Machine Learning base Cyberbullying Classification Using Tweets Data

Table of Contents

Table of Figures

Table of Tables

Introduction

1. Introduction

Digital media has enabled numerous movements around the globe in the last ten years. It is a highly impactful technology of the time and a terrific method to broaden somebody's horizons in terms of events and cultural engagement. Social networking is a dual-edged sword, though. On digital media, it's common to see unwelcome conduct like online stalking, internet bullying, and online harassment. These types of harassment, abuse, and surveillance are now accepted aspects of coming of age. Furthermore, anyone can become a victim; it is not just a problem for kids and teenagers.

Thanks to considerable advancements in Internet technologies (*The History of Social Media and Its Impact on Business - ProQuest*, n.d.), digital networking websites like Twitter or Facebook have gained popularity and have a big impact on how people live their lives. Digital media platforms in particular have blended everyday activities including commerce, schooling, amusement, including e-government into everyday life. By 2025, there are expected to be more than 3.62 billion monthly dynamic digital media users worldwide (Bauman, 2015). One-third of all people on planet Earth will be represented by these numbers. For some researchers, Twitter is the crucial network and the crucial data collection source among many currently available digital networks. Twitter is the well-known, real-time, and free, publicly available blogging platform where information frequently appears before it does in authoritative sources. Approximately per day tweets by the tweeter is 550 million, Twitter use has quickly grown, especially across all occurrences (Pereira-Kohatsu et al., 2019). Its quick text restriction (currently 280 letters) and uncensored stream are its defining features. At the moment, digital media is a necessary part of everyday life. Unquestionably, there are numerous behavioral or psychological hazards associated with some young people using electronics, especially digital media. Online bullying, the pervasive social assault that takes place on digital media networks, is one of the threats.

Additionally, online bullying has been linked to negative impacts human on mental health conditions, such as digital and emotional issues thinking of suicide, the attempt of suicide, melancholy, anxiousness, and different other forms of self-harming (*Cyberbullying: Experiences, Impacts and Coping Strategies as Described by Australian Young People | Youth Studies Australia*, n.d.; Miller, 2016; Smith, 2012).

Additionally, a sharp rise in reported occurrences of online bullying has brought attention to the risk of this behavior, specifically among kids and teenagers who can be rude and immature. Kids and teenagers take bullying very seriously, but they lack the digital skills to deal with it as well, which makes them vent their feelings in hurtful ways on digital media. According to (Sampasa-Kanyinga et al., 2014), numerous researchers have demonstrated how extremists frequently experience psychological issues, which cause individuals to harass and cause misery to everyone. In light of this, cyberbullying is comparable to the epidemic and has the potential to create an aggressive community, especially among high technology college and secondary school pupils.

Additionally, a sharp rise in internet bullying instances has brought attention to the risk of internet bullying, specifically among kids and teenagers who can be disrespectful and immature. Since they lack the digital skills to deal with digital difficulties that take harassment very seriously, kids and teenagers often vent their feelings in hurtful ways on digital media. Numerous research investigations have revealed, according to (*Automated Hate Speech Detection and the Problem of Offensive Language | Proceedings of the International AAAI Conference on Web and Social Media*, n.d.), as abusers frequently experience psychological issues, which cause individuals to harass and cause misery to others. Online bullying is therefore akin to the epidemic and can contribute to the aggressive environment, especially with respect to the high-tech development of school kids.

Conclusively, large number of international efforts are suggested to deals with this issue of online bullying. Such initiatives focus on improving safety measures for online consumers, in supporting kids. As example, A reputed university institute of turku which is located in Finland developed the safety program (KIVA) related to online bullying (Vaillancourt et al., 2017), French governmental authorities launched a campaign against abusive behavior [10], and Belgian governmental authorities launched the anti-cyberbully action (Görzig et al., 2013). Cyberbullying detection and filtration are thought to be supplementary to regulation with intervention strategies, despite Internet's huge and challenging-to-control material. As a result, it is critical to identify online bullying on digital media and pay particular attention to doing so in order to safeguard youngsters or communities from its negative impacts. With the goal of detecting, reducing, and controlling online bullying through digital media, online bullying is becoming a research issue. By examining insulting language depending on several characteristics, including the structure and distinctive material, moreover INTERNET consumer’s way of write something, additional goal in the current field is to determine the user's purpose in uploading offensive content. The identification of textual content employing deep learning for inappropriate and abusive language identification and categorization is also another area of online bullying investigation.

1.1. Online bullying

"Deliberate and repetitive damage done via computers, smartphones, and some other digital devices" is an official definition of online bullying. In other words, abusers frequently employ internet-based communication to torment victims. Rage, disappointment, retaliation, or the simple wish to dominate other people and feel more dominant may cause these kinds of activities like harassment. To work for the less self confidence for blend it along with the friends, children will find occasionally internet bully strangers. Allegations received via e-mail or uploaded via digital-media, shameful images and some videos, and threatening, disrespectful, and tormenting messages published on digital networks are all illustrations of online bullying. It is incredibly challenging to remove similar offensive things from digital media platforms once they have been made. The strikers can do online attack anytime in a day or night or may be any day in a week, when the person is all alone, out there in the school playground field (*Survey: Majority of Cyber Bullying Incidents Go Unreported | TopNews,* n.d.). With internet bullying, the bully is given the ability to ridicule and harm the target in social media groups while going unnoticed. Moreover, complainants are discouraged from investigating rape out of concern for repercussions or for becoming social outcasts. It becomes challenging to manage this issue.

Online bullying is not only unpleasant, but it can also have disastrous results sometimes. The publication Psychosocial Investigation on the Internet Space conducted important studies, which revealed the "critical effects" that occurred in most of the respondents' situations in the form of less confidence level, hopelessness, discontentment, as well as suspiciousness about various peoples: A much excessive implication was harming himself or aggressive behaviors with the family and personal friends. Additionally, This shows that there are different victims created "faring mechanisms." Internet bullying victims frequently attempt to handle the matter on their own, which can be distressing. It is also challenging for victims' families to monitor what the youngster is doing online.

Social support must spot internet bullying or early warning symptoms of it is intended to facilitate the victim. They shouldn't anticipate hearing from the victim regarding online bullying. This necessitates the creation of autonomous internet bullying monitoring tools which might warn relatives about online bullying.

1.2. The use of digital media as a defense

Digital networks, therefore, provide a safe internet experience to a certain extent. The following are some technologies that can aid with privacy protection:

- Enabling the users to unsubscribe, blocked, or ban bothersome friends.
- Twitter offers the following services to consumers.
- Internet users can apply different filters on alerts to remove any unwelcome comments or comments from such accounts they are not following.
- Tweeting about the inappropriate behavior.
- Whenever displaying sensitive information, give a warning to the user about this. It merely functions with pictures including videos.
- By setting some privacy options for their tagging, users can control people tagging or cannot tag them in pictures.
- There are just 140 letters on Twitter.

The following options are available on Facebook:

- Customers are assured by Facebook that it "eliminates cyberbullying material when we would become informed and can disable Facebook account anybody who abuses or assaults someone."
- Options to blacklist or unfollow the bothersome individual.
- Facebook's data privacy options let individuals choose the viewership for the posts. Additionally, it offers individuals tagging confidentiality. Furthermore, individuals can examine posts including tags prior to sharing them on their timelines.
- There is an active link at the lower end of every post which enables users to report offensive material. Use this link to do so.
- The story which shows in the user's Fb Newsfeed can be hidden.
- Digital media platforms do not have any internal systems for identifying online bullying.

Only incidents that are documented promptly them to take action. Therefore, it is really the responsibility of victims or onlookers to investigate online bullying.

1.3. Autonomous Detection

Despite the fact that modern digital networking platforms offer mechanisms as well as some strict rules, and they are regulating these things smoothly to combat online bullying, a majority of incidents are not documented. However, no system in place to automatically spot this type of activity. One of the generally acknowledged issues that affect sufferers in a lasting way is online bullying. The approach to this issue is good social interaction, however, digital media networks should think about incorporating technologies and systems that really can aid in the detection and avoidance of these instances. Consequently, it is essential to create a smart platform or digital patrolling that will forbid inappropriate activity by observing and screening the offensive, divisive, and inappropriate information from digital media postings in order to create a safer and much more positive social atmosphere.

1.4. Defining the problem and its motivation

The school students in some particular and people of the community at large have started to think about online bullying as "not a big problem" and that it is okay to abuse someone continuously if a great number of incidents go unaddressed. But on the other side, whether there are suitable repercussions for such behavior, people will think twice considering taking quite a step. Additionally, everyone in the playground will be aware of the seriousness of the situation and the fact that these behaviors have repercussions. Nevertheless, depending on a victim and bystander to disclose the occurrences would not eventually accelerate this. This is a cause of the majority of online bullying instances getting unregistered and consequently punishable at the moment. Due to the enormous number of datasets created on digital media, the traditional categorization of current work can be very challenging to manage. Autonomous, 24/7, and precise detection techniques are essential for combating the issue of online bullying, as was mentioned in the preceding section.

Although a lot of digital media platforms are available nowadays in the world, Facebook, Ask.FM, including Twitter, are the best and most frequent places where online bullying occurs. So conclusively, a primary purpose of the project to identify online bullying activity in a publicly accessible Twitter database. Twitter is a popular platform for blogging. Those who have subscribed can view or write "Tweets," or comments with the character limit of 140. Those tweets are automatically made public. Authorized members of Twitter can also share images as well as videos. Users who are not signed up can see publicly accessible tweets. The identification of online bullying is the main purpose of our thesis is restricted to the text-based cyberbullying in the Twitter database, despite the fact that internet bullying can take numerous different forms, just like publishing humiliating comments, and images, including videos. Consequently, the tweet carries vulgarity, and caution must be exercised in determining whether it represents online bullying or not. Sequencing techniques are frequently used to identify online bullying. Due to the following factors, the distributed methodology is much more appropriate for the identification of online bullying on digital media sites like Twitter:

- Because the Twitter dataset is most of the time distributed as well as asynchronous, it is easier to identify instances of online bullying within the network.
- The point of the failure with bottlenecking may affect the sequential detection methods.

- By taking advantage of inherent parallelism connected with the independent development of the tweets, the distributed identification can shorten analysis time (Mangaonkar et al., 2015).
- The network-wide knowledge-based employed for identification is dispersed.

Apps including Twitter as well as Facebook are most of the time distributed since their dataset is created in an asynchronous, geographically diverse manner. These solutions are being inundated with the incoming dataset from numerous sources as the dataset is being generated simultaneously. A sequential method is unquestionably problematic in this circumstance. Therefore, the processing method used to handle this dataset needs to be flexible enough to work in the distributed context in addition to being rapid and economical. Therefore, if identification of online bullying must be done online, the procedure must start first before data transferred on centralized database. A main aim of this suggested study will train existing machine learning model. The proposed study will also develop a customized artificial neural network algorithm for the contribution of cyberbullying classification in term of accuracy. For the fair comparison of trained model, the proposed study will also present the comparative analysis of all the trained models.

1.5. Organization

The content of this report is divided in six different chapters. First one chapter is gives a brief introductory overview about proposed problem. The second chapter will discuss the published study on the in the field of cyberbullying and present a comprehensive literature review of the proposed problem. The chapter 3 will discuss about the dataset and chapter 4 will present the data exploratory analysis of the dataset. Next chapter will describe the complete methodology for the classification of cyberbullying types using tweets data. The last chapter will present the results of proposed methodology followed by the conclusion section.

Research Work

2. The Background and Literature Survey

This chapter explains a challenge of internet bullying. People faced this issue emerging in digital media a few years ago. It gives a brief overview of some suggested cyberbullying behavior detection techniques taken from existing literature. Different techniques proposed up till now are machine learning based, and some of them are related to lexicon-based techniques. Both of them are explained in this chapter.

2.1. Existing Cyberbullying Detection Research Work

To prevent or lessen cyberbullying on digital media forums, academics have been working hard on cyberbully identification for several years. The psychological strain of the violent, frightening, demeaning, and angry texts on recipients makes cyberbullying problematic. The phenomena of cyberbullying provide key information in regards to identification, protection, and reduction to lessen its negative impacts.

Numerous international programs are currently being implemented with the goal of reducing cyberbullying and enhancing the overall protection of internet consumers, especially kids (Leon-Paredes et al., 2019; Mangaonkar, 2017). There is research to minimize cyberbullying, which could be referred to as treatment and preventative measures in related literature. These strategies have their roots in the psychological and educational sectors. Those methods, though, are uncommon worldwide. Additionally, children who are bullied frequently avoid talking to their parents (Ho et al., 2019), teachers (Ibn Rafiq et al., n.d.), or some other people (Cheng et al., n.d.). They tend to consume a huge amount of time online surfing (Nahar et al., 2014), frequently seek out confidential support, and frequently post requests, including online support (Rahman et al., 2020). Nevertheless, using the World Wide Web to deploy anti-cyberbullying measures is more efficient. Additionally, patients can employ internet-based techniques anywhere and anytime they want to choose. For example, the University of Turku, located in Finland, has launched a Kiva anti-cyberbullying program, while the French government has launched an Anti-Harassment movement (Vaillancourt et al., 2017) and the Belgian state has launched an anti-cyberbullying strategy (Görzig et al., 2013).

These preventative and responsive measures must preferably be: (1) raise necessary awareness about potential cyberbullying dangers by employing specialized, comprehensive intervention techniques depending on the needs of the sufferers (Bhattacharya & Lindgreen, 2020; Zinovyeva et al., 2020), (2) convey health knowledge and develop psychological self-control (Pawar et al., 2018),

(3) make people afraid more conscious of both reactionary (e.g., filtering, removing, and disregarding communications) and preventative (e.g., enhanced vigilance and protection) actions that offer helpful tactics but also resource to help sufferers deal with the stress and bad feelings (Acı et al., 2019), (4) Because victims frequently participate in these two types of cyberbullying, it is also important to eliminate conventional cyberbullying (Pawar & Raje, 2019) and (5) comprise Online tagging, and compassion training, including positive online conduct (Louppe, 2014; Novalita et al., 2019). Online bullying is quite challenging to stop, so far, in this case. The majority of parents and their teachers depend on kids' knowledge of the causes and effects of online bullying. According to a few parents, mentoring is a good approach to stop online bullying, especially in children's teen years when friends have a bigger influence over their family and schools. As a result, additional specialized strategies or internet tools must be created to assist the sufferers (García-Recuero, 2016). For instance, Stauffer et al (Waseem & Hovy, n.d.), giving caution on mitigation, said that programs for stopping bullying only had a small impact on students' behavior.

Corresponding to this, experts in (Waseem & Hovy, n.d.) recommend that most schools implement the following steps while creating the program to avoid cyberbullying: Establish clear regulations, characterize cyberbullying, educate school staff, children, including families, on how to spot it, then educate them to report it if they do. Employ web filtration technology to comply fully. According to previous studies (Tu et al., 2017), social reinforcement (beneficial interpersonal contact that enhances the child's behavior) may be an important protection element in reducing the negative consequences of cyberbullying, according to previous studies (Chatterjee et al., 2019). They need to ask for assistance in order to receive all the necessary support to lessen any negative impacts of online bullying. Nevertheless, other studies indicate that internet victims of online bullying are unwilling to disclose incidents of online bullying and instead choose to remain quiet. Several adolescents occasionally ask their instructors or counselors for help (Rafiq et al., 2015; Tarwani et al., 2019).

Depending on the aforementioned prevention-related concerns, it is crucial to identify and block internet bullying on digital media. As a result, the focus of this part is on reviewing the cyberbully detection methods. The next subdivisions describe how the natural language processing and latest machine learning-based techniques, which are both primary approaches identified by literature studies, can be used to identify cyberbullies. The Lexicon-based techniques and some famous machine learning-based techniques are the two areas in which online bullying techniques are studied in the literature.

2.2. The Lexicon-based approaches

The textual classification techniques for identifying internet bullying are based on the straightforward bags-of-word technique. This produces the bulk of hurtfulness, violence, as well as offensive textual wording. The internet material which needs to be evaluated is again searched for using techniques. Here are some illustrations of the bag-of-word model: Lexicon Syntactic Features (LSF) is a technique that (Raza et al., 2020) suggested for detection of the cyberbullying. It can be viewed as a more intelligent and lexicon-based filtering technique which considering seriously the consumer's past record of unfavorable conduct on digital media platforms. It predicts the user's likelihood of sending abusive messages by combining message-level bully behavior detection and user-level bully behavior detection. This approach largely depends on Bag-of-Word and N-Grams algorithms enabling message-level offense detection. Assessments of the user-level obnoxiousness are done with a user's chat record. In the 2013 study, Kontostathis et al. looked at certain textual words online bullies use and the context. These terms are then utilized to create searches that are used to look through material related to online bullying.

2.3. Machine Learning based Techniques

Online bullying detection methods rely on textual classification algorithms produced with the help of artificial intelligence specially (ML) based algorithm. Intelligent algorithms produced with the (ML) techniques that are trained on the datasets that are typically constructed employing hand annotated digital media data. To make it much more informative, they also do sort of preprocessing. The research from these categories is listed below:

2.3.1. Cyberbullying detection by using NLP

One goal throughout this subject is the use of NLP for detect objectionable information. These tiers of some lingual method (Chen et al., 2012) describes much more illustrative way to explain whatever occurs inside Natural Language Processing technology. Users hire different layers to glean sense from writing or speaking the language. This scaling alludes to the rationale behind speech processing's predominance of theoretical approaches or information presentations attuned to differing stages.

- Nomenclature Grade (Information about a language sounds)
- Morphological Grade (Information of a meaningful part of the words)
- Vocabulary Grade (it handles the simple definition of difficult wording and different elements of the human language)
- Grammatical Grade (Information about structural connections between the words)
- Linguistic Grade (Information about a word's meaning)
- Language Used Grade (Information of language structures larger than the single speech)
- Practical Grade (Information on the connection between interpretation and the speaker's objectives and target)

For instance, (Akhter et al., 2019) usage of the common-sense information base and related thinking methods. Employing search phrases that are frequently utilized in internet bullying scenarios, identified internet bullying literature utilizing the Formspring.me datasets. In order to identify signals of bullying (any new word referring to the online internet-based references which may be harassment incidents itself or internet-based references referring to the off-line cyberbullying incidents), (*Modeling the Detection of Textual Cyberbullying | Proceedings of the International AAAI Conference on Web and Social Media*, n.d.) employ a variety of natural language processing techniques. They employed Latent Dirichlet Analysis to pinpoint topics or themes and sentimental analysis characteristics to pinpoint internet bullying behaviors. The goal of the authors in (Chen et al., 2012) is to lay the groundwork for numerous tasks connected to online bullying identification and also to issue a challenge to additional scholars to improve those particular methodologies. In light of all this, (Akhter et al., 2019) was the initial researchers who worked with NLP internet bullying and its detection, and they researched forecasting power n-grams, segments of human speech data, moreover, sentimental data depending upon expletive lexical items for the task (both with or without TF-ID weighting). Similar characteristics were employed in (Mccallum & Nigam, n.d.) for both the detection of the text types and cyberbullying-related occurrences.

To sum up, Terms Frequency (TF), Terms Frequencies Inverse Documents Frequencies (TF-IDF), the global Vector (GloVe) for Words Representing (Word2Vec), as well as others are frequently employed word representation strategies that have been shown to increase categorization accuracy (Buczak & Guven, 2016). Context expert understanding is a primary NLP restriction. For example, there are many questionable claims regarding detecting sarcasm, so how will anyone recognize it in a brief paragraph such as "The Great Game" in response to the loss? As a result, it is not about language; rather, it is also about having information that is pertinent to discourse.

Another method of detecting internet bullying that has been extensively employed by many academics is machine learning-based internet bullying buzzwords. Additionally, machine learning (ML) is the subfield of AI which enables computable systems for learning or improving autonomously from experience in the absence of the need for static programming (Zhang, 2011). It is frequently divided into supervised learning, semi-supervise learning, moreover unsupervised learning techniques. In a supervise approach, the model that produces the intended projection (i.e., depending on annotation or labeled dataset) is built using a number of training examples. Unsupervised learning approach, secondly, not data Dependant and typically used for clustering issues (Joachims, 1998).

A strategy for discovering inappropriate statements on digital media through screening or alerting individuals affected was put forth by the (Chatzakou et al., 2019). For the training of this algorithm, researchers utilized remarks on Twitter and Ask.fm that contained abusive language. Some other authors (Hosseinmardi et al., 2015; Irena & Setiawan, 2020) developed communication networks depending on intelligent robots who give comfort and help to the victims of internet bullying. proposed a method for identifying internet bullying that focuses on identifying the aggressive patterns in user posts by analyzing objectionable phrases and utilizing the rating stage of a threat. In a similar vein, 50 decision trees achieved an accuracy of 82.7%.

The researchers of (van Hee et al., n.d.) explain the installation of the online tool for parents, including school employees in Japan, who are responsible for identifying subpar information on unofficial supplementary web pages. The objective of their work was to provide existing data on internet bullying to the federal government. SVMs were used, and the accuracy rate was 79.9%. The Facebook structure has been developed by (Kowsari et al., 2019) to shield young users against sex harassment and internet bullying. In order to track behavioral changes, this system analyzes the content of photos, including videos, as well as user activity. The Formspring.me site's 3815 submitted comments were used to compile the list of objectionable phrases in (Sahlgren et al., 2018). About 58.4% accuracy was found in their investigation (*Salminen*, n.d.).

Implementing the genuine scenario of internet bullying detection in Twitter utilizing by supervised learning approach. Sequential Minimum Optimization (SMO) classification achieved (68.47 percent), the greatest accuracy amongst others, an investigation using two different characteristics extraction approaches and numerous machine learning-based models. The researchers of a study have suggested a method for detecting online bullying that is dependent on the digital media of Instagram. The investigations were conducted using consumer comments and digital image feature analysis. The outcome demonstrates that using a variety of features can increase its linear support vector machine's (SVM) classification performance. By adding digital image pixel categories as the additional characteristics, the overall accuracy of SVM increased from 0.71 to 0.77. (Snakenborg et al., 2011) suggests developing a weighted and directed graphical model for internet bullying. It can be utilized to determine every user's attacker and sufferer ratings whilst utilizing the weighted TF-IDF technique with the text features (2nd person pronoun or foul language).

Different digital media platforms could benefit from the hate speech detection method, according to (Patchin & Hinduja, 2010). The writers used 198,566 responses in total, with 79% of them being deemed benevolent and the remaining 21% being violent, across four different digital platforms such as YouTube, Facebook, Wikipedia, and Instagram. Numerous machine learning methods were used in these studies to test every feature individually and assess the accuracy depending on the attributes chosen. In contrast, machine learning-based classifiers (Tenenbaum et al., 2011) proposed a suitable technique for integrating users, content, and responsibilities characteristic of internet bullying. The outcomes demonstrated improved performance when every aspect was utilized jointly. The collection of the German digital media posts was created by (Olweus, 2012) and labeled as various forms of internet bullying, including threats as well as insults. The researchers went into great detail about individuals who participated in online bullying (identifying the victims, internet predators, and eyewitnesses). By extending the term to include cyberbullying cues depending on word embedding, In (Hinduja & Patchin, 2008), we were able to use the SVM classifier to get an f-measure of 0.79. The vocabulary of common phrases utilized with neurotics in digital networks was also employed to help develop the unique characteristics. The Word2Vec integrating algorithm-dependent neural networks were employed by the researchers in (Hemphill et al., 2012) to provide textual health data with the semantic framework.

Additionally, the Word2Vec approach incorporates distinct area ontologies. More information on the neural network-based model that detects the semantic significance of unusual terms is provided by such ontologies. To correctly differentiate between unorganized and organized healthcare data, additional semantic information based on the BiLSTM framework is used. To identify and categorize fake stories over Twitter, a distinct work, a decision tree c4.5 classification based on the TFID method is utilized to add characteristics to the recommended C4.5 classification algorithm, such as N-gram, that is used in (Casas et al., 2013). The writers of (Ang & Goh, 2010) have suggested the novel approach which deals a most part of pertinent news stories, papers, comments, and retweets from digital media. Additionally, researchers combined topic2vec with Word2Vec to produce a word embedding algorithm that gives every word in the document the low-dimensional vectors and the semantic contextual meaning. The machine learning algorithms previously described were utilized by the researchers to classify data by using machine learning. This research has used a number of supervised learning-based algorithms to improve the classifier accuracy and effectiveness in identifying internet bullying in digital media, particularly on Twitter, as it is a classification issue (For example, labeling the given examples as these are harmful or not harmful case). A following classification techniques are used herein this research:

2.3.2. Logistic Regression Technique

A very well-known statistical method that machine learning has developed is logistic regression (Barlińska et al., 2013). By using this logistic function, a process of logistic regression creates the hyperplane which connects two different datasets (Ybarra et al., 2006). This logistic-regression method uses characteristics as an input and generates the predictions based on the likelihood that the class would match this given input. According to the equation, the forecast would be for another group (negative class) if the likelihood is greater than 0.5, in which case the example categorization will be a positive class (1). Using logistic regression, internet bullying prediction algorithms were implemented in (Smith et al., 2008).

$$h_\theta(x) = \frac{1}{1 + e^{-\theta \mathrm{T} x}}$$

The value of h (x) is > or = 0.5, which shows it falls into a positive class. Similarly, if the value of h (x) is less than or equal to 0.5, it means it falls into a negative class. So, as it is described in (Raisi et al., 2016), logistic regression performs much better on binary classification tasks. If the size of the data increases, it will perform better. An error function is attempted to be minimized while logistic regression updates a set of various parameters repeatedly.

2.3.3. Logistic Light Gradient Boosting Machine (Logistic LGBM)

Logistic LGBM, the gradient boosting architecture that employs the tree-based supervised learning method, is the most efficient boosting approach in machine learning (Bosse & Stam, 2011). When contrasted with XGBoost and CatBoost, this works significantly better (Reynolds et al., 2011). In

LGBM, the measurements that were utilized to calculate the difference are classified using gradient-based one-side sampling (GOSS). The main benefit of LGBM is that it alters the training method, which speeds up all processes considerably (Raisi et al., 2016) but frequently results in a model that is much more effective (Rybnicek et al., 2013). Several classification areas, including cyber behavior identification (Codina et al., n.d.) and abnormalities prediction in huge financial information (Bayzick, n.d.), have adopted LGBM. Nevertheless, LGBM was not frequently applied to the detection of internet bullying. So, the described research work, we try to investigate the LGBM in online bullying detection to assess the precision of its categorization.

2.3.4. Stochastic Gradient Descent (SGD)

The optimization approach called the stochastic gradient descent utilized to determine the model related values the function ‘f’ that reduced the overall function of the cost. As shown, the equation, SGD, in comparison, updates the parameters for every training instance x^i with a tag y^i.

$$\theta = \theta - \eta . \triangle_{\theta}\, j\left(\theta, x^{i}, y^{i}\right)$$

SGD was employed throughout [90] to create an online bullying forecasting algorithm on digital networking sites. According to the authors in this paper (Salminen et al., 2020), SGD is much faster as compared to Naive Bayes and Logistic Regression, although errors may not be as small as they are in Logistic Regression.

2.3.5. The Random Forest based approach

This is a classifier and technique (Dinakar et al., 2012) which compares different decision taken on various dataset examples or parts of these examples while utilizing the dataset to improve prediction accuracy as well as regulate the fitting (Dadvar et al., 2012). For this purpose of classification information, ensemble techniques integrate multiple techniques of the same or multiple types (van Hee et al., 2018). The research conducted by (van Hee et al., 2018) provides instances of how RF is frequently using in literature for creation online attack forecasts. As a result, Random Forest uses the number of forests to randomly select the classifier information parameters. The building of Random Forest is carried out in four processes that follow. The number of instances as N or cases for training, while M shows number of classification attributes.

- N shows the number of instances or cases, while M shows total the total classification characteristics of training dataset.
- A collection of arbitrary decision trees is produced by choosing characteristics on a random basis. The training dataset chosen, every branch with selecting many times from among the N

examples already present. By estimating their categories, the remaining examples in the learning dataset utilized to calculate an overall error of these trees.

- In order to build a decision at every tree node, M random parameters are picked. Utilizing particular characteristics, the training package's best outstanding division is identified. Every tree is constructed totally rather than being trimmed, as is possible when creating a standard tree classification.
- A lot of decision trees are produced by architecture. These decision trees cast their votes for the simplest class. These procedures are known as RFs. In order to determine which category is most preferred, every tree-structured classification in the Random Forest's framework casts one vote. The subclass with the most votes is the one that is chosen as the output.

2.3.6. AdaBoost algorithm

Need to increase an performance of binary classification, adaptive AdaBoost, the common enhancing strategy, was first created (Havas et al., 2011). It employs the iterative process to turn very weak classifiers into very strong ones by learning from their mistakes. Every training assessment, therefore, is given equal weight in the beginning. It utilizes a number of inadequate frameworks and gives the experimental wrong classification observations more significance. The reliability of incorrectly classified data is increased as results of definitive parameters produced over a convergence rate are integrated by using a number of low assumptions. As a result, the entire iteration's precision is improved (Jacobs et al., 2014). In the first figure, it gives the description of implementation of the AdaBoost classifier. It depicts similar data with 2 features as well as the same number of classes, weak learner number 2 improving on weak learner number 1's error, moreover an accuracy for the non-classified observational data being more upgraded couple of week classification algorithms are merged to make a strong learner. Furthermore, Adaboost have utilized internet bullying identified by various investigators, including (Chatzakou et al., 2019), where they utilized it for online bullying detection but also have got the precision accuracy of 77.38 percent by the Adaboost, using the unigram, remarks, as well as communication as characteristics.

2.3.7. Multi nominal naive Bayes

For classifying issues involving textual documents, this is frequently utilized. Nevertheless, multinomial NB was most frequently employed to develop internet bullying forecasts in the detection of online bullying, as seen in (Hinduja & Patchin, 2008; Irena & Setiawan, 2020). By using the Bayes theorem between characteristics, multinomial Naive Bayes classifiers were created. This algorithm employed training data to derive overall Bayes-optimized parameters estimates for the model as well as assumes that the parametric approach generates plain text. It classifies the generated testing data using the assumptions. Multinomial NB classifications are capable of supporting the infinite variety of distinct categorical parameters. The job of predicting the multi-dimensional examples simplified for predicting another single dimensional kernel density under the assumption that functions are different. The multinomial NB technique is the learning approach built on the Bayes theorem using strong (naive) independent assumptions. Consequently, multinomial NB was covered in depth.

2.3.8. Support vector machine (SVM)

The SVM is the very famous and it is supervised ML technique designed for classification of various text-based data (Joachims, 1998). The initial feature set is transformed into the higher-dimensional structure with the user-defined kernel by SVM, which further looks for the support vectors to minimize averaged distance (difference) among the 2 groups. SVM initially makes the approximation of the hyperplane dividing the 2 groups. As a result, SVM chooses support vector samples from both groups that are closest to the hyperplane.

SVM aims to effectively separate these two groups (e.g., the positive class and the negative class). Support Vector Machine implements particular kernels to convert its function space suitably if the dataset can be divided along irregular borders. For datasets it is difficult to separate, soft margins are used to avoid overfitting by assigning categorization mistakes at decision borders reduced weight (Havas et al., 2011). The study describes, the SVM by using a kernel which is a linear as a basis ‘f’ or a function is used. The Support Vector Machine classifier framework is illustrated in the figure Figure 2 for dataset containing couple of elements and two classes, while the complete training dataset represented either circles or dots.

These training samples' support vectors (also known as the stars) were for every of two groups, making them training specimens that really are closest to the hyperplane. Due to being on the incorrect side of the hyperplane, 2 training outcomes were misidentified.

SVM was, therefore, employed to create online bullying forecasting models in (Havas et al., 2011) and it was discovered to be successful and more efficient. Nevertheless, the research in (Joachims, 1998) found overall efficiency dropped as data quantity expanded, indicating how SVM maybe not best solution for handling the numerous linguistic difficulties included in online bullying.

Related Knowledge

3. Related Knowledge

3.1. The term frequencies inverse documented frequency (TF-IDF)

It used for transformation of text data into numeric features. For the extraction of numeric features, This find importance of words unlike Count Vectorizer that simply count the frequencies of the words in the certain document. Significance about a term for a document is not accurately reflected by utilizing a word's word count as its only feature value. For instance, if a term occurs often throughout all of the articles in a dataset, its count value in various documents is useless for differentiating between them. However, if a word is only found in a small number of papers, then that word's count value in those documents can assist in differentiating them from the other documents. As a result, a word's feature value, or relevance, for a text depends on both the frequency of its use in that document as well as the corpus as a whole. This method of finding word importance in the certain document is called as completed document. It is a TF-IDF weighting procedure.

Word "frequency" refers to the ratio between the rounds of the times, the words shown in the pharases. As a result, it is a normalized measure that accounts for document length. Let's use tf_{ij} to display the frequency of words in document j. The quantity of the certain documents as bulk that contain word A is indicated by document frequencies of the word A. Let's use df_i to denote the document frequencies for the word I. The following formula mostly using for culculation of TF-IDF's weights related to words A in the whole document J with n shows the amount of documents in bulk:

$$w_{ij} = tf_{ij} * \left(1 + log\frac{1+N}{1+df_{ij}}\right)$$

3.2. The Random Forest (RF)

The RF is a tree-based ensemble training model about the classifications as well as regression problems. RF develops the multiple trees during the training of the model. The final output is the class that selected by the majority trees. It can also be used to grade these qualities according to overall significance. There are numerous types of the decision-trees in RF . "forest" is created by random-forest method and the trained by using the bagging and bootstrap aggregating. Bagging, the aggregation of the meta-algorithm, improves effectiveness of the machine-learning methods. RF model chooses the outcome based on predictions made by decision trees. Through the averaging or average out outcomes from various branches, it provides forcasting. As there are more trees, overall accuracy of outcome improves. The RF algorithm overcomes the drawbacks of decision-tree technique. It increases accuracy and reduces the data overfitting.

Features of Random Forest model are following:

- Compared to the decision tree algorithm, it is more accurate.
- It offers a practical method for the working by the missed dataset.
- Without using the hyper parameter settings, it can generate the appropriate forcast.
- This fixes a overfitting issue with decision trees.
- The point where network node is split up in each RF tree, the set of characteristics is selecting randomly.

3.3. Suport Vector Machine (SVM)

This is a supervise ML algorithm that is called SVM can apply for the data classification and regression problem. However, classification issues are where it is most frequently utilized. SVM uses different types of kernels for the classification of data like Linear, Non-Linear, Polynomial and Sigmoid. Kernel is a function in SVM that helping in solving the problem. SVM is a very useful model high dimensional dataset.

When using the SVM classifier, every data point is represented the point in the multi dimensional space (where n is the total numbers of characteristics you have), by every value of the given feature being the number of the certain coordinate. Furthermore we perform classify by identifying these hyperplane which is effectively discriminate both groups (shown in the diagram below).

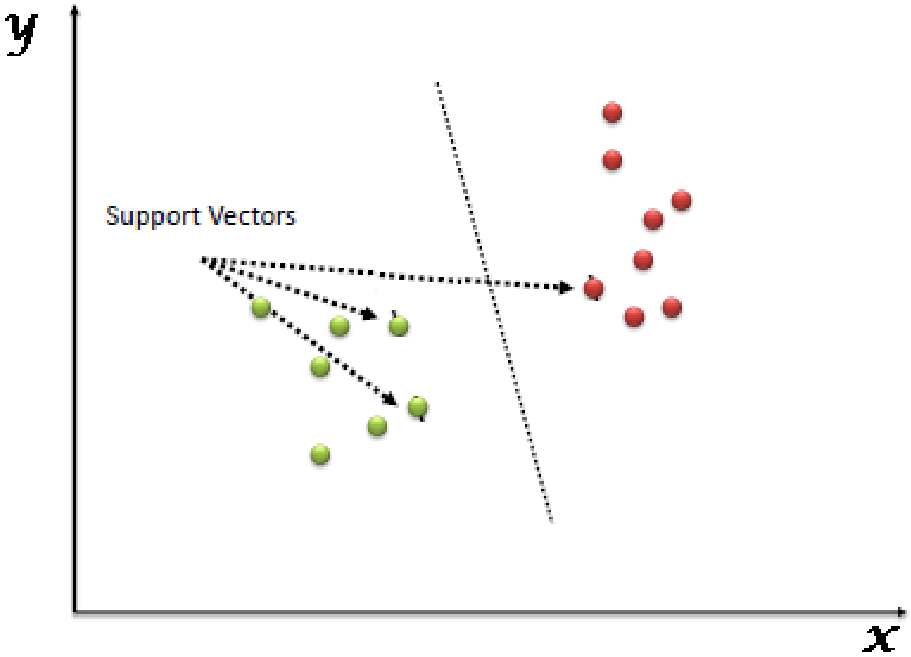


Figure *1***:** Classification method of Support Vector Machine.

3.4. K-Nearest Neighbor Model

The KNN model believes that related things are located nearby. In other words, related things are located close to one another. The majority of the time, related data items in the below image are near to one another. This presumption must be true enough for the KNN algorithm to be effective. KNN uses the arithmetic we may have learned as children—calculating the distance between points on a graph—to encapsulate the idea of similarity (also called distance, proximity, or closeness).

KNN Algorithm

- One of the simplest machine learning techniques, based on the supervised learning method, is K-Nearest Neighbor.
- The K-NN algorithm makes the assumption that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories.
- A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. This means that utilizing the K-NN method, fresh data can be quickly and accurately sorted into a suitable category.
- Although the K-NN approach is most frequently employed for classification problems, it can also be utilized for regression.
- Since K-NN is a non-parametric technique, it makes no assumptions about the underlying data.

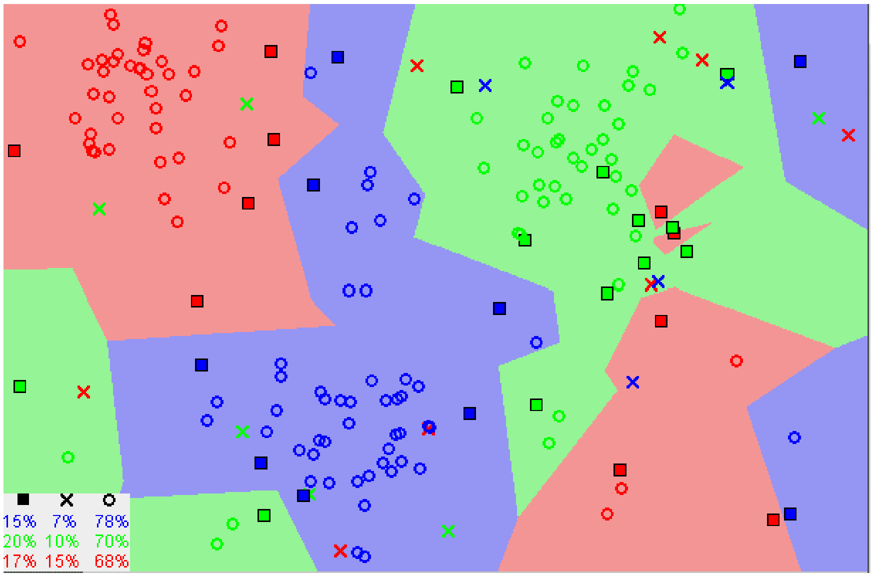


Figure *2***:** Classification Method of K Nearest Neighbor.

3.5. Neural Network Model

Artificial neural networks are made to mimic the functioning of neural networks in the brains of humans and other animals. Machine learning acquires the model architecture needed to handle increasingly complicated data by mimicking and modelling the function of neurons. Artificial neural networks come in a wide variety of forms, with many early incarnations appearing straightforward in comparison to contemporary methods. For advanced deep learning models, artificial neural networks are utilized as the architecture. The brain's neurons are modelled as simplified versions of artificial neurons or nodes. The number and strength of connections between each artificial neuron and other nodes varies depending on the type of artificial neural network. Between the input and output layers of the network, there are often layers of nodes. Because of the density of these layers, this multi-layered network architecture is also referred to as a deep neural network. These many layers in artificial neural network models can pick up on various data aspects. Hidden hierarchical layers enable complicated concepts or patterns to be understood from processed data.

There are five types of neural networks models:

- Feedforward artificial neural networks
- Perceptron and Multilayer Perceptron neural networks
- Radial basis function artificial neural networks
- Recurrent neural networks
- Modular neural networks

Dataset

4. Dataset

4.1. Introduction

The Cyberbullying classification dataset was downloaded from Kaggle that was based on the approximately 47 thousand samples. The samples of selected dataset were labeled with the type of cyberbullying. The dataset was based on the two columns: the tweet text and the label of the text. The text tweet was labeled with the six different categories, five categories belong to cyberbullying types and six categories belong to tweets of normal tweets (without any cyberbullying attacks). The tweets in the dataset are based on average 20 words. The content of the tweets in dataset was available English language. Collectively, our selected dataset was based on ~47,000 tweets in English language with different labels of cyberbullying. The sample content from each class is represented in Table 1.

Table *1***:** Samples of Dataset from different classes.

Text	**Class**
In other words #katandandre, your food was crapilicious! #mkr	Not-Cyberbullying
rape is real..zvasiyana nema jokes about being drunk or being gay or being lesbian...rape is not ones choice or wish..thtz where the sensitivity is coming from	Gender
Sudeep, did she invite him though? No right? Why are you getting worded up? You're okay with Parvesh Verma cause he speaks against Muslims but against an idiot like Imam because he called for chakka jam?	Religion
@ikralla fyi, it looks like I was caught by it. I'm not a botter, so...	Other-Cyberbullying
Here at home. Neighbors pick on my family and I. Mind you my son is autistic. It feels like high school. They call us names attack us for no reason and bully us all the time. Can't step on my front porch without them doing something to us	Age
Hey dumb fuck celebs stop doing something for people for publicity on Facebook... Wtf happen to life u niggers are cowards.	Ethnicity

4.2. Preprocessing

The first stage in preparing a dataset for machine learning and deep learning models is to clean truncate and transform the dataset. In this context, we employ a variety of data pretreatment approaches, including data cleaning, tokenization and lemmatization.

For data cleansing, we eliminate all stop words from the tweets. Additionally, we eliminate punctuation, symbols, integers, and Hyperlinks from the tweets of Cyberbullying dataset. In addition, the tweets' emojis symbols and hash tags were removed, and the residual text was changed to lower case. The final step in the preprocessing stage was stemming, which changed words from their alternate forms to their conventional ones. We utilized the Porter Stemmer Algorithm from the NLTK tool to stem tweets. Additionally, it aids in the extraction of key features. Table 2 displays an example of a text after preprocessing.

Table *2***:** Sample of Tweet Text before and after preprocessing.

Tweet Text (Before Preprocessing)	**Tweet Text (After Preprocessing)**
In other words #katandandre, your food was crapilicious! #mkr	word katandandr food crapilici mkr

4.3. Data Exploratory Analysis

4.3.1. Target Variable Description

The cyberbullying dataset was based on the tweets labeled with the cyberbullying types. Collectively the dataset was based on the two features that was tweets text and the label of the text. The label of the tweets text was the most suitable features for the classification of cyberbullying attacks. After the initial understanding of the of dataset, we selected the label of the tweets as a target variable. The target variable contains the six unique values that translated that the six types of tweets are exist in cyberbullying dataset.

Table *3***:** Description of input and target variable.

Features	ID, Tweet Text, Cyberbullying Type
Target	Cyberbullying Type
Classes	Not-Cyberbullying, Gender, Religion, Other-Cyberbullying, Age, Ethnicity

4.3.2. Most Frequent Words

For the understanding of mostly usually words in dataset, we perform the most frequently used words analysis. In this process, we found the count of each word and then sort all the words based on their frequency in ascending order. In this regard, different forms of words are considered two different words. For instance, plays and playing will be considered two different words. For the fair counting of the words, we convert the all words into their basic form by the use of stemming technique (Preprocessing section). After converting all the words into their stem form, we count the frequency of all words in dataset. Lastly, we Plot the top 10 words that are most frequently used in the dataset by matplotlib library. The bar chart of top 10 most frequently used words is shown in Figure 3.

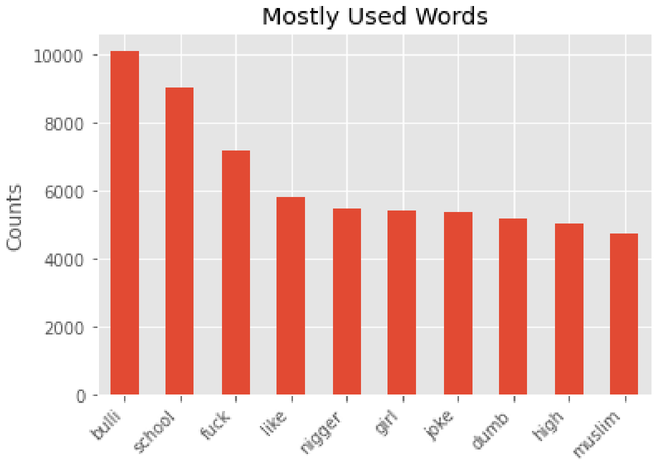


Figure 3: Top 10 Most Frequently used Words in Tweets.

4.3.3. Class Distribution Analysis

For the better understanding of the dataset, we find out the class distribution in cyberbullying dataset. Class distribution analysis reveal that how many samples are fall against each class in target variable. For the class distribution analysis, we use “value-count” built-in function of pandas. It calculated the number of tweets against each class. We found the six different classes in dataset and approximately 8000 samples against each class. Further, the class distribution data was used to plot the bar chart of class distribution by the matplotlib library. The class distribution bar chart of cyberbullying dataset in shown in Figure 4.

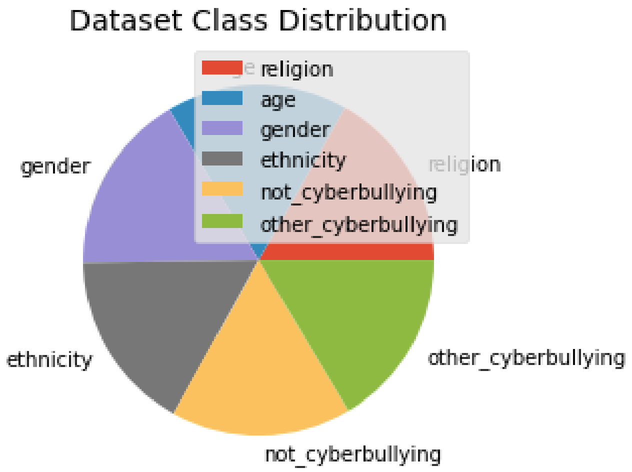


Figure 4: Class Distribution of Tweets Dataset.

4.3.4. Length of text in Class Distribution

We also analyze the length of the tweets in each category. We used the processed text for the calculation of each tweet length. For the length calculation of tweets, we count the total words in that tweet and represent the count as tweet length. We use the tweets lengths data to plot the length of tweets collectively and separately for each category of cyberbullying. We see that the length of tweets in gender and ethnicity class are medium in size (average 14 words). We also reveal that the tweets labeled with not cyberbullying class having the length of average 7-8 words while the religion and age cyberbullying tweets are very large in size and having the average 17-18 words. The histogram plot of collective dataset and individual class against text length is presented in Figure 5.

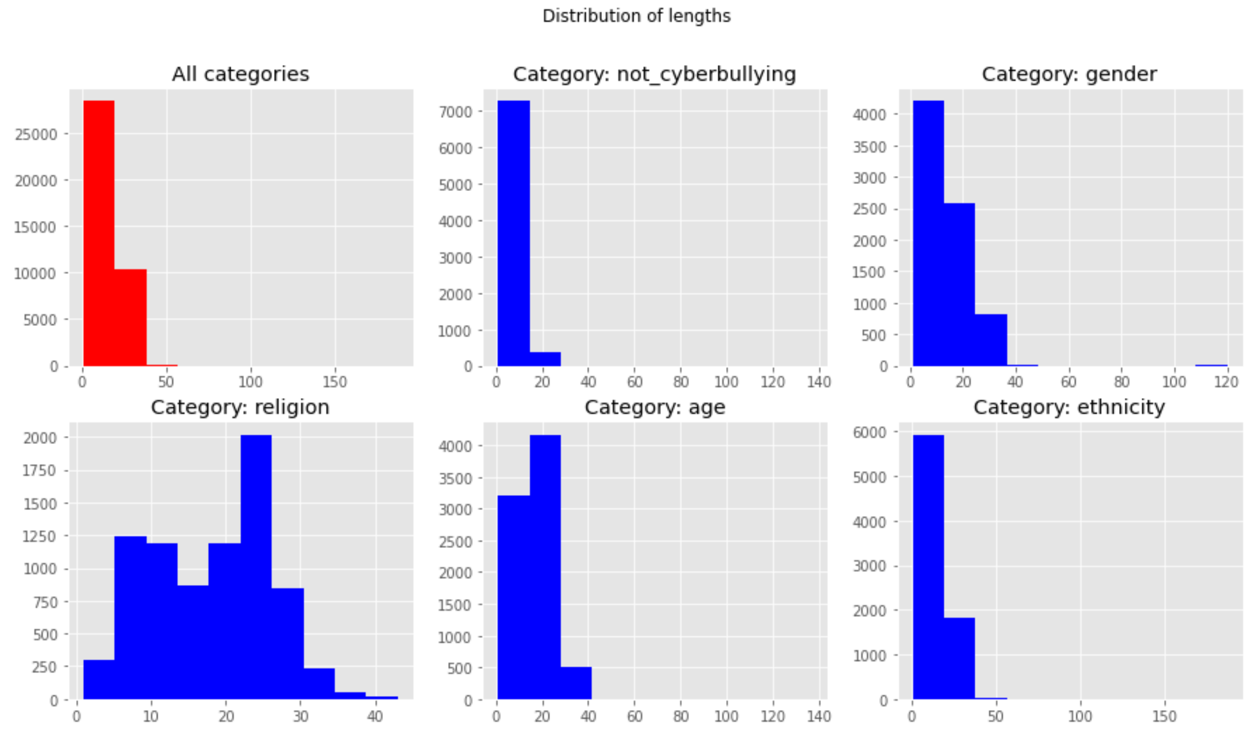


Figure *5***:** Length of words in each type of cyberbullying.

4.3.5. Count of tweet against text Length

We also count the tweets count against the text length 1 to 10. For this, we used the tweets length data. After calculating the tweets length, we group-by all the tweets based on their length and then count the tweets for each value. Further, we sort the frequency of tweets against each value and extract the tweets counts for 1 to 10 words. The extracted data was further used to plot the bar chart for enhanced representation. The bar chart of the tweets count for 1 to 10 words length is shown in Figure #.

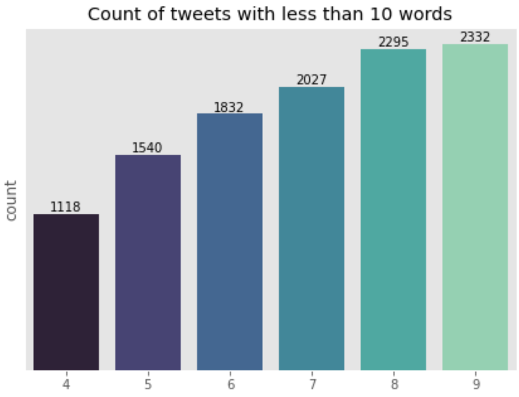


Figure *6***:** Count of tweets with 4 to 10 words.

Methodology

5. Methodology

In this section we will discuss the proposed methodology for the classification of cyberbullying types based on the tweets of cyberbullying. The rest of the section will discuss the methodology of feature extraction, model development, model training and evaluation of trained models.

5.1. TFIDF Feature Extraction

Feature extraction technique refer to the method of extracting the features from unstructured and unformatted data. As the selected dataset for proposed study was based on the non-structural text data and data need to pass from feature extraction process. We convert the text data into structural data by extracting the features from text data. From the well-known methods (Count Vectorizer, TF-IDF, and word2vec) in NLP domain, we used the TF-IDF vectorizer for the extraction of features from text data.

TF-IDF extract the features based on the occurrence of the word in the tweets (section 3.1). By using the TF-IDF, we convert each tweet into the 200 features base on the frequency of the words. After the complete training of the TF-IDF vectorizer, it extracts the 200 features for each tweet.

5.2. Train Test Split

After the preprocessing and feature extraction process, dataset need to be split in different subset for the training, testing and validation of the model. We used the built-in train-test-split function od scikit-learn library for dividing the dataset into three different subsets. The built-in train test split function randomly selects the samples from each class with some ratio for one subset and rest of samples for second subset. It did not override the samples into the divided subset. We also used the extracted features dataset from TFIDF and split into three different subsets. The extracted features dataset was dividing into training, testing and validation set with the ratio of 70%, 20% and 10%. After splitting the dataset, the train, test and validation set contain the 26721, 7423 and 2970 samples.

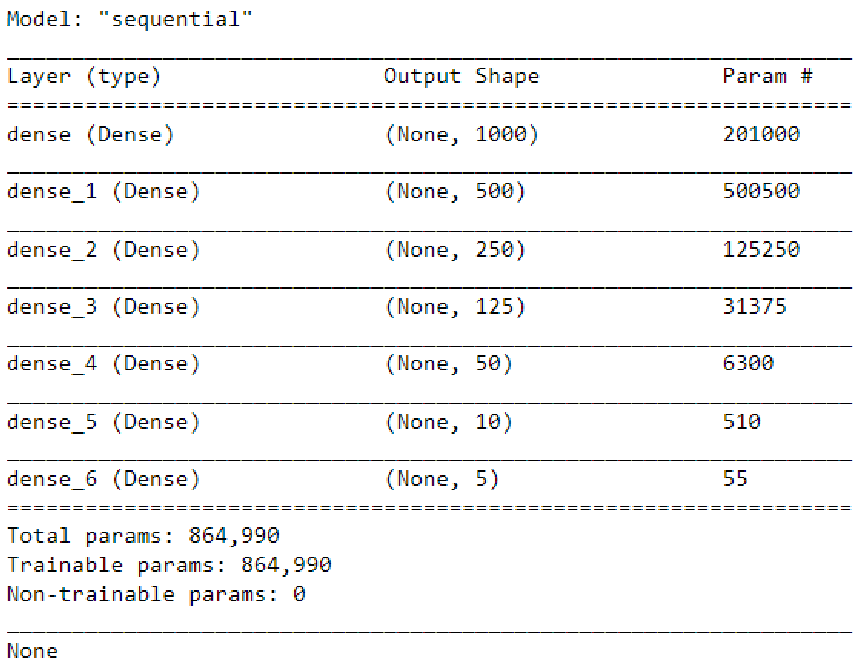
Table 4: Samples in train, test and validation set.

	Training Set	Testing Set	Validation Set
Number of Samples	26,721	7423	2970
Number of Features	200	200	200

5.3. Models Development

For the classification of cyberbullying types with tweets data, we used different machine learning and deep learning models. From the machine leaning models, we used the Random Forest, Support Vector Machine, and K Nearest Neighbor algorithm for the classification of cyberbullying. We did not develop the machine learning model from scratch. We initialize the already developed the machine learning models by using the scikit learn library of python. SVM model was initialized with the 200 hyper parameter value of max-iteration. Random forest was initialized with the max-depth value of 5 and bootstrap value of 100. While the KNN model was initialize with the 5 n-neighbor value. Rest of the hyper parameters of machine learning models were used with their default values.

For the development of the neural network model, we used the input layer, output layer and five hidden layers. We used the 1000 and 5 neurons on input layer and output layer respectively. The hidden layer also used the 500, 250, 125, 50, and 10 neurons respectively with ReLU activation function. As the proposed problem is the multiclass classification problem and Softmax is most suitable activation function for multiclass problem, we used the Softmax as an activation function on the output layer of the proposed model. The summary of customize develop model is presented in Figure 7.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 1000)              201000
_________________________________________________________________
dense_1 (Dense)              (None, 500)               500500
_________________________________________________________________
dense_2 (Dense)              (None, 250)               125250
_________________________________________________________________
dense_3 (Dense)              (None, 125)               31375
_________________________________________________________________
dense_4 (Dense)              (None, 50)                6300
_________________________________________________________________
dense_5 (Dense)              (None, 10)                510
_________________________________________________________________
dense_6 (Dense)              (None, 5)                 55
=================================================================
Total params: 864,990
Trainable params: 864,990
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 7: Proposed Model Architecture.

5.4. Model Training

For the training of the machine learning models, simply the train set was passed to the initialized machine learning models. After the complete training of the models, all the trained models were evaluated using the 7423 samples of test set. While for the training of customized neural network model, Adam optimizer was used with batch size of 32. The learning rate of the model was also tunned with the value of 0.001. Lastly, the model was trained on the 50 epochs.

5.5. Model Evaluation

For the evaluation of the trained models, we used different evaluation measures including accuracy, precision, recall and f1-score. Recall or sensitivity is the ratio of real positive cases that correctly predict positive with the total real positive cases. Contrarily, precision or confidence refers to the percentage of predicted positive instances that are actually real positives. So, we can mention the recall means “how many samples of particular class you find over the all samples of that class," and the precision will be “how many are correctly classified among that class." The f1-score is the harmonic mean between precision & recall. The test set with 7423 samples was used for the evaluation of the trained model. The evaluation measures were calculated by using the formula of that measure. The equation of calculating the measures is presented in eq 1-4.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Eq. 1}$$

$$Precision = \frac{TP}{TP+FP} \qquad \text{Eq. 2}$$

$$Recall = \frac{TP}{TP+FN} \qquad \text{Eq. 3}$$

$$F1Score = \frac{2*(recall*Precision)}{Recall+Precision} \qquad \text{Eq. 4}$$

Experiments and Results

6. Experiments and Results

6.1. Environmental Setup

We will discuss the development environment, used libraries, programming language and all other related setting for the training of the proposed models. All the experiments were performed by using the python programming language. The python 3.7 version was use to perform the experiments. A Conda environment was established with python version 3.7 for the training of the models. All the essentials library were installed in the developed environment.

From the well-known framework for deep learning models, we used the TensorFlow framework for the development and the training of the proposed models. We installed different libraries in our established environment. The list of all core libraries with their version number is listed in below table (Table 5).

Table 5: List of libraries used in environment

Library Name	**Version**
TensorFlow	2.3.0
Matplotlib	3.5.2
scikit-learn	1.10.1
Pandas	1.3.5
Numpy	1.19.0
NLTK	3.6.6

6.2. Experimental Details

We perform number of experiments for the classification of Cyberbullying types using tweets text. The detail of all experiments in presented in Table 6.

Table 6: List of Proposed Experiments in proposed study

Experiment no.	Experiment Name	Details
Experiment 1	Feature Extraction	Extract features from text by using the frequency of words through TF-IDF vectorizer.
Experiment 2	Train Test Split	Split the dataset with the ration of 70%, 20%, and 10% in training, testing and validation set. Use the Train test Split function of scikit learn library.

Experiment 3	Random Forest	Train the Random Forest model with the 26,721 samples. Use the default value for all hyper parameters of Random Forest. After the Complete training of the model, Evaluate the model on test set.
Experiment 4	SVM	Train the Support Vector Machine model with the 26,721 samples. Use the default value for all hyper parameters of SVM. After the Complete training of the model, Evaluate the model on test set.
Experiment 5	KNN	Train the KNN model with the 26,721 samples. Use the default value for all hyper parameters of KNN. After the Complete training of the model, Evaluate the model on test set.
Experiment 6	NN	We train the customize Neural Network model with the training and validation set. Model was trained with 50 Epochs and 0.001 learning rate. After the complete training of the model, use the test set to calculate the performance of the model.
Experiment 7	Comparative Study	After the training of the all models, we perform the comparative study. In this study, we plot the comparison graph to evaluate the best model.

6.3. Experimental Results

The experimental result section will present the result of all proposed experiments by using different tables and graphs.

6.3.1. Train Test Split

By splitting the dataset into three different set, we get the train, test and validation set. Train Test Split function of scikit learn generate these subsets from the tweets dataset. The class distribution in the generated train test and validation set is presented in Figure 8.

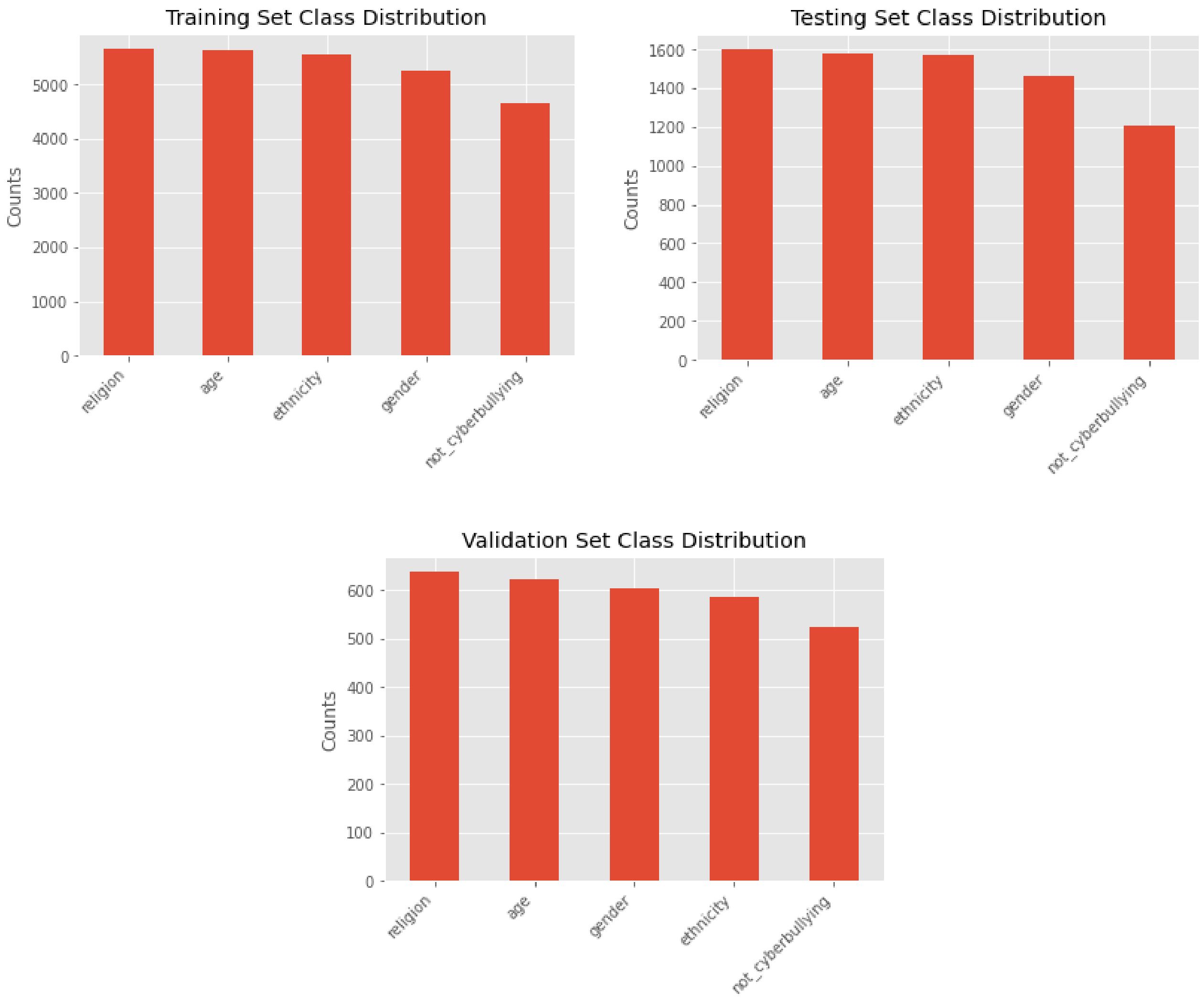


Figure *8*: Class Distribution in split datasets

6.3.2. Support Vector Machine

After the complete training of the SVM model, random forest showed the 0.82% test accuracy on test set. The complete classification report of SVM model on test set in presented in Table 7. The confusion matrix of SVM model on the 7423 samples of test is also presented in Figure 9.

Table *7*: Cyberbullying Classification - SVM Report

Classification Report for SVM Model				
	precision	recall	f1-score	support
Not-Cyberbullying	0.95	0.96	0.97	1588
Gender	0.97	0.99	0.98	1489
Religion	0.79	0.75	0.77	1446

Age	0.66	0.67	0.67	1300
Ethnicity	0.92	0.92	0.92	1600
accuracy			0.87	7423
macro avg	0.86	0.86	0.86	7423
weighted avg	0.87	0.87	0.87	7423

Accuracy (Train Set): 0.8892
Accuracy (Test Set): 0. 8670

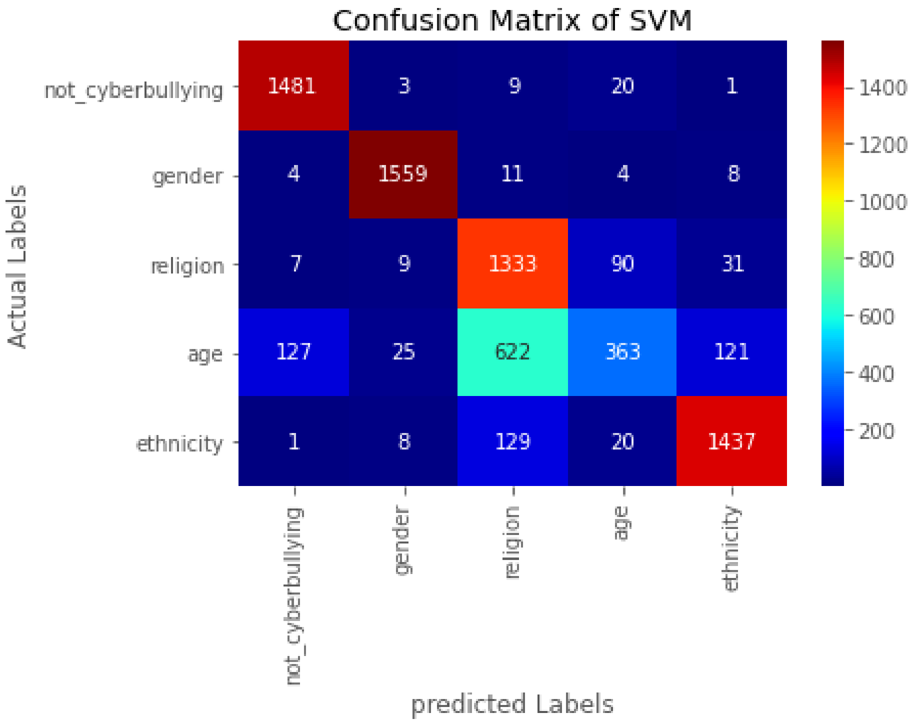


Figure *9***:** Cyberbullying Classification - SVM Confusion Matrix

6.3.3. Random Forest

After the complete training of the random forest model, random forest showed the 0.87% test accuracy on test set. The complete classification report of random forest model on test set in presented in Table 8. For the visual understanding of the trained model, the confusion matrix of the model on test set is also presented in Figure 10.

Table *8*: Cyberbullying Classification - RF Report

Classification Report for RF Model				
	precision	recall	f1-score	support
Not-Cyberbullying	0.86	0.99	0.92	1588
Gender	0.97	0.94	0.95	1489
Religion	0.95	0.74	0.85	1446
Age	0.65	0.72	0.68	1300
Ethnicity	0.94	0.93	0.93	1600
accuracy			0.87	7423
macro avg	0.87	0.86	0.86	7423
weighted avg	0.88	0.87	0.87	7423

Accuracy (Train Set): 0.8813
Accuracy (Test Set): 0. 8707

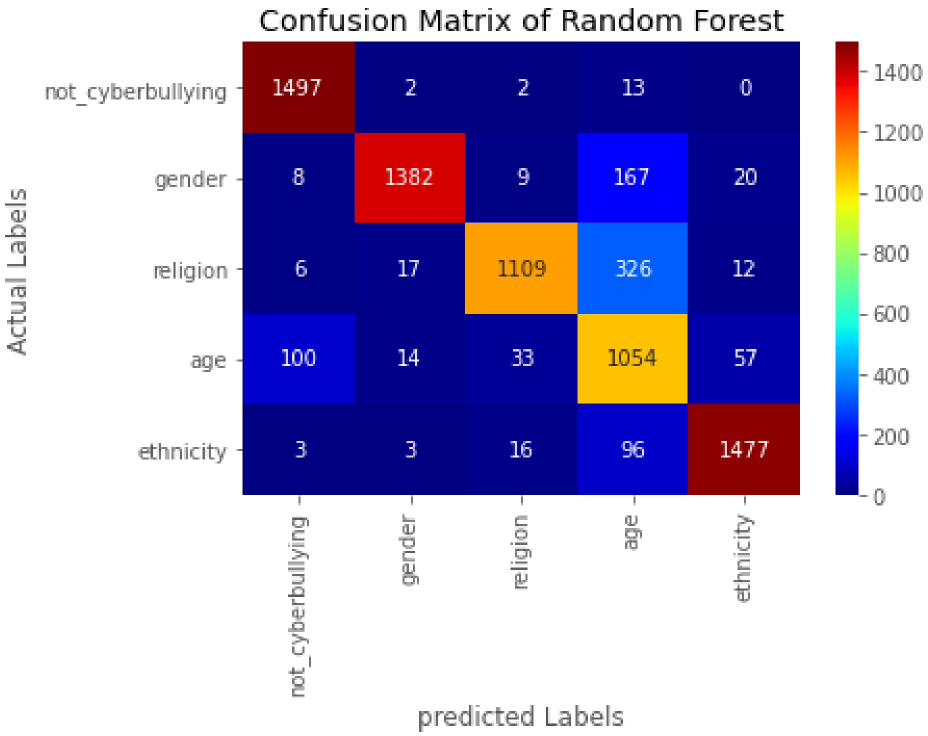


Figure *10***:** Cyberbullying Classification - RF Confusion Matrix

6.3.4. KNN Model

KNN model showed the 0.82% test accuracy on test set after the complete training on train set. The complete classification report of KNN model is shown in Table 9. The confusion matrix of KNN model is also shown in Figure 11.

Table *9***:** Cyberbullying Classification – KNN Report

Classification Report for KNN Model

	precision	recall	f1-score	support
Not-Cyberbullying	0.92	0.89	0.90	1588
Gender	0.96	0.92	0.94	1489
Religion	0.84	0.76	0.79	1446
Age	0.56	0.76	0.64	1300
Ethnicity	0.93	0.79	0.85	1600
accuracy			0.83	7423
macro avg	0.84	0.82	0.83	7423

weighted avg	0.85	0.83	0.83	7423

Accuracy (Train Set): 0.8408
Accuracy (Test Set): 0.8269

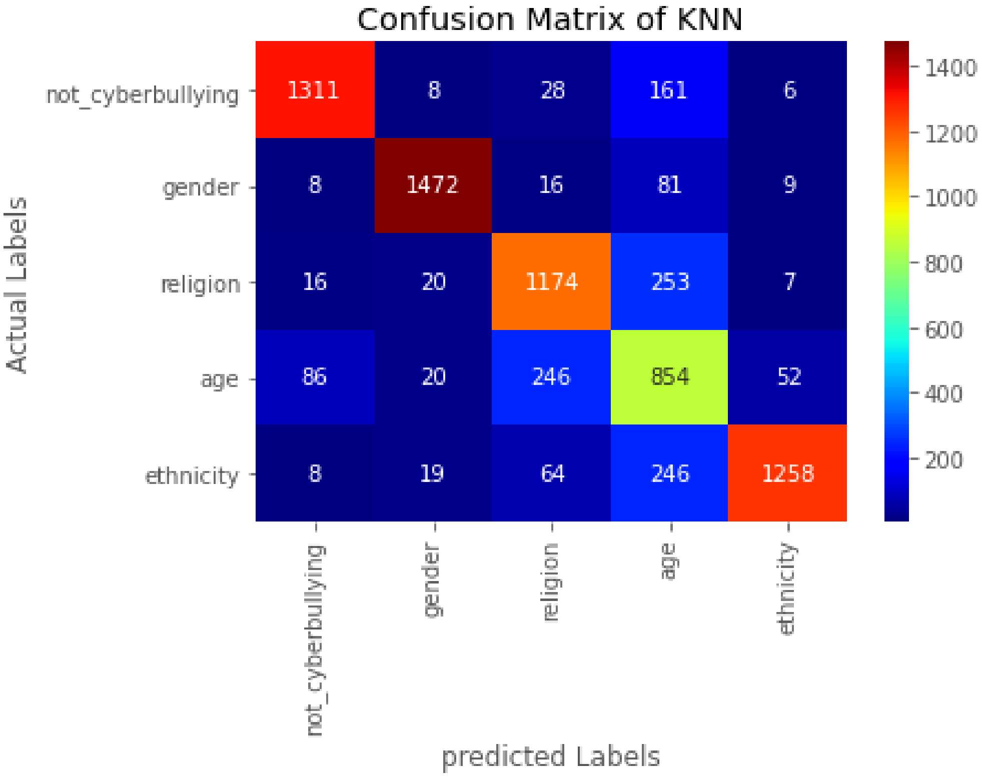


Figure *11*: Cyberbullying Classification - KNN Confusion Matrix

6.3.5. Customize Model (NN)

We train the customize model with samples training and validation set. Training set was used for the training of the model, while the validation set was used to validate the model during training. After the complete training of the model, the trained customized model was evaluated by using the evaluation measures on test set. Customized NN model showed the 0.90% accuracy on test set. The complete classification report of the trained model on test set in shown in Table 10.

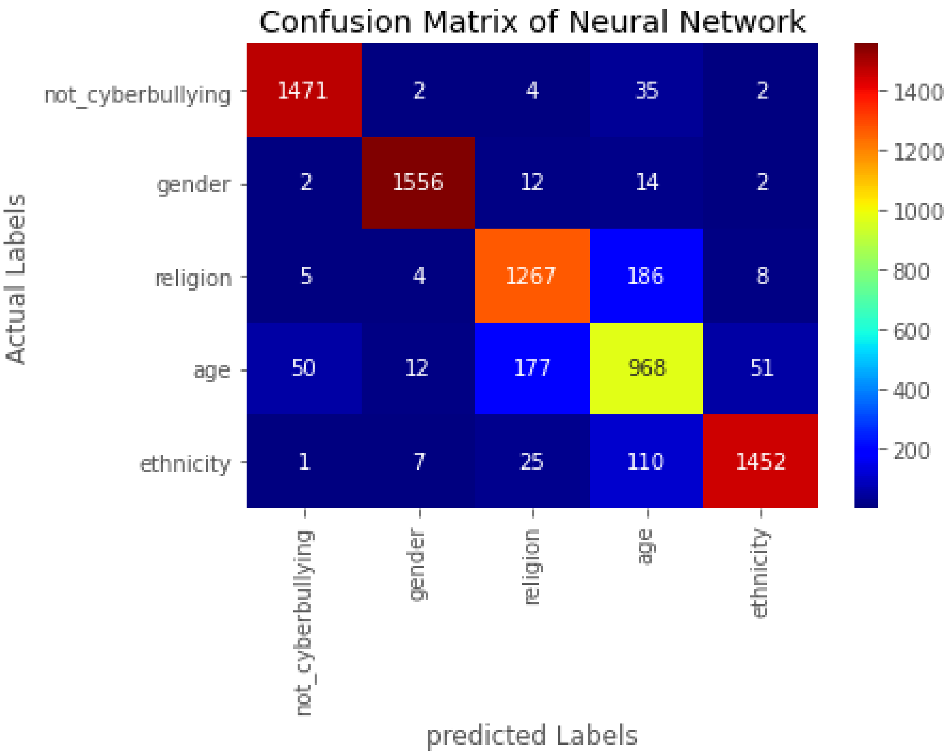
Table *10***:** Cyberbullying Classification – Customize NN Report

Classification Report for NN Model

	precision	recall	f1-score	support
Not-Cyberbullying	0.96	0.97	0.97	1588
Gender	0.99	0.98	0.98	1489
Religion	0.91	0.78	0.84	1446
Age	0.71	0.83	0.76	1300
Ethnicity	0.95	0.92	0.94	1600
accuracy			0.90	7423
macro avg	0.90	0.90	0.90	7423
weighted avg	0.91	0.90	0.90	7423

Accuracy (Train Set): 0.9740
Accuracy (Test Set): 0.9019

We also plot the confusion metrics for the clear understanding of the model predictions. The confusion metrics of the customize NN on test set in shown in Figure 12.



	not_cyberbullying	gender	religion	age	ethnicity
not_cyberbullying	1471	2	4	35	2
gender	2	1556	12	14	2
religion	5	4	1267	186	8
age	50	12	177	968	51
ethnicity	1	7	25	110	1452

Figure *12***:** Cyberbullying Classification - NN Confusion Matrix

Evaluation and Conclusion

6.3.6. Comparative Study

Lastly, we perform the comparative study on the results of the trained model. After the training of the proposed model, the results were compiled on test set. The comparative view of all trained models is presented in Table 11.

Table *11***:** Cyberbullying Classification – Comparative Classification Report

Metrics	SVM	RF Classifier	KNN	NN
accuracy	0.867035	0.870672	0.826889	0.901926
precision	0.858708	0.873465	0.840209	0.901866
recall	0.859307	0.86361	0.824602	0.898107
f1-score	0.858877	0.864581	0.827423	0.89798

Further, the comparison bar chart for all evaluation measure was plotted for the comparison of the results. The comparative bar chart of all trained models is presented in Figure 13.

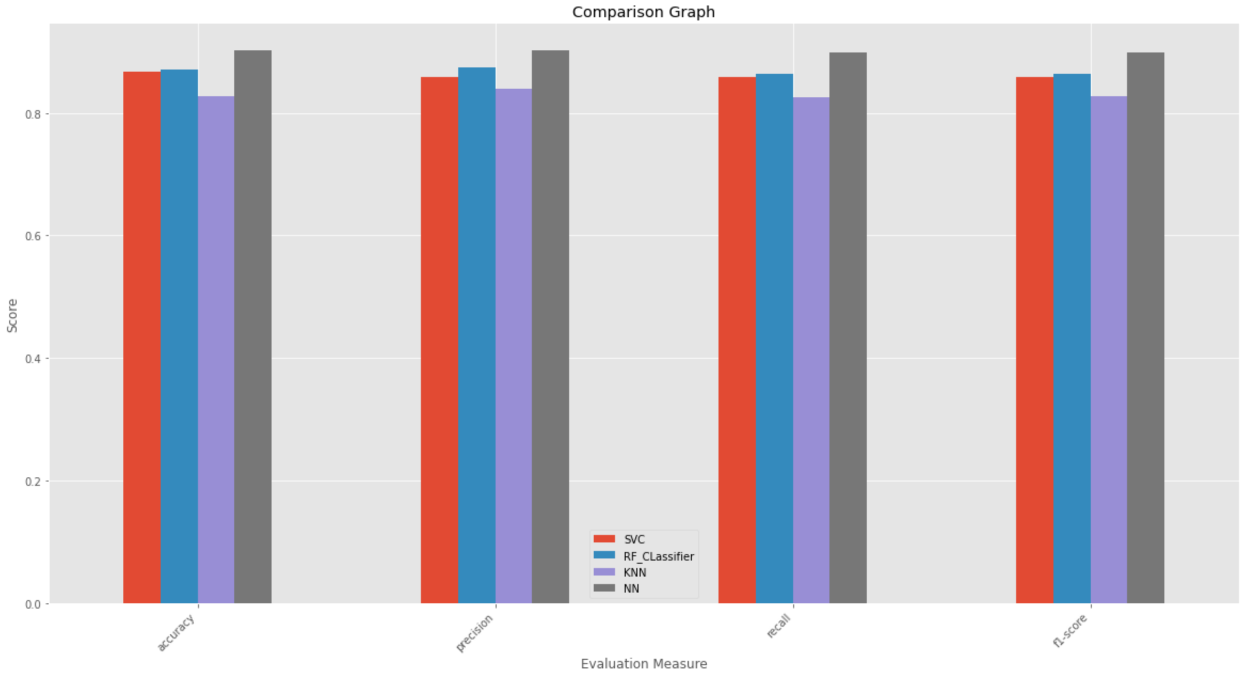


Figure *13***:** Evaluation Scores of trained models.

7. Conclusion

In this proposed study, the aim objective was to classify the cyberbullying types using tweets data. The more accurate classification of cyberbullying types will help to reduce this by different NLP related techniques. For the classification of cyberbullying types, we trained numerous machine learning models. Lastly, we also proposed a customized neural network based deep learning model of the classification of cyberbullying types. By using the different evaluation measure like accuracy, precision, recall and f1-score, we evaluate all the trained models. As the selected dataset for the proposed study was the approximately class balanced dataset, we select the accuracy as our base evaluation measure. On the basis of all models accuracy score, we discover that the customized deep learning model is more robust for cyberbullying classification related to other machine learning models. Our proposed neural network model showed the 90% test accuracy that was the highest accuracy score compare to the other trained models. By analyzing the all-evaluation measures of proposed model, we hypothesized that our model is robust enough to deploy in real world environment for the classification of cyberbullying types using tweets data.

8. References

Acı, Ç. İ., Çürük, E., & Saraç Eşsiz, E. (2019). AUTOMATIC DETECTION OF CYBERBULLYING IN FORMSPRING.ME, MYSPACE AND YOUTUBE SOCIAL NETWORKS. *Turkish Journal of Engineering*, *3*(4), 168–178. https://doi.org/10.31127/TUJE.554417

Akhter, A., Acharjee, U. K., & Polash, M. A. (2019). Cyber Bullying Detection and Classification using Multinomial Naïve Bayes and Fuzzy Logic. *I.J. Mathematical Sciences and Computing*, *4*, 1–12. https://doi.org/10.5815/ijmsc.2019.04.01

Ang, R. P., & Goh, D. H. (2010). Cyberbullying among adolescents: The role of affective and cognitive empathy, and gender. *Child Psychiatry and Human Development*, *41*(4), 387–397. https://doi.org/10.1007/S10578-010-0176-3/FIGURES/1

Automated Hate Speech Detection and the Problem of Offensive Language | Proceedings of the International AAAI Conference on Web and Social Media. (n.d.). Retrieved August 2, 2022, from https://ojs.aaai.org/index.php/ICWSM/article/view/14955

Barlińska, J., Szuster, A., & Winiewski, M. (2013). Cyberbullying among Adolescent Bystanders: Role of the Communication Medium, Form of Violence, and Empathy. *Journal of Community & Applied Social Psychology*, *23*(1), 37–51. https://doi.org/10.1002/CASP.2137

Bauman, S. (2015). Types of Cyberbullying. *Cyberbullying*, 53–58. https://doi.org/10.1002/9781119221685.CH4

Bayzick, J. (n.d.). *Detecting the Presence of Cyberbullying Using Computer Software Submitted to the faculty of Ursinus College in fulfillment of the requirements for Distinguished Honors in Computer Science*.

Bhattacharya, I., & Lindgreen, E. R. (2020). A SEMI-SUPERVISED MACHINE LEARNING APPROACH TO DETECT ANOMALIES IN BIG ACCOUNTING DATA. *ECIS 2020 Research Papers*. https://aisel.aisnet.org/ecis2020_rp/100

Bosse, T., & Stam, S. (2011). A normative agent system to prevent cyberbullying. *Proceedings - 2011 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2011*, *2*, 425–430. https://doi.org/10.1109/WI-IAT.2011.24

Buczak, A. L., & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys and Tutorials*, *18*(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Casas, J. A., del Rey, R., & Ortega-Ruiz, R. (2013). Bullying and cyberbullying: Convergent and divergent predictor variables. *Computers in Human Behavior*, *29*(3), 580–587. https://doi.org/10.1016/J.CHB.2012.11.015

Chatterjee, R., Datta, A., & Sanyal, D. K. (2019). Ensemble Learning Approach to Motor Imagery EEG Signal Classification. *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging*, 183–208. https://doi.org/10.1016/B978-0-12-816086-2.00008-4

Chatzakou, D., Leontiadis, I., Blackburn, J., de Cristofaro, E., Stringhini, G., Vakali, A., & Kourtellis, N. (2019). Detecting Cyberbullying and Cyberaggression in Social Media. *ACM Transactions on the Web (TWEB)*, *13*(3), 51. https://doi.org/10.1145/3343484

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, 71–80. https://doi.org/10.1109/SOCIALCOM-PASSAT.2012.55

Cheng, L., Li, J., Silva, Y. N., Hall, D. L., & Liu, H. (n.d.). XBully: Cyberbullying Detection within a Multi-Modal Context. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. https://doi.org/10.1145/3289600

Codina, J., Davison, B. D., Yin, D., Xue, Z., Liangjie, H., Kontostathis, A., Edwards, L., & Edu, L. (n.d.). *Content analysis in web 2.0 Related papers Det ect ion of harassment on web 2.0 A Review of Cyberbullying Det ect ion : An Overview Nurfadhlina Mohd Sharef Approaches for Mining YouTube Videos Met adat a in Cyber bullying Det ect ion Shivraj Marat he Detection of Harassment on Web 2.0*.

Cyberbullying: Experiences, Impacts and Coping Strategies as Described by Australian Young People | Youth Studies Australia. (n.d.). Retrieved August 2, 2022, from https://search.informit.org/doi/abs/10.3316/IELAPA.213627997089283

Dadvar, M., De, F., Roeland, J., & Dolf Trieschnigg, O. (2012). *Improved Cyberbullying Detection Using Gender Information*. http://www.noswearing.com/dictionary

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(3), 1–30. https://doi.org/10.1145/2362394.2362400

García-Recuero, Á. (2016). Discouraging Abusive Behavior in Privacy-Preserving Online Social Networking Applications. *WWW 2016 Companion - Proceedings of the 25th International Conference on World Wide Web*, 305–309. https://doi.org/10.1145/2872518.2888600

Görzig, A., Ólafsson, K., Gö, A., & Lafsson, K. O. ´. (2013). What Makes a Bully a Cyberbully? Unravelling the Characteristics of Cyberbullies across Twenty-Five European Countries. *Https://Doi.Org/10.1080/17482798.2012.739756*, *7*(1), 9–27. https://doi.org/10.1080/17482798.2012.739756

Havas, J., de Nooijer, J., Crutzen, R., & Feron, F. (2011). Adolescents' views about an internet platform for adolescents with mental health problems. *Health Education*, *111*(3), 164–176. https://doi.org/10.1108/09654281111123466/FULL/PDF

Hemphill, S. A., Kotevski, A., Tollit, M., Smith, R., Herrenkohl, T. I., Toumbourou, J. W., & Catalano, R. F. (2012). Longitudinal Predictors of Cyber and Traditional Bullying Perpetration in Australian Secondary School Students. *Journal of Adolescent Health*, *51*(1), 59–65. https://doi.org/10.1016/J.JADOHEALTH.2011.11.019

Hinduja, S., & Patchin, J. W. (2008). Cyberbullying: An Exploratory Analysis of Factors Related to Offending and Victimization. *Http://Dx.Doi.Org/10.1080/01639620701457816*, *29*(2), 129–156. https://doi.org/10.1080/01639620701457816

Ho, S., Kao, D., Chiu-Huang, M.-J., Li, W., Lai, C.-J., & Ankamah, B. (2019). Charged language on Twitter: A predictive model of cyberbullying to prevent victimization. *WISP 2019 Proceedings*. https://aisel.aisnet.org/wisp2019/21

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Detection of Cyberbullying Incidents on the Instagram Social Network. *MobiSys*, 2014. https://doi.org/10.48550/arxiv.1503.03909

Ibn Rafiq, R., Hosseinmardi, H., Han, R., Lv, Q., & Mishra, S. (n.d.). Scalable and Timely Detection of Cyberbullying in Online Social Networks. *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 10. https://doi.org/10.1145/3167132

Irena, B., & Setiawan, E. B. (2020). Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(4), 711–716. https://doi.org/10.29207/RESTI.V4I4.2125

Jacobs, N. C., Völlink, T., Dehue, F., & Lechner, L. (2014). Online Pestkoppenstoppen: Systematic and theory-based development of a web-based tailored intervention for adolescent cyberbully victims to combat and prevent cyberbullying. *BMC Public Health*, *14*(1), 1–16. https://doi.org/10.1186/1471-2458-14-396/FIGURES/7

Joachims, T. (1998). *Text categorization with Support Vector Machines: Learning with many relevant features*. 137–142. https://doi.org/10.1007/BFB0026683

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms: A Survey. *Information 2019, Vol. 10, Page 150*, *10*(4), 150. https://doi.org/10.3390/INFO10040150

Leon-Paredes, G. A., Palomeque-Leon, W. F., Gallegos-Segovia, P. L., Vintimilla-Tapia, P. E., Bravo-Torres, J. F., Barbosa-Santillan, L. I., & Paredes-Pinos, M. M. (2019). Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language. *IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies, CHILECON 2019*. https://doi.org/10.1109/CHILECON47746.2019.8987684

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*. https://doi.org/10.48550/arxiv.1407.7502

Mangaonkar, A. (2017). *COLLABORATIVE DETECTION OF CYBERBULLYING BEHAVIOR IN TWITTER DATA*.

Mangaonkar, A., Hayrapetian, A., & Raje, R. (2015). Collaborative detection of cyberbullying behavior in Twitter data. *IEEE International Conference on Electro Information Technology*, *2015-June*, 611–616. https://doi.org/10.1109/EIT.2015.7293405

Mccallum, A., & Nigam, K. (n.d.). *A Comparison of Event Models for Naive Bayes Text Classification*.

Miller, K. (2016). Cyberbullying and Its Consequences: How Cyberbullying Is Contorting the Minds of Victims and Bullies Alike, and the Law's Limited Available Redress. *Southern California Interdisciplinary Law Journal*, *26*. https://heinonline.org/HOL/Page?handle=hein.journals/scid26&id=395&div=&collection=

Modeling the Detection of Textual Cyberbullying | Proceedings of the International AAAI Conference on Web and Social Media. (n.d.). Retrieved August 2, 2022, from https://ojs.aaai.org/index.php/ICWSM/article/view/14209

Nahar, V., Li, X., Zhang, H. L., & Pang, C. (2014). Detecting cyberbullying in social networks using multi-agent system. *Web Intelligence and Agent Systems: An International Journal*, *12*(4), 375–388. https://doi.org/10.3233/WIA-140301

Novalita, N., Herdiani, A., Lukmana, I., & Puspandari, D. (2019). Cyberbullying identification on twitter using random forest classifier. *Journal of Physics: Conference Series*, *1192*(1), 012029. https://doi.org/10.1088/1742-6596/1192/1/012029

Olweus, D. (2012). Cyberbullying: An overrated phenomenon? *Http://Dx.Doi.Org/10.1080/17405629.2012.682358*, *9*(5), 520–538. https://doi.org/10.1080/17405629.2012.682358

Patchin, J. W., & Hinduja, S. (2010). Traditional and Nontraditional Bullying Among Youth: A Test of General Strain Theory: *Https://Doi.Org/10.1177/0044118X10366951*, *43*(2), 727–751. https://doi.org/10.1177/0044118X10366951

Pawar, R., Agrawal, Y., Joshi, A., Gorrepati, R., & Raje, R. R. (2018). Cyberbullying Detection System with Multiple Server Configurations. *IEEE International Conference on Electro Information Technology*, *2018-May*, 90–95. https://doi.org/10.1109/EIT.2018.8500110

Pawar, R., & Raje, R. R. (2019). Multilingual cyberbullying detection system. *IEEE International Conference on Electro Information Technology*, *2019-May*, 040–044. https://doi.org/10.1109/EIT.2019.8833846

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and Monitoring Hate Speech in Twitter. *Sensors 2019, Vol. 19, Page 4654*, *19*(21), 4654. https://doi.org/10.3390/S19214654

Rafiq, R. I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., & Mattson, S. A. (2015). Careful what you share in six seconds: Detecting cyberbullying instances in Vine. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, 617–622. https://doi.org/10.1145/2808797.2809381

Rahman, S., Irfan, M., Raza, M., Ghori, K. M., Yaqoob, S., & Awais, M. (2020). Performance Analysis of Boosting Classifiers in Recognizing Activities of Daily Living. *International Journal of Environmental Research and Public Health 2020, Vol. 17, Page 1082*, *17*(3), 1082. https://doi.org/10.3390/IJERPH17031082

Raisi, E., Tech, V., & Huang, B. (2016). *Cyberbullying Identification Using Participant-Vocabulary Consistency*. https://doi.org/10.48550/arxiv.1606.08084

Raza, M. O., Memon, M., Bhatti, S., & Bux, R. (2020). Detecting Cyberbullying in Social Commentary Using Supervised Machine Learning. *Advances in Intelligent Systems and Computing*, *1130 AISC*, 621–630. https://doi.org/10.1007/978-3-030-39442-4_45/COVER

Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. *Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011*, *2*, 241–244. https://doi.org/10.1109/ICMLA.2011.152

Rybnicek, M., Poisel, R., & Tjoa, S. (2013). Facebook watchdog: A research agenda for detecting online grooming and bullying activities. *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, 2854–2859. https://doi.org/10.1109/SMC.2013.487

Sahlgren, M., Isbister, T., & Olsson, F. (2018). Learning Representations for Detecting Abusive Language. *2nd Workshop on Abusive Language Online - Proceedings of the Workshop, Co-Located with EMNLP 2018*, 115–123. https://doi.org/10.18653/V1/W18-5115

Salminen. (n.d.). Retrieved August 2, 2022, from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/viewPaper/17885

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. gyo, Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-Centric Computing and Information Sciences*, *10*(1), 1–34. https://doi.org/10.1186/S13673-019-0205-6/FIGURES/8

Sampasa-Kanyinga, H., Roumeliotis, P., & Xu, H. (2014). Associations between Cyberbullying and School Bullying Victimization and Suicidal Ideation, Plans and Attempts among Canadian Schoolchildren. *PLOS ONE*, *9*(7), e102145. https://doi.org/10.1371/JOURNAL.PONE.0102145

Smith, P. K. (2012). Cyberbullying: Challenges and opportunities for a research program—A response to Olweus (2012). *Http://Dx.Doi.Org/10.1080/17405629.2012.689821*, *9*(5), 553–558. https://doi.org/10.1080/17405629.2012.689821

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, *49*(4), 376–385. https://doi.org/10.1111/J.1469-7610.2007.01846.X

Snakenborg, J., Acker, R. van, & Gable, R. A. (2011). Cyberbullying: Prevention and Intervention to Protect Our Children and Youth. *Https://Doi.Org/10.1080/1045988X.2011.539454*, *55*(2), 88–95. https://doi.org/10.1080/1045988X.2011.539454

Survey: Majority of cyber bullying incidents go unreported | TopNews. (n.d.). Retrieved August 2, 2022, from https://topnews.in/survey-majority-cyber-bullying-incidents-go-unreported-2375044

Tarwani, S., Jethanandani, M., & Kant, V. (2019). Cyberbullying Detection in Hindi-English Code-Mixed Language Using Sentiment Classification. *Communications in Computer and Information Science*, *1046*, 543–551. https://doi.org/10.1007/978-981-13-9942-8_51/COVER

Tenenbaum, L. S., Varjas, K., Meyers, J., & Parris, L. (2011). Coping strategies and perceived effectiveness in fourth through eighth grade victims of bullying: *Http://Dx.Doi.Org/10.1177/0143034311402309*, *32*(3), 263–287. https://doi.org/10.1177/0143034311402309

The History of Social Media and its Impact on Business - ProQuest. (n.d.). Retrieved August 2, 2022, from https://www.proquest.com/openview/f828806820e0b99fcbda9c765788e137/1?cbl=25565&pq-origsite=gscholar&parentSessionId=3I%2FpheP28zbwlwCZQ2Z1nXAfMJU%2BNyCs%2FmwWd0o9o38%3D

Tu, C., Liu, H., & Xu, B. (2017). AdaBoost typical Algorithm and its application research. *MATEC Web of Conferences*, *139*, 00222. https://doi.org/10.1051/MATECCONF/201713900222

Vaillancourt, T., Faris, R., & Mishna, F. (2017). Cyberbullying in Children and Youth: Implications for Health and Clinical Practice. *Canadian Journal of Psychiatry*, *62*(6), 368–373. https://doi.org/10.1177/0706743716684791

van Hee, C., Jacobs, G., Emmery, C., DeSmet, B., Lefever, E., Verhoeven, B., de Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PLOS ONE*, *13*(10), e0203794. https://doi.org/10.1371/JOURNAL.PONE.0203794

van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., de Pauw, G., Daelemans, W., & Hoste, V. (n.d.). *Detection and Fine-Grained Classification of Cyberbullying Events*. 672–680. Retrieved August 2, 2022, from https://www.gnu.org/software/wget

Waseem, Z., & Hovy, D. (n.d.). *Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter*. 88–93. Retrieved August 2, 2022, from http://github.com/zeerakw/hatespeech

Ybarra, M. L., Mitchell, K. J., Wolak, J., & Finkelhor, D. (2006). Examining Characteristics and Associated Distress Related to Internet Harassment: Findings From the Second Youth Internet Safety Survey. *Pediatrics*, *118*(4), e1169–e1177. https://doi.org/10.1542/PEDS.2006-0815

Zhang, H. (2011). EXPLORING CONDITIONS FOR THE OPTIMALITY OF NAÏVE BAYES. *Http://Dx.Doi.Org/10.1142/S0218001405003983*, *19*(2), 183–198. https://doi.org/10.1142/S0218001405003983

Zinovyeva, E., Härdle, W. K., & Lessmann, S. (2020). Antisocial online behavior detection using deep learning. *Decision Support Systems*, *138*, 113362. https://doi.org/10.1016/J.DSS.2020.113362

9. Appendix

9.1. Import Libraries

```
import os
import re
import glob
import math
import string
import statistics
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from tqdm import tqdm # Progress Bar

# nltk toolkit
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.stem.porter import PorterStemmer

# Sklearn toolkit
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.pipeline import make_pipeline, Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.metrics import confusion_matrix,classification_report, accuracy_score,precision_score,
recall_score, f1_score


# tensorflow toolkit

Import tensorflow as tf

import tf.python.keras.backend as K

from tf.keras.utils import to_categorical

from tf.keras.models import Sequential, Model

from tf.keras.callbacks import ModelCheckpoint, ReduceLROnPlateau, EarlyStopping

from tf.keras.layers import Dense, Activation, Dropout, BatchNormalization, Input, Embedding, LSTM
```

9.2. Preprocessing

```
#Clean emojis from text

def remove_emoji(tweet):

   return re.sub(emoji.get_emoji_regexp(), r"", tweet) # remove emoji


#Remove punctuations, links, stopwords, mentions and \r\n new line characters

def strip_all_entities(tweet):

   tweet = tweet.replace('\n', '').replace('\r', ' ').lower() # remove format specifier and convert to
lowercase

   tweet = re.sub(r"(?:\@|https?\://)\S+", "", tweet) # remove URL’s

   tweet = re.sub(r'[^\x00-\x7f]',r'', tweet)

   banned_words = string.punctuation

   table = string.maketrans('', '',     banned_words)

   tweet = tweet.translate(table)

   tweet = [w for w in tweet.split() if w not in stop_words]

   tweet = “ “.join(tweet)

   tweet = “ “.join(w for w in tweet.split() if len(word) < 12) # remove words longer than 12 characters

   return tweet


# remove helping words
```

```
def remove_HW(tweet):
    tweet = re.sub(r"\'re", " are ", tweet)
    tweet = re.sub(r"can\'t", " can not ", tweet)
    tweet = re.sub(r"\'s", " is ", tweet)
    tweet = re.sub(r"\'d", " would ", tweet)
    tweet = re.sub(r"\'ll", " will ", tweet)
    tweet = re.sub(r"\'t", " not", tweet)
    tweet = re.sub(r"\'ve", " have", tweet)
    tweet = re.sub(r"\'m", " am", tweet)
    return tweet

def remove_hash_tags(tweet):
    tweet = " ". join (w. strip() for w in re. split('#(?!(?:hashtag)\b)[\w-]+(?=(?:\s+#[\w-]+)*\s*$)', tweet))
    tweet = " ". join (w. strip() for w in re.split('#|_', tweet))
    return tweet

def filter_special_chars(tweet):
    words_array = []
    for w in tweet. split (' '):
        if ('&' in w) or ('$' in w):
            words_array.append('')
        else:
            words_array.append(w)
    return ' '. join (words_array)

#Remove extra consecutive spaces
def remove_extra_spaces(tweet):
    return re. sub("\s\s+" , " ", tweet)

# Words Stemming
```

```
def text_stemmer(tweet):
    tokenized_words = nltk.word_tokenize(tweet)
    stemmer = PorterStemmer()
    return ' '. join([stemmer.stem(w) for w in tokenized_words])

# Words Lemmatization
def text_lemmatize(tweet):
    tokenized_words = nltk.word_tokenize(tweet)
    lemmatizer = WordNetLemmatizer()
    return ' '. join([ lemmatizer.lemmatize(w) for w in tokenized_words])

def preprocess(tweet):
    tweet = remove_emoji(tweet)
    tweet = remove_HW (tweet)
    tweet = strip_all_entities(tweet)
    tweet = remove_hashtags(tweet)
    tweet = filter_special_chars(tweet)
    tweet = remove_extra_spaces(tweet)
    tweet = text_stemmer(tweet)
   tweet = text_lemmatizer(tweet)
    return tweet

cleaned_texts = []
for tweet in df[“text”] :
    cleaned_texts. append(preprocess(tweet))
df['text_clean'] = cleaned_texts
```

9.3. Drop Duplicate Samples

```
df.drop_duplicates("text_clean", inplace=True)

print(df["text_clean"].duplicated().sum())
```

9.4. Label Encoding

```
le = LabelEncoder()
df['encoded_label'] = le.fit_transform(df['label'])
```

9.5. Feature Extraction

```
features = df['text_clean']
label = df['encoded_label']

tfidf_vectorizer = TfidfVectorizer(max_features=200)
transform_features = tfidf_vectorizer.fit_transform(features)
features_df = pd.DataFrame(data=transform_features_tf.todense(), columns=vectorizer.get_feature_names_out())
```

9.6. Train Test Split

```
def split_df(features, label, split_index):
    test_ratio = 0.2
    if split_index==0:
        trainX, testX, trainY, testY = train_test_split (features, label, test_size = test_ratio)
        return trainX, testX, trainY, testY
    if split_index==1:
        train_X, test_X, train_y, test_y = train_test_split(features, label, test_size = test_ratio)
        train_X, val_X, train_y, val_y = train_test_split(train_X, train_y, test_size=0.1)
        return train_X, test_X, val_X, train_y, test_y, val_y
index = 1
train_X, test_X, val_X, train_y, test_y, val_y = split_df(features_df, label, index)
```

9.7. Machine Learning Models Code

9.7.1. Support Vector Machine

```
from sklearn.svm import SVC as SVM
svm = SVM(max_iter=200)
svm. fit (train_X, train_y)
preds = svm.predict(test_X)
```

9.7.2. Random Forest

```
from sklearn.ensemble import RandomForestClassifier as Random_Forest
RF = Random_Forest (bootstrap=100, max_depth=5)
RF. fit(train_X, train_y)
preds = RF.predict(test_X)
```

9.7.3. K-Nearest Neighbor

```
from sklearn.neighbors import KNeighborsClassifier as k_Neighbor
knn = k_Neighbor (n_neighbors=5,)
knn. fit(train_X, train_y)
preds = knn. predict(test_X)
```

9.8. Deep Learning Code

9.8.1. Customized Neural Network

```
batch_size = 32
activation_func = “relu”
model = Sequential()
model.add(Dense(1000, input_shape=(200,), activation= activation_func))
model.add(Dense(500, activation= activation_func))
model.add(Dense(250, activation= activation_func))
model.add(Dense(125, activation= activation_func))
model.add(Dense(50, activation= activation_func))
model.add(Dense(10, activation= activation_func))
model.add(Dense(5,activation= activation_func))
model.compile(loss = 'categorical_crossentropy', optimizer='adam',metrics = ['accuracy'])

y_train1 = to_categorical(train_y, 5)
y_test1 = to_categorical(test_y, 5)
y_val1 = to_categorical(val_y, 5)
model.fit(train_X, y_train1, validation_data=(val_X, y_val1), epochs = 15, batch_size=batch_size)
```

```
# make prediction using customized NN model

preds = model. predict(test_X)

# calculate accuracy and other evaluation measures

y_pred = [np.argmax(prediction) for prediction in preds]
```