*Section: Research Paper*

# Bullying Tweet Classification using Naive Bayes, Bi-LSTM and Bert for Disease Classification

**Mr.PRAKHAR NAGPAL[1], SRMIST, KATTANKULATHUR – 603203.**
M.SALOMI[*], DEPARTMENT OF COMPUTATIONAL INTELLIGENCE, SRM INSTITUTE OF SCIENCE AND TECHNOLOGY, FACULTY OF ENGINEERING AND TECHNOLOGY, KATTANKULATHUR – 603203, CHENGALPATTU DISTRICT, TAMIL NADU, INDIA.
Ms.UNNATI MOSHRA[2], SRMIST, KATTANKULATHUR – 603203.
Mr.DEVENDRA BANSAL[3], SRMIST, KATTANKULATHUR – 603203.
Mr.VARUN SUNIL PATHAK[4], SRMIST, KATTANKULATHUR – 603203.
Mr.SAYAN MOHATTY BHAVISHYA DHIMAN, SRMIST, KATTANKULATHUR – 603203

*Abstract*— **The advantages of social networking sites are obvious, and they have given us more possibilities than ever before. Despite the positive effects, peers, outsiders, and anonymous users frequently humiliate, abuse, bully, and torment people. The term "cyberbullying" describes the use of technology to degrade and disparage others. Hateful letters and social media posts are examples of this. Cyberbullying has surfaced as a form of bullying through social media due to the exponential rise in social media users tweets as well. The problem of harassment has been brought up by people's use of Twitter, particularly teens and young people. Lack of research exists on the characteristics influencing the spread of tweets about cyberbullying as well as the kinds of guidance and support offered in tweets for victims. 7,315 tweets related to cyberbullying were collected and examined for this research. The findings showed that tweets with specific characteristics, such as more URLs, keywords, or friends, did not always result in more retweets. The study of the tweets' emotions revealed users' attitudes towards cyberbullying to be mixed. 400 tweets were carefully selected for this study's content analysis. Tweets on harassment spanned a range of topics, from user views to current events. According to the findings, 33% of tweets included encouragement and guidance for cyberbullying sufferers. Compared to tweets discussing other aspects of harassment, these tweets received the most shares. By using ML Classification algorithms like Naive Bayes, Bi-LSTM, and BERT Classification, our research seeks to identify cyberbullying in Twitter as a potential answer to the aforementioned issue. Also, this model uses Word2Vec a technique for Natural Language Processing (NLP) for word embedding.Also,the diseases can be classified using Bert algorithm.**

**Keywords:user,embedding,tweets,algorithms,networking,benefits,classification,Bi-LSTM,Naïve bayes**

## I. INTRODUCTION

Cyberbullying has become more prevalent, particularly among teenagers, on account of people using social networking sites (SNS) like Facebook or Twitter to interact with one another and the rest of the world. Researchers have published numerous studies addressing the problems related to teens and harassment.

Cyberbullying is the intentional, persistent, and hostile use of ICT (information and communication technology) to torment and damage. (Stopbullying.gov, 2014). Social media bullying, harassment, cyberstalking, deceit, impersonation, and delivering hurtful messages through chat groups and instant messaging are examples of cyberbullying types. More than 10,008 adolescents were polled by Ditch the Label (2014), which found that 28% of the 43% who used Twitter had encountered harassment.

Social scientists have investigated a variety of topics in the study of cyberbullying involving teenagers, including

Cyberbullying risk factors, the personalities of actors (victims, bullies, and bystanders) involved in cyberbullying. Comparatively, computer scientists have concentrated on creating automatic methods for detecting cyberbullying using data and text extraction. Characterizing bullying-related hashtags, identifying cyberbullying messages through feature selection and analysing the linguistic uniqueness of Indonesian bullying phrases on Twitter identify Indonesian bullying patterns are some research studies that are specifically focused on Twitter.

Tarapdar and Kellett (2011) highlighted that additional to identifying cases of online abuse online, adolescent victims of cyberbullying seek support from a range of individuals, including peers, parents, instructors, the authorities, peer leaders, and social network businesses. In their 2013 article, Notar, Padgett, and Roden stressed the value of collaborating with parents to resolve harassment situations. Giving parents information about cyberbullying can help them better comprehend adolescents who are being victimized. Children and teenagers need to be urged to report cases of harassment. According to Holfeld and Grabe's (2012) research of 665 US middle school students' encounters with cyberbullying, 64% of the 333 students who had been bullied online decided to report it. The majority of students (64%) revealed this to their peers, followed by their parents (50%) and then their siblings (20%), instructors (8%) and relatives (5%). Despite having a low rating, 80% of students said that once they were made aware of cyberbullying, instructors attempted to halt it.

This article was inspired by the need for more research into the ways in which tweets on Twitter are used to spread information about cyberbullying, particularly the guidance and support related to combating it. This guidance can offer a network of support for sufferers of cyberbullying of all ages, including many teens. The following areas of additional study will be

looked into:
1.Automating the extraction of tweets related to intimidation in order to investigate the relationship between the characteristics of the tweets and their dissemination through Twitter;
2.Describing the kinds of information about cyberbullying that was circulated in the excerpted tweets and examining the tweets for proof of the help and guidance offered to victims of cyberbullying

## II. PROBLEM STATEMENT

The social media network offers us fantastic contact platforms, but it also makes young people more susceptible to danger online. Because The number of users on social media networks is so high, cyberbullying on them is a worldwide problem. A daily rise in harassment on social media is seen in the trend. Cyberbullying is a growing problem for youth, according to recent study. Due to the information inundation on the Web, intelligent systems are needed to automatically identify potential dangers. Effective avoidance relies on the accurate identification of potentially damaging communications. Therefore, the goal of this project is to create a model for the automated identification of cyberbullying in tweets by abusers.

## III. LITERATURE SURVEY

This section addresses the issues with cyberbullying and reviews some earlier studies done in the discovery of cyberbullying.

| Title | Problem | Solution | Result |
|---|---|---|---|
| An Effective Approach for Cyberbullying Detection and avoidance | The biggest problem regarding cyberbullying is that the age group of the offenders range from as young as eight to the legal adult age of eighteen and beyond. Once happen this activity then victims are often left permanently then difficult to find them. | In this paper focused on the issues of robust system and objectives are 1) Automatic detection and avoidance of cyberbully attack in internet. 2) Effective age authentication for website browsing and categorizing the links based on age | Represented a novel method on the current scenario of cyber-bullying and various methods available for the detection and prevention of cyber harassment. Our concept depends upon the text analysis, the data which is uploaded or text written by any user is first analyzed. |
| Using Machine Learning to Detect Cyberbullying | Teens and young adults, are finding new ways to bully one another over the Internet. in a study conducted by Symantec reported that, to their knowledge, their child has been involved in a cyberbullying incident | Used machine learning algorithm to detect cyberbullying. For training the data downloaded from website. The data was labeled using a web service. the labeled data, in conjunction with machine learning techniques provided by the Weka tool kit, to train a computer to recognize bullying content | Used a language-based method of detecting cyberbullying. By recording the percentage of curse and insult words within a post. |
| Cyberbullying Detection System on Twitter | Increased cyberbullying attacks on the social network services. To prevent these activities a system was proposed. | In this system, the users can identify the cyberbullying related tweets based on the keywords and populate it in a news feed form. By doing this, it allows users to determine the identities of the cyberbullies and the victims from the cyberbullying tweets | With the advent of this cyberbullying detection and solution system on Twitter, it will help the authorities to monitor, regulate or at least decrease the harassing incidents in cyberspace |

*Figure 1 Literature Survey*

The article, according to the Literature Review [1], concentrated on the problems of a robust system and goals like automated cyberbully attack detection and prevention on the internet, as well as effective age verification on websites for viewing and categorizing the links based on age. The study's findings demonstrated a new approach to the problem of cyberbullying in the present day and to the different techniques for identifying and avoiding online harassment. Their conceptualization is also reliant on text analysis, which examines submitted data as well as user-written text.

They employed a machine learning programme to find harassment, according to Literature Review [2]. Data that was obtained from a website and used for training was marked using a web application. the labelled data in combination with machine learning methods made available by the Weka tool set, to teach a computer to recognize bullying-related material. According to the findings of this study, they used a language-based technique to identify cyberbullying by keeping track of the frequency of derogatory and offensive terms in posts.

The users can use the terms to find tweets linked to cyberbullying and populate them in a news stream form, claims study paper [3]. By doing this, it enables users to identify the victims and perpetrators of harassment from the tweets. The finding indicates that the establishment of this method for detecting and resolving cyberbullying on Twitter will assist the law enforcement in keeping track of, regulating, or at the very least reducing cyberbullying occurrences

## IV.SYSTEM ARCHITECURE

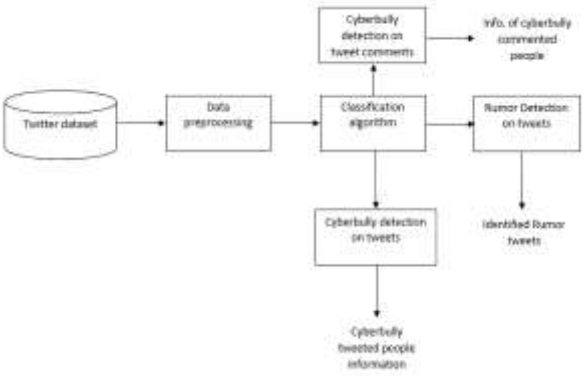


*Figure 2 Architecture Diagram*

Firstly, a twitter dataset is obtained from a public dataset site such as Kaggle, Google Dataset Site etc.
After that various preprocessing is done through the dataset such as in this module, we are removing all punctuations, links, stop words, mentions, and new line characters. We are removing Contractions such as can't, won't and replacing them with their complete forms and also special characters like '&'.Also we are removing the hashtags from the end of the sentences and keeping only those which are in the middle of the sentence
After data cleaning and data pre-processing, carious data visualizations and analysis is done using Seaborn Libraries.
Two Graphs are created in this which depict the Count of Tweets less than 10 words and Count of tweets with high

number of words.

The first algorithm we implement is Naive Bayes, which is used as a simple baseline model. We must preprocess the text input first before using this method. First, we use Count Vectorizer to produce a word bundle.

Then we generate a classification report for Naïve Bayes and a confusion matrix

The algorithm's achievement ratings are excellent, with a total accuracy of 84%. We can see that while the predictions for the more crowded courses have very high F1 scores (over 84%), the result is much lower for the class "non-cyberbullying."

Next we implement a more complex algorithm to perform the classification, aiming to achieve higher accuracy than the baseline Naive Bayes model.

The Word Embedding Matrix is created here. utilizing the pre-trained model Word2vec and the original text tweets. Initially, a list of words from the previously created X_train vector is created.

The number of features in each transformed word is the dimension of the embedding words that we specify.

We Determined an Embedding Word Dimension, Which May Be Interpreted As The Quantity Of Features Of Each Transformed Word.

We Must Decide on The Maximum Number of Words Before Defining The Embedding Matrix. We will extract the specified number of words from the previously developed Python dictionary for vocabulary. Finally, The Embedding Matrix Can Be defined.

Now we will define a custom training loop, where we include an early stopping functionality, and save only the best models in terms of validation accuracy.

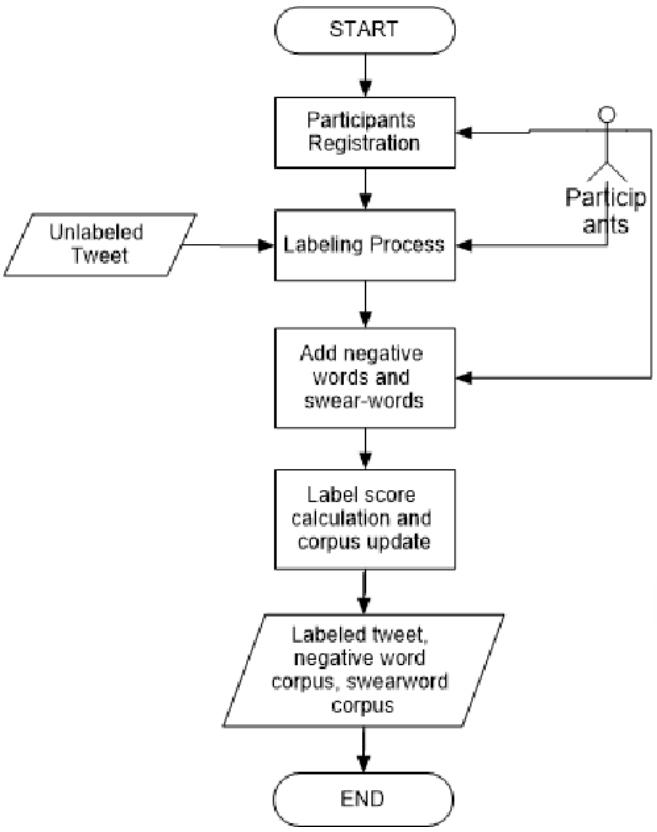


*Figure 3 System Flow Diagram*

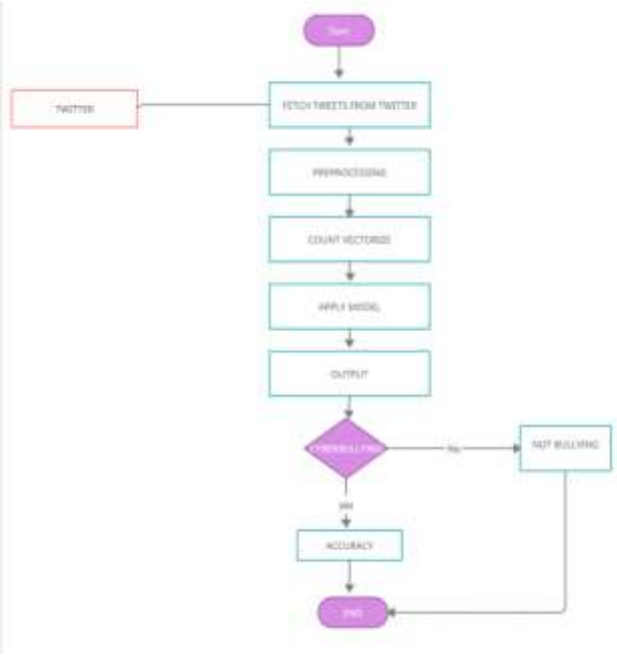

*Figure 4 Flow Chart*

## V. METHOLDOLGIES/ALGORITHMS IMPLEMETED

### Naïve Bayes Theorem

As a straightforward default model, Naive Bayes is the first method we employ. We must first preprocess the text input in order to use this method. First, we use Count Vectorizer to produce a word bundle.

Next, we create a confusion matrix and a categorization report for Naive Bayes.

The algorithm's achievement ratings are excellent, with a total accuracy of 84%. We can see that while the predictions for the more crowded courses have very high F1 scores (over 84%), the result is much lower for the class "non-cyberbullying."

It is a Bayesian supervised learning approach for classification problems. With a large training set, it is mostly used in text classification.

One of the most easy and efficient classification methods available today is the Naive Bayes Classifier. It contributes to the quick development of machine learning models capable of producing accurate forecasts.

It makes forecasts as a probabilistic predictor based on the assumption that an item will happen.

Algorithms like Naïve Bayes are widely utilized in Spam detection, article analysis and sentiment analysis categorization.

It is referred regarded as naive since it implies that the appearance of one characteristic is unrelated to the appearance of other traits.

### Word Embedding Using Word2Vec

Word embedding is one of the most used methods for describing text vocabulary. It may determine a word's

location in a document, its semantic and syntactic similarity, its link to other terms, and so on.
What precisely are word embeddings? They can be thought of as loose vector depictions of specific words. Having said that, how do we produce them is what follows. How do they catch the background, more importantly?
Word2Vec is a popular approach for learning word embeddings using shallow neural networks.
Consider the following equivalent phrases: I wish you a good day. They barely transmit any contradictory notions. If we were to invent a phrase, it would be "Have a nice, fantastic day," V.

**Bi-LSTM(Bi-Directional Long Short-term memory)**

Bidirectional recurrent neural networks (RNN) are essentially the combination of two distinct RNNs. This structure allows the networks access to both forward and backward information about the series at each time point.
The inputs we provide while utilizing bidirectional will be handled in two separate ways: One goes from the present to the future, while the other goes from the future to the present. By combining the two concealed states, we may keep data from both the present and the future at any given time. which distinguishes this approach from unidirectional methods in that it preserves future data in the LSTM that runs backward.

**Bert Classification and Modelling**

BERT, or Bidirectional Encoder Representations from Transformers, is an abbreviation. Several hints about what BERT is about are provided by the word alone.
Transformer encoders are layered one on top of the other to create the BERT architecture. Each Transformer encoder contains two sub-layers: a feed-forward layer and a self-attention layer.
There are two distinct BERT designs:
BERT base, a BERT model, is made up of 110M parameters, 768 concealed sizes, 12 attention heads, and 12 levels of Transformer encoders.
BERT large, a BERT model, has 340 parameters and has 24 levels of Transformer encoders, 16 attention centers, and 1024 concealed sizes.

## VI. MODULES AND THEIR IMPLEMENTATIONS

### MODULE 1-DATA PRE-PROCESSING AND CLEANING

a. In this module we are removing all punctuations, links, stop words, mentions, and new line characters.
b. We are removing Contractions such as can't, won't and replacing them with their complete forms and also special characters like '&'.

Also, we are removing the hashtags from the end of the sentences and keeping only those which are in the middle of the sentence.

### MODULE 2- TWEET LENGTH ANALYSIS USING SEABORN LIBRARY

i. After data cleaning and data pre-processing, carious data visualizations and analysis is done using Seaborn Libraries.
ii. Two Graphs are created in this which depict the Count of Tweets less than 10 words and Count of tweets with high number of words.

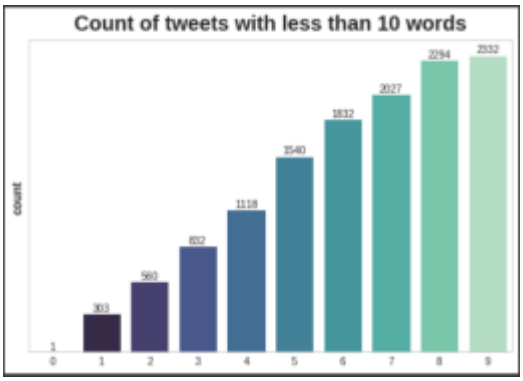


*Figure 5. Count of Tweets Less than 10 Words*

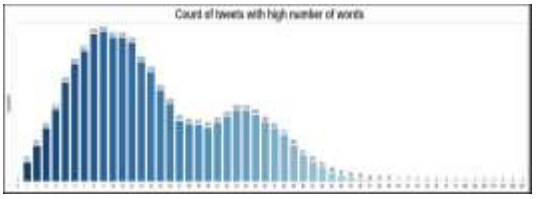


*Figure 6. Count of Tweets with high number of Words*

### MODULE 3-CREATE A BASELINE USING NAÏVE BAYES THEOREM

A. The first algorithm we implement is Naive Bayes, which is used as a simple baseline model. We must preprocess the text data beforehand before using this approach. Count Vectorizer is used to first construct a bag of words.
B. Then we generate a classification report for Naïve Bayes and a confusion matrix
C. The algorithm's execution scores are good, with an overall accuracy of 84%. We can observe that, although the predictions for the more populated classes have exceptionally high F1 values (over 84%), the score for the class "non-cyberbullying" is significantly lower.

Classification Report for Naive Bayes:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| religion | 0.82 | 0.96 | 0.89 | 1589 |
| age | 0.79 | 0.97 | 0.87 | 1577 |
| ethnicity | 0.88 | 0.91 | 0.89 | 1548 |
| gender | 0.86 | 0.84 | 0.85 | 1521 |
| not bullying | 0.85 | 0.49 | 0.62 | 1527 |
| accuracy | | | 0.84 | 7762 |
| macro avg | 0.84 | 0.83 | 0.82 | 7762 |
| weighted avg | 0.84 | 0.84 | 0.83 | 7762 |

*Figure 7. Classification Report for Naïve Bayes*

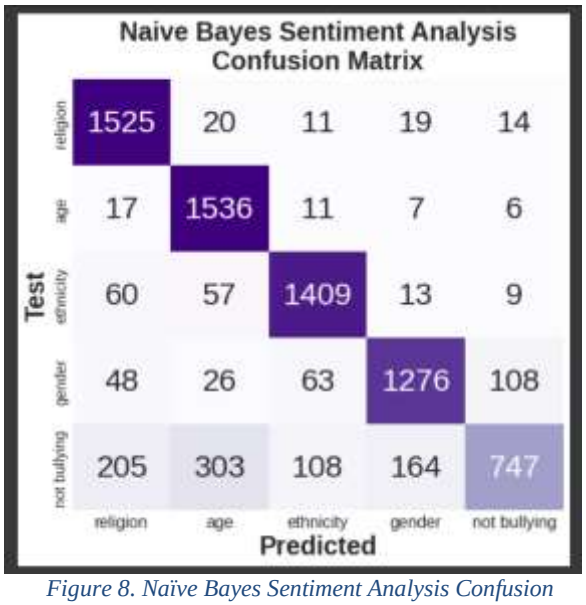


*Figure 8. Naïve Bayes Sentiment Analysis Confusion Matrix*

**MODULE 4-WORD EMBEDDING USING WORD2VEC**

A. The Word Embedding Matrix is created here. use the pre-trained model Word2vec and the original text tweets. The X_train Vector is first used to create a list of words.

B. We defined an Embedding Word Dimension, which may be interpreted as the number of features in each transformed word.

C. We Determine The Dimension Of The Embedding Words, Which Is The Number Of Features Of Each Transformed Word.

D. We also need to decide on the maximum number of words before defining the embedding matrix. We'll take the word count from the previously created Python dictionary for vocabulary. We can now define the embedding matrix, at last.

**MODULE 5- Py-Torch Bi-LSTM Modelling and LSTM Training Loop**

Now we will define a custom training loop, where we include an early stopping functionality, and save only the best models in terms of validation accuracy.

Classification Report for Bi-LSTM :

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| religion | 0.97 | 0.94 | 0.95 | 1589 |
| age | 0.96 | 0.98 | 0.97 | 1577 |
| ethnicity | 0.98 | 0.98 | 0.98 | 1548 |
| gender | 0.94 | 0.87 | 0.90 | 1521 |
| not bullying | 0.81 | 0.88 | 0.84 | 1527 |
| accuracy | | | 0.93 | 7762 |
| macro avg | 0.93 | 0.93 | 0.93 | 7762 |
| weighted avg | 0.93 | 0.93 | 0.93 | 7762 |

*Figure 9. Classification Report of Py-Torch Bi-LSTM Modelling*

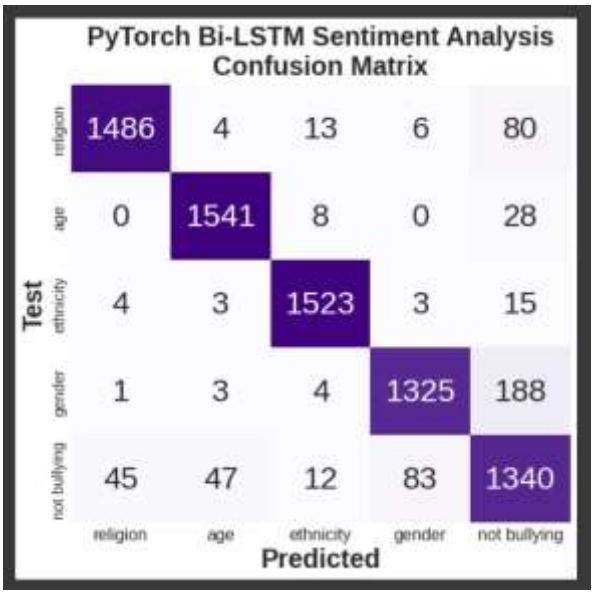


*Figure 10. Confusion Matrix of Py-Torch Bi-LSTM Modelling*

**MODULE 5- BERT CLASSIFICATION**

In this section, we will load a pre trained BERT model from the Hugging Face library and fine tune it for our classification task.

First, we split the dataset into train - validation again since we need to tokenize the sentences differently from before (Naive Bayes and LSTM).Since we need to tokenize the tweets (get "input ids" and "attention masks") for BERT, we load the specific BERT tokenizer from the Hugging Face library. Since we need to specify the length of the longest tokenized sentence, we tokenize the train tweets using the "encode" method of the original BERT tokenizer and check the longest sentence. With an overall accuracy of about 94% and F1 scores well over 95%, BERT Classifier's performance scores are fairly good and higher than those obtained using the LSTM model.

Classification Report for BERT :

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| religion | 0.94 | 0.96 | 0.95 | 1589 |
| age | 0.98 | 0.98 | 0.98 | 1577 |
| ethnicity | 0.99 | 0.99 | 0.99 | 1548 |
| gender | 0.88 | 0.91 | 0.90 | 1521 |
| not bullying | 0.86 | 0.81 | 0.83 | 1527 |
| accuracy | | | 0.93 | 7762 |
| macro avg | 0.93 | 0.93 | 0.93 | 7762 |
| weighted avg | 0.93 | 0.93 | 0.93 | 7762 |

*Figure 11. Classification Report of Bert Classification*

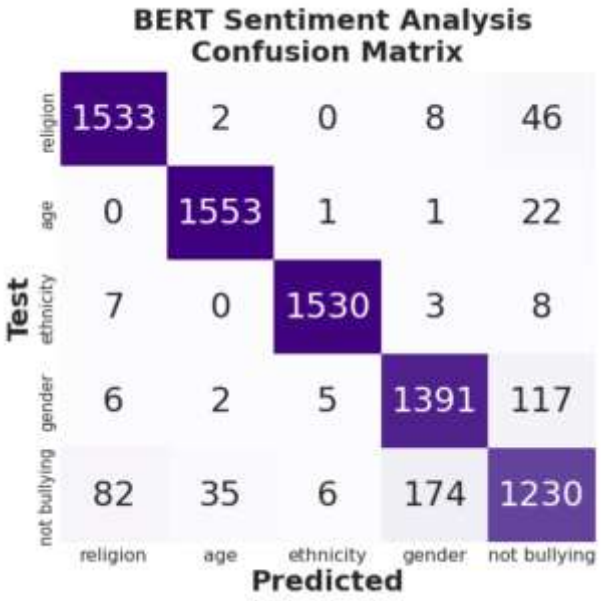


*Figure 12. Confusion Matrixs of Bert Classification*

## VII. RESULTS AND DISCUSSIONS

With an overall accuracy of 84%, the Naive Bayes model performs admirably. We can see that while the predictions for the more crowded courses have very high F1 scores (over 84%), the result is much lower for the class "non-cyberbullying."

With an overall accuracy of 93%, the Bi-LSTM has excellent performance ratings. while the grade is 84% for the class "non-cyberbullying."

With an overall accuracy of about 94% and F1 scores over 95%, the performance scores of the BERT Classifier are fairly good and greater than those obtained using the LSTM model.

## VIII. CONCLUSION AND FUTURE WORK

The results of our poll show that the huge volumes of data created by Web 4.0 cannot be handled by the machine learning algorithms in use today, making it challenging to identify cyberbullying content.

Many researchers have lately been interested in deep learning techniques including deep recurrent neural networks, convolutional neural networks, and stacking auto-encoder.

Future study may focus on using these deep learning methods to precisely identify cyberbullying in social media. Our algorithm for detecting cyberbullying depends on binary classification (bullying or non-bullying), therefore a multi-class classification approach may also be used in future research.

## IX. REFERENCES

[1]. M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, "Cyberbullying detection on social networks using machine learning approaches," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020.

[2]. A. Muneer and S. M. Fati, "A comparative analysis of Machine Learning Techniques for cyberbullying detection on Twitter," Future Internet, vol. 12, no. 11, p. 187, 2020.

[3]. S. Bastiaensens, H. Vandebosch, K. Poels, K. Van Cleemput, A. DeSmet, and I. De Bourdeaudhuij, "Cyberbullying on social network sites. an experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully," Computers in Human Behavior, vol. 31, pp. 259–271, 2014.

[4]. S. Kemp, "Digital 2019: Global Digital Overview - DataReportal – global digital insights," DataReportal, 13- Apr-2019.[Online].Available:https://datareportal.com/reports/digi tal- 2019-global-digital-overview. [Accessed: 16-May2022].

[5]. R. Pawar and R. R. Raje, "Multilingual cyberbullying detection system," 2019 IEEE International Conference on Electro Information Technology (EIT), 2019.

[6]. V. Jain, V. Kumar, V. Pal, and D. K. Vishwakarma, Detection of cyberbullying on social media using machine learning," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC),2021.

[7]. B. Haidar, M. Chamoun, and A. Serhrouchni, "Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content," 2017 1st Cyber Security in Networking Conference (CSNet), 2017.

[8]. Alduailej, A. H., & Khan, M. B. (2017, September). The challenge of cyberbullying and its automatic detection in Arabic text. In 2017 International Conference on Computer and Applications (ICCA) (pp. 389-394). IEEE.

[9]. Kargutkar, S., & Chitre, V. Implementation of Cyberbullying Detection using Machine Learning Techniques.

[10]. Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., ... & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. PloS one, 13(10), e0203794.

[11]. Çiğdem, A. C. I., Çürük, E., & Eşsiz, E. S. (2019). Automatic detection of cyberbullying in Formspring. me, Myspace and Youtube social networks. Turkish Journal of Engineering, 3(4), 168-178.

[12]. Al-Ajlan, M. A., & Ykhlef, M. (2018). Deep learning algorithm for cyberbullying detection. International

Journal of Advanced Computer Science and Applications, 9(9), 199-

[13]. Volume 7, Issue 6, June – 2022 International Journal of Innovative Science and Research Technology ISSN No:- 2456-2165 IJISRT22JUN1040 www.ijisrt.com 262

[14]. H. Nurrahmi and D. Nurjanah, "Indonesian twitter cyberbullying detection using text classification and user credibility," 2018 International Conference on Information and Communications Technology (ICOIACT), 2018.

[15]. S. A. Ozel, E. Sarac, S. Akdemir, and H. Aksu, "Detection of cyberbullying on social media messages in Turkish," 2017 International Conference on Computer Science and Engineering (UBMK), 2017.

[16]. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2020). Deep Learning based Recommender System. ACM Computing Surveys, 52(1), 1–38.