighbors (KNN), and Logistic Regression (LR) were used to compare the performance of the count vectorizer and term frequency-inverse document frequency and to select the best feature extraction model. In this paragraph, we outline:

Analysis of two feature extraction algorithms in comparison. According to the below plot, the two algorithms that provide the best accuracy are random forest classifier and logistic regression. Random Forest provides 90%, respectively.

Even though Random Forest has the highest accuracy (90%), training and output prediction require longer.

As a result, we were able to effectively extract the data, clean it, and visualize it using different Python libraries.

**Comparing different Machine Learning Techniques:**

****

**Fig 6 Comparing different Machine Learning Techniques**

1. CONCLUSION AND FUTURE WORK

* Long Execution time: Running all of the 450 experiments in the first set was challenging since it required a huge amount of time to complete. To overcome the slow execution time of the models, the experiments were conducted in parallel on high-performance compute nodes in Compute Canada clusters.
* Along with the laws that are used to punish those people who cause cyber violence having an system that automatically detects the context of cyberbullying and changes it into a positive comment will be of great help as the saying goes prevention is the best .Using this will prevent lot of people from depression ,low self-esteem and also suicides.it also never let the users to use the social media as a tool to humiliate or bully others .The cyberbullying detecting system will lead to healthy environment on social media .

1. REFERENCES
2. A. Weiler, M. Grossniklaus, M. H. Scholl, "An evaluation of the run-time and task-based performance of event detection techniques for Twitter," Information Systems, vol. 62, pp. 207 - 219, 2016
3. Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety.
4. Van Hee C, Verhoeven B, Lefever E, De Pauw G, Daelemans W, Hoste V. Guidelines for the Fine-Grained Analysis of Cyberbullying, version 1.0. LT3, Language and Translation Technology Team–Ghent University; 2015. LT3 15-01.
5. M. Pittman, B. Reich, "Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words," Computers in Human Behavior, vol. 62, pp. 155 - 167, 2016..
6. H. Watanabe, M. Bouazizi, and T. Ohtsuki, “Hate speech on twitter:A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection,” IEEE access, vol. 6, pp. 13 825–13 835,2018.
7. L. Ketsbaia, B. Issac, and X. Chen, “Detection of hate tweets using machine learning and deep learning,” in 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, 2020, pp. 751–758..
8. Gahagan, Kassandra, J. Mitchell Vitellius, and Libby R. Frost. “College student cyberbullying on social networking sites: Conceptualization, prevalence, and perceived bystander responsibility.” Computers in human behavior 55 (2016): 1097-1105..
9. S. Hinduja and J. Patchin, “Bullying, Cyberbullying, and Suicide”, Archives of Suicide Research, vol.14, no. 3, pp. 206-221, 2010.
10. V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, “Detection of cyberbullying using deep neural network,” in 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS).IEEE, 2019, pp.
11. J Yadav, D Kumar, D Chauhan Cyberbullying detection using pre-trained bert model 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), p. 1096 – 1100 Posted: 2020.
12. M. Dadvar, D. Trieschnigg, R. Ordelman, F. Jong, “Improving Cyberbullying Detection with User Context,” Advances in Information Retrieval, vol. 78, pp. 693 – 696, 2013.
13. Singh, A., Rajput, N., & Kumar, V. (2020). Cyberbullying detection using machine learning techniques: A review. Journal of Ambient Intelligence and Humanized Computing, 11(10), 4583-4603
14. Mishra, S., Kumari, S., & Bhagat, S. (2020). Cyberbullying detection using machine learning: A review. Journal of Intelligent & Fuzzy Systems, 39(3), 4185-4200.
15. Carrasco, R. C., Plaza, L., & Ramos-Garijo, J. (2020). Cyberbullying detection using machine learning: A systematic review. IEEE Access, 8, 15015-15030.
16. Basile, A., Caputo, A., Semeraro, G., & Basile, V. (2019). Deep learning for cyberbullying detection in social media texts. Information Processing & Management, 56(3), 1009-1022