##### A Project report on

###### **Detection of Cyberbullying on Social Media using Machine Learning Approach in Python**

###### A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

**Bachelor of Technology**

**in**

**Computer Science and Engineering**

Submitted by

B. SWATHI

(19H51A0531)

G. DIANA

(19H51A0537)

D. KEERTHI

(19H51A0538)

Under the esteemed guidance of

Mr.B.K. Chinna Maddileti

(Assistant Professor)



**Department of Computer Science and Engineering**

**CMR COLLEGE OF ENGINEERING & TECHNOLOGY**

(An Autonomous Institution under UGC & JNTUH, Approved by AICTE, Permanently Affiliated to JNTUH, Accredited by NBA.)

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

#### 2019- 2023

**CMR COLLEGE OF ENGINEERING & TECHNOLOGY**

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**



#### CERTIFICATE

This is to certify that the Major Project Phase-1 report entitled **"Detection of Cyberbullying on Social Media using Machine Learning Approach in Python"** being submitted by B. Swathi (19H51A0531), G. Diana (19H51A0537), D. Keerthi (19H51A0538) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

###### The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

**Mr. B.K. Chinna Maddileti Dr. Siva Skandha Sanagala**

**Assistant Professor Associate Professor and HOD**

**Dept. of CSE Dept. of CSE**

#### ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express our heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Mr.B.K. Chinna Maddileti,** **Assistant Professor,** Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete our project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academic, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the Teaching & Non- teaching staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Mr. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

B.Swathi 19H51A0531

G.Diana 19H51A0537

D.Keerthi 19H51A0538

**TABLE OF CONTENTS**

**CHAPTER**

**NO. TITLE PAGE NO.**

LIST OF FIGURES iii

LIST OF TABLES iv

ABSTRACT v

**1** **INTRODUCTION** 1

1.1 Problem Statement 3

1.2 Research Objective 3

1.3 Project Scope and Limitations 3

**2** **BACKGROUND WORK** 5

2.1. Cyberbullying Severity Detection: A Machine Learning approach 6

2.1.1.Introduction 6

2.1.2.Merits and Demerits 6

2.1.3.Implementation of Cyberbullying Severity Detection:

A Machine Learning approach 7

2.2. Detection of Cyber-Aggressive Comments on Social Media

Networks: A Machine Learning and Text mining approach 11

2.2.1.Introduction 11

2.2.2.Merits and Demerits 11

2.2.3.Implementation of Detection of Cyber-Aggressive

Comments on Social Media Networks: A Machine

Learning and Text mining approach 11

2.3. Social Media Cyberbullying Detection using Machine Learning 16

2.3.1.Introduction 16

2.3.2.Merits and Demerits 16

2.3.3.Implementation of Social Media Cyberbullying Detection

using Machine Learning 16

**3 RESULTS AND DISCUSSION** 20 3.1. Comparison of Existing Solutions 21

3.2. Data Collection and Performance Metrics 21

**4** **CONCLUSION** 25

**5** **REFERENCES** 27

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **FIGURE**  **NO.** | **TITLE** | **PAGE** **NO.** |
| 1 | Cyberbullying Severity Detection Framework | 7 |
| 2 | Detection of Cyber-Aggressive Comments on Social Media Networks framework | 12 |
| 3 | Social Media Cyberbullying Detection using Machine Learning framework | 17 |
| 4 | True classification summary for test data | 23 |
| 5 | Comparison between the Best Classifiers in Terms of Accuracy | 24 |
| 6 | Comparison between the Best Classifiers in Terms of F-Measure | 24 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **FIGURE**  **NO.** | **TITLE** | **PAGE** **NO.** |
| 1 | Annotated tweets by category | 7 |
| 2 | Cyberbullying tweets categorized per severity level | 8 |
| 3 | Summary of training performance of different models with different feature extraction and selection methods | 15 |
| 4 | Classifies performance under various settings in multi-class classification | 22 |
| 5 | Classifier performance under various settings in binary classification | 22 |

**ABSTRACT**

Cyberbullying is a major problem encountered on internet that affects teenagers and also adults.It has lead to mishappenings like suicide and depression.Regulation of content on social media platforms has become a growing need. Our proposed model will be using data from two different forms of cyberbullying i.e.,Hate speech tweets from Twitter and comments based on personal attacks from Wikipedia forums to build a model based on detection of Cyberbullying in text data using Natural Language Processing and Machine learning. Three methods for feature extraction and six classifiers are studied to outline the best approach.

# **CHAPTER 1**

**INTRODUCTION**

1. **INTRODUCTION**

The real world stopped and the reel world started what could be more appropriate to describe the life of today’s generation. The reel world works as a web-connected people from all over the globe. The medium which should have been used to connect and communicate with people has started to become a place where people, especially teens and young adults bully one another. Though the positive side where social media helps people to connect, it also exposes them to a threatening situation like aggressive cyberbullying. Quite half of teenagers have reported being the victims of cyberbullying. Moreover, research has found that links between experiences of cyberbullying and negative outcomes are decreased performance in school, violent behavior, and potential devastating psychological effects like depression, low self-esteem, suicide ideation which can have future effects within the longer-term lifetime of victims. Incidents of cyberbullying with extreme consequences like suicide are now routinely reported within the favored press. The results of cyberbullying has on its victims and it is rapidly spread among high school students, there is a need for research to know how cyberbullying occurs in social network today to those effective techniques are often developed to automatically predict cyberbullying. It reports that experts within the sector of cyberbullying could favor automatic detection of cyberbullying on social media networking sites and propose effective follow-up techniques and methods. So in the era of widespread usage of social networking, cyber-attacks are also on the rise. Cyberbullying is at the forefront of this, which is the act of insulting or defaming a person personally. Moreover, it includes harassment and flaming communications. Through text messaging, comments, etc. by cyber communication, cyberbullying can occur. Different models are generated to resolve this issue by detecting cyberbullying and a lot of detection has been done using many techniques. But still, cyberbullying is seen as a major issue on many social media.

* 1. **Problem Statement**

Cyberbullying frequently leads to serious mental and physical distress, particularly for children, and even sometimes force them to attempt suicide. Therefore, the identification of bullying text or message on social media has gained a growing amount of attention.

Our work focuses on data from two different forms of cyberbullying i.e.,Hate speech tweets from Twitter and comments based on personal attacks from Wikipedia forums to build a model based on detection of Cyberbullying in text data using Natural Language Processing and Machine learning. Three methods for Feature extraction and six classifiers are studied to outline the best approach.

* 1. **Research Objective**
* The main aim of detecting cyberbullying will help to improve manual monitoring for cyberbullying on social networks.
* In this project we fetch the tweets from twitter accounts and comments based on personal attacks from wikipedia forums then preprocess the tweets and comments by necessary algorithms, and then we detect whether it is an offensive text/message or not.
  1. **Project Scope and Limitations**

Cyberbullying is the use of electronic communication to bully a person by sending harmful messages using social media, instant messaging or through digital messages. Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying. So to overcome these issues, detection of cyberbullying is very important these days, which will help to stop menace on social media networks.

We could not perform depth analysis in relation to user’s behavior because the dataset we used for this study did not provide any information (i.e. time of the tweet, favorite, followers etc.) other than just content (tweets). Moreover, we could have performed the meta-analysis on the effects of cyberbullying severity, however, the studies that we reviewed did not provide necessary information that would enable this

type of analysis. Furthermore, present study is only focused on twitter and wikipedia. Other social network platforms (such as Facebook, YouTube etc) need to be investigated to see the same pattern of cyberbullying severity.

**CHAPTER 2**

**BACKGROUND WORK**

1. **BACKGROUND WORK**

There are three existing systems that deal with the detection of cyberbullying. They are Cyberbullying Severity Detection: A Machine Learning approach, Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach, and Social Media Cyberbullying Detection using Machine Learning

**2.1 CYBERBULLYING SEVERITY DETECTION: A MACHINE LEARNING APPROACH**

1. **INTRODUCTION**

A supervised machine learning technique for cyberbullying detection and multiclass categorization is proposed. They applied sentiment, embedding and lexicon features with PMI-semantic orientation and then applied the selected features to Decision Tree, SVM, KNN, Naive Bayes, and Random Forest algorithms. The developed model classified the Twitter contents as cyberbullying or non-cyber bullying along with the severity level as low, medium, high, or none. Lexicons related to the five topics (sexuality, racism, physical-appearance, intelligence and politics) were utilized to annotate tweets.

1. **MERITS AND DEMERITS**

*Merits:*

* Supervised machine learning algorithms have an ability to collect data and produce output from the previous experience
* The existing method to detect cyberbullying behavior in binary classification performs better than several feature engineered techniques
* Feature selection contributes to boosting prediction accuracy by reducing dimensionality of the dataset and used to yield improved results

*Demerits:*

* Decision boundary might be overstrained if training set doesn't have enough examples
* Meta-analysis is not performed because the studies that were reviewed did not provide necessary information that would enable this type of analysis

1. **IMPLEMENTATION OF CYBERBULLYING SEVERITY DETECTION: A MACHINE LEARNING APPROACH**

All steps of the existing system framework are presented in Fig 1 and discussed in the following section.

pone.0240924.g001

Fig 1. Cyberbullying Severity Detection Framework

***Methodology:***

*Data collection step*

* Annotated dataset collected by <https://github.com/Mrezvan94/Harassment-Corpus/blob/master/Harassment%20Lexicon.csv> is used for this project.
* Out of total 50,000 collected tweets, 24,189 tweets were annotated. Three indigenous English-speaking annotators subsequently determined whether or not a particular tweet is a) harassing with respect to the type of harassment content and b) allocated one of three labels “yes”, “no”, and “other”.
* Dataset was already categorized into different topics of harassment content: i) sexual, ii) racial, iii) appearance-related, iv) intelligence and v) political

pone.0240924.t001

Table 1: Annotated tweets by category

* Table 1 represents the binary classification of the mentioned topics of the dataset. Categorizing the annotated cyberbullied tweets into 4 levels; low, medium, high and non-cyberbullying.
* After that, sexual and appearance related tweets as high-level cyberbullying severity; political and racial tweets as medium-level; intelligence tweets as low-level cyberbullying severity, and all the tweets that were labelled as ‘non-cyberbullying’ in each category were consolidated into one category as non-cyberbullying tweets. This resulted in a dataset with characteristics shown in (Table 2).

pone.0240924.t002

Table 2: Cyberbullying tweets categorized per severity level

*Pre-processing step*

* The collected data is pre-processed before assigning severity levels. Tweets were converted to lower case to avoid any sparsity issue, reduced repeated letters, standardized URLs and @user mention to remove noise in the tweets.
* Tokenization was applied with Twitter-specific tokenizer and only words with minimum frequency of 10 were kept. Tokenization is the process of breaking a text corpus up into most commonly words, phrases, or other meaningful elements, which are then called tokens.
* Finally, stop-words and stemming procedures were performed before feature extraction.
* Stop words are defined as the insignificant words that appear in document which are not specific or discriminatory to the different classes.
* Stemming refers to the process of reducing words to their stems or roots.

*Feature extraction step*

* All tweets were represented with bag-of-words which is one of the most
* appropriate and quickest approaches. In this approach, text is represented by set of words and each word is treated as an independent feature.
* We applied part-of-speech (POS) tagging with Twitter-specific tagger based on the CMU TweetNLP library.

*Feature generation step*

* Document level classification and measured semantic orientation of each word in the corpus are applied.
* In the document level classification, phrases were extracted using the POS tags. Once phrases have been extracted from the dataset, then their semantic orientation in terms of either cyberbullying or non-cyberbullying are determined.
* In order to achieve this goal, the concept of pointwise mutual information (PMI) is used to calculate the semantic orientation for each word in a corpus of tweets. The PMI between two words, word1 and word2, is defined as follows:

PMI(word1, word2) = log2

*Feature engineering and selection step*

* Feature engineering is the process of generating or deriving features from raw data or corpus. Creation of additional features inferring from existing features is known as feature engineering.
* One of the most common approaches to improve cyberbullying detection is to perform feature engineering, and most common features that improve quality of cyberbullying detection classifier performance are; textual, social, user, sentiment, word embeddings features.
* Since social and user features were not available in the dataset provided by, they attempted to build features based on the textual context and their semantic orientation. As a consequence, they proposed the following features to improve cyberbullying detection:

1. Embedding Feature Vector: In this study, tweet-level feature representation using pre-trained Word2Vec embeddings were applied.Used 400 dimension embeddings of 10 million tweets from the

Edinburgh corpus.

1. Sentiment Feature Vector: SentiStrength was used to calculate positive and negative score of each tweet.
2. Lexicon Feature Vector: Multiple phrase level lexicons were applied in this study that identify positive and negative contextual polarity of sentiment expression in the dataset.
3. PMI-Semantic Orientation: Processed previously generated domain specific lexicon which contained mutual information of each word in the corpus. This PMI input approach assigns a PMI score to each word in the document. PMI-Semantic Orientation is then calculated for each document by subtracting the PMI of the target word.

*Machine learning algorithms selection step*

Choosing the best classifier is the most significant phase of the text classification pipeline. In order to select the best classifier, we tested several machine learning algorithms namely: Naïve Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest, and K-Nearest Neighbors (KNN).

*Performance evaluation*

Performance measures generally evaluate specific aspects of the performance of classification tasks and do not always present the same information. Understanding how a model performs is an essential part of any classification algorithm. There are several methods to measure performance of a classifier: example metrics are recall, precision, accuracy, F-measure, Kappa statistics and AUC. These metrics are based on “Confusion Matrix” that includes true positive (TP): the number of instances correctly labelled as belonging to the positive class; true negative (TN): negative instances correctly classified as negative; false positive (FP): instances incorrectly labelled as belonging to the class; false negative (FN): instances that are not labelled as belonging to the positive class but should have been.

**2.2 DETECTION OF CYBER-AGGRESSIVE COMMENTS ON SOCIAL MEDIA NETWORKS: A MACHINE LEARNING AND TEXT MINING APPROACH**

1. **INTRODUCTION**

First approach is to identify and filter cyber aggressive comments of social media networks in three different categories like: hate speech, offensive speech, or neither. This experiment has collected social media text data from the Data world website. 3000 text comments are there in original dataset, but 1000 text comments are selected for doing the analysis. Some very popular and useful feature extraction techniques like: bag of word, N-gram and TF-IDF are used as feature extraction methods. In this study, ensemble machine learning method like random forest, logistic regression and support vector machine are applied to perform cyber aggressive comment classification on social media networks.

1. **MERITS AND DEMERITS**

*Merits:*

* The experiment collects the social media comments and identifies the comments are agressive or not using 3 machine learning algorithms
* Quick in predicting
* Trains the data speedily

*Demerits:*

* It is important to train the model. Without proper training, it is impossible to gain expected results
* If the input data is more it takes lot of processing time & decreases the classification accuracy as well

1. **IMPLEMENTATION OF DETECTION OF CYBER-AGGRESSIVE COMMENTS ON SOCIAL MEDIA NETWORKS: A MACHINE LEARNING AND TEXT MINING APPROACH**

All steps of the Existing system framework are presented in Fig 2 and discussed in the following section.

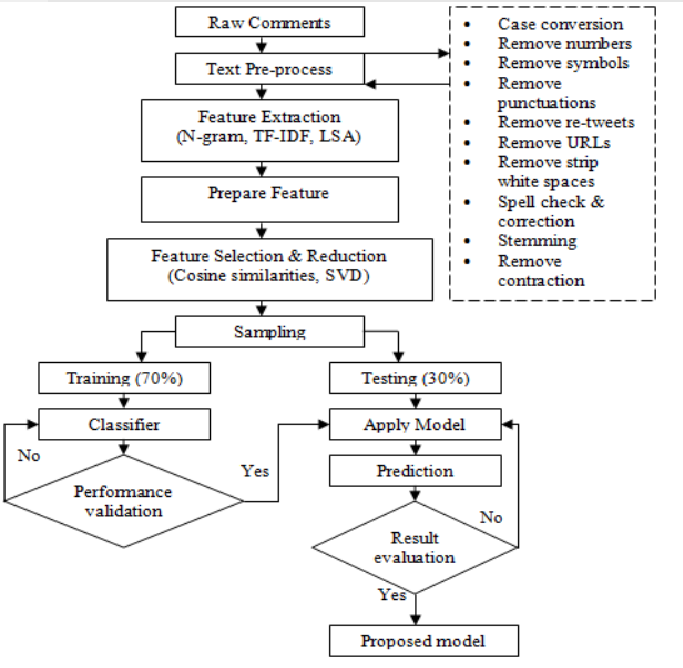


Fig 2. Detection of Cyber-Aggressive Comments on Social Media Networks Framework

***Methodology:***

*Data collection step*

To undertake the experiments we have collected social media text data of twitter comments from Data world website (https://data.world/crowdflower/hate-speech-identification). The original data set contains 2 attributes and 3000 instances. Attributes are ‘label’ and ‘twitter text’. There are three types of labeled data, like hate speech, offensive and neither. We use only 1000 text documents from total data set with the ratio of offensive– 35%, neither -30%, and hate speech – 35%. We have sampled our experimental data into two sets; training data and testing data. The training dataset has contained 70% of total data and test dataset has contained rest of the 30%. Training documents contain 701-labeled examples and test documents contain 299 unlabeled examples.

*Pre-processing step*

Following text pre-processing tasks are done on labeled documents:

* Converting all the words into lower case
* Remove numbers, symbols, punctuations, retweets, URLs, strip white spaces from all documents.
* Use spellchecker to correct spellings
* Break each document into a set of single words (Tokenize)
* Remove all single words listed in stop word list (Stop word removal)
* Convert variants of words into their root (Stemming)
* Remove the contractions

*Feature extraction step*

* We have applied bag-of-words, N-gram (bi-, tri-), and TF-IDF feature extraction methods to create optimized feature vector and input data.
* N-grams refer to a sequence of N words or characters. They often play a key role in enabling machines to understand the context of the given text.
* TF-IDF which stands for Term Frequency – Inverse Document Frequency. It is one of the most important techniques used for information retrieval to represent.
* TF-IDF use two statistical methods, first is Term Frequency and the other is Inverse Document Frequency.
* Term frequency refers to the total number of times a given term ‘t’ appears in the document doc against (per) the total number of all words in the document.
* The inverse document frequency measure of how much information the word provides. It measures the weight of a given word in the entire document. IDF show how common or rare a given word is across all documents.

*Feature reduction step*

* Feature reduction, also known as dimensionality reduction, is the process of reducing the number of features in a resource heavy computation without

losing important information.

* Reducing the number of features means the number of variables is reduced making the computer’s work easier and faster.
* There are many techniques by which feature reduction is accomplished. Some of the most popular are generalized discriminant analysis, autoencoders, non-negative matrix factorization, and principal component analysis.
* We use Latent semantic analysis (LSA) as a feature reduction tool. It is helpful for text that is a series of these three steps a TF-IDF vectorization, a PCA (SVD in this case to account for the sparsity of text) and Row normalization.

*Machine learning algorithms selection step*

Choosing the best classifier is the most significant phase of the text classification pipeline. In order to select the best classifier, we tested several machine learning algorithms namely: Support Vector Machine (SVM), Random Forest, and Logistic Regression.

*Performance Evaluation:*

* It can be clearly seen the gradual increase of results after applying feature reduction using Latent Semantic Analysis (LSA).
* After doing all possible pre-process analysis, we have got a huge feature vector, containing 2084 features. After meticulous analysis with these reduced feature vectors we have identified that 320-feature vector work better that others to produce significantly less erroneous classification.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature  extraction | Feature selection | Document | Feature | Accuracy | Model |
| Bag-of-words | None | 701 | 2084 | 81% | Decision Tree |
| Unigram using  TF-IDF | None | 701 | 2084 | 81% | Decision Tree |
| Bigram using  TF-IDF | None | 701 | 6056 | 81% | Decision Tree |
| TF-IDF + LSA | None | 701 | 320 | 77% | Decision Tree |
| TF-IDF + LSA | None | 701 | 320 | 80% | Random Forest |
| TF-IDF + LSA | Cosine Similarity | 701 | 320 | 95% | Random Forest |
| TF-IDF + LSA | Cosine Similarity | 701 | 320 | 85% | SVM |
| TF-IDF + LSA | Cosine Similarity | 701 | 320 | 94% | Logistic Regression |

Table 3: Summary of training performance of different models with different feature extraction and selection methods

**2.3 SOCIAL MEDIA CYBERBULLYING DETECTION USING MACHINE LEARNING**

1. **INTRODUCTION**

Given the consequences of cyberbullying on victims, it is urgently needed to find a proper actions to detect and hence to prevent it. One of the successful approaches that learns from data and generates a model that automatically classifies proper actions is machine learning. Machine learning can be helpful to detect language patterns of the bullies and hence can generate a model to detect cyberbullying actions. This experiment have collected data from the Kaggle. In this study, ensemble machine learning method like Support vector machine and Neural Network are applied.

1. **MERITS AND DEMERITS**

*Merits:*

* It is a framework where it classifies the messages as bullying or not just by using 2 algorithms.
* With the use of Neural Network, there is an advancement in algorithm which improves the efficiency
* With the use of SVM, there is memory efficiency.
* It is very effective

*Demerits:*

* The Existing framework will not support big Sized conversations.
* The distribution of annotated class/data is unbalanced
* Neural network is quite uninterpretable
* This is quite expensive when compared to any other traditional algorithms.

1. **IMPLEMENTATION OF SOCIAL MEDIA CYBERBULLYING DETECTION USING MACHINE LEARNING**

All steps of the existing system framework are presented in Fig 3 and discussed in the following section.

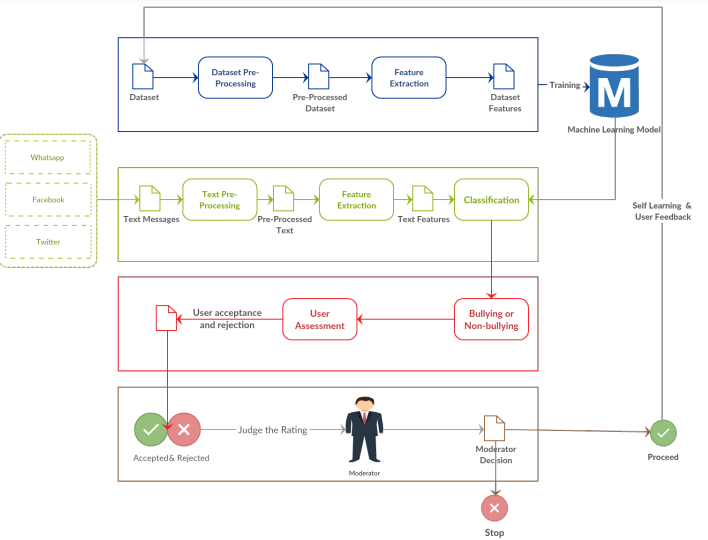


Fig 3. Social Media Cyberbullying Detection using Machine Learning Framework

***Methodology:***

*Data collection step*

Cyberbullying dataset from Kaggle which was collected and labeled by the authors Kelly Reynolds is used. This dataset contains in general 12773 conversations messages collected from Formspring. The dataset contains questions and their answers annotated with either cyberbullying or not. The annotation classes were unbalanced distributed such that 1038 question-answering instances out of 12773 belongs to the class cyberbullying, while 11735 belongs to the other class.

*Pre-processing step*

In the preprocessing step we clean the data by removing the noise and unnecessary text. The preprocessing step is done in the following:

1. Tokenization: In this part we take the text as sentences or whole paragraphs and then output the entered text as separated words in a list.
2. Lowering text: This takes the list of words that got out of the tokenization and then lower all the letters Like: 'THIS IS AWESOME' is going to be 'this is awesome'.
3. Stop words and encoding cleaning: This is an essential part of the preprocessing where we clean the text from those stop words and encoding characters like \n or \t which do not provide a meaningful information to the classifiers.
4. Word Correction: In this part we used Microsoft Bing word correction API that takes a word and then return a JSON object with the most similar words and the distance between these words and the original word.

*Feature extraction step*

In this step the textual data is transformed into a suitable format applicable to feed into machine learning algorithms. First we extract the features of the input data using TFIDF and put them in a features list. The key idea of TFIDF is that it works on the text and get the weights of the words with respect to the document or sentence. In Addition to TFIDF, we use sentiment analysis technique to extract the polarity of the sentences and add them as a feature into the features list containing the TFIDF features. The polarity of the sentences means that if the sentence is classified as positive or negative. For that purpose we extract the polarity using Text Blob library which is a pre-trained model on movie reviews. In addition to the feature extraction using TFIDF and sentiment polarity extraction, the propose approach uses N-Gram to consider the different combinations of the words during evaluation of the model. Particularly, we use 2- Gram, 3-Gram and 4-Gram.

*Machine learning algorithms selection step*

Choosing the best classifier is the most significant phase of the text classification pipeline. In order to select the best classifier, we tested several machine learning algorithms namely: Support Vector Machine (SVM) and

Neural network.

*Performance Evaluation*

The evaluation of classifiers is done using several evaluation matrices depends on the confusion matrix. Among of those criteria are Accuracy, precision, recall and f-score. They are calculated according to the following equations:

Accuracy =

Precision =

Recall =

F - Score =

Where TP represents the number of true positive, TN represents the number of true negatives, FP represents the number of false positives, and FN represents the number of false negatives classes.

**CHAPTER 3**

**RESULTS AND DISCUSSIONS**

1. **RESULTS AND DISCUSSION**
2. **COMPARISON OF EXISTING SOLUTIONS**

The first existing solution CYBERBULLYING SEVERITY DETECTION has the best overall classifier performance in multi-class setting achieved by Random Forest with SMOTE of having kappa statistic of 0.711, overall classifier accuracy 91.153, and

f-measure 0.898. The best overall classifier performance in binary setting was achieved by Random Forest for having f-measure 0.929. For DETECTION OF CYBER-AGGRESSIVE COMMENTS the analysis is done by three different models. Among the all, Random forest performed better in true classification of hate speech, offensive speech and neutral speech with 93% accuracy and the third existing solution SOCIAL MEDIA CYBERBULLYING DETECTION found Neural Network performing better with the highest accuracy of 91.76% and an f-score of 91.9%.

1. **DATA COLLECTION AND PERFORMANCE METRICS**

Bandeh Ali Talpur proposed a supervised machine learning technique for cyberbullying detection and multiclass categorization for Cyberbullying Severity Detection: A Machine Learning approach. They applied sentiment, embedding and lexicon features with PMI-semantic orientation and applied the selected features to Decision Tree, SVM, KNN, Naive Bayes, and Random Forest algorithms. The developed model classified the Twitter contents as cyberbullying or non-cyber bullying along with the severity level as low, medium, high, or none. Used an annotated dataset collected by <https://github.com/Mrezvan94/Harassment-Corpus/blob/master/Harassment%20Lexicon.csv> . The reasons for selecting this dataset include: (a) it is publicly available on git repository along with lexicon; (b) it is well-suited for our study as it contains the topics of cyberbullying that we are interested in. Lexicons related to the five topics (sexuality, racism, physical-appearance, intelligence and politics) were utilized to annotate tweets between December 18th, 2016 to January 10th, 2017. They found that Random Forest is the best classifier having f-measure of 0.929.

pone.0240924.t004

Table 4: Classifiers performance under various settings in multi-class classification

pone.0240924.t006

Table 5: Classifiers performance under various settings in binary classification

Risul and Nasrin proposed a framework for Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach to identify and filter cyberbullying comments in three different categories such as hate speech, offensive speech, and neither. Feature extraction methods such as bag-of-word, N- gram, and TF-IDF are used to create feature vectors and input data. They used logistic regression, random forest, and SVM models. During the training phase, they found that logistic regression and random forest have the same performance in classifying comments. After the training phase, the testing data is applied to the three models. To undertake the experiment we have collected social media text data of twitter comments from Data world website <https://data.world/crowdflower/hate-speech-identification> . The original data set contains two attributes and 3000 instances. Attributes are ‘label’ and ‘twitter text’. Finally, the experiment results have shown that the proposed model can classify cyber-aggressive comments more than

93% accurately. And, among all the models that are built, random forest outperforms logistic regression slightly, but it completely outperforms support vector machine very high in performance range. The significant findings from this study is, fine tuned combination of TF-IDF, LSA and cosine similarity analysis can produce optimized feature vector for quantitative analysis of text data for cyber-aggressive comments classification on social media networks.

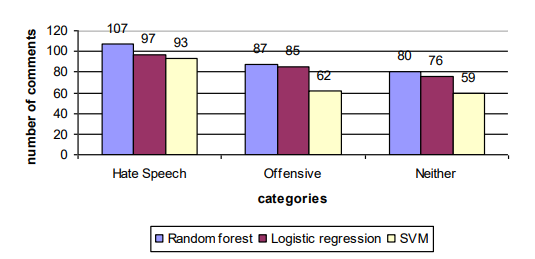


Fig 4: True classification summary for test data

John and Ammar proposed a framework for Social Media Cyberbullying Detection using Machine Learning by collecting dataset from Kaggle. We have used cyberbullying dataset which was collected and labeled by the authors Kelly Reynolds https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset?select=kaggle\_parsed\_dataset.csv . They applied feature extraction methods such as TF-IDF and sentimental analysis. Different N-gram language model-2- gram, 3-gram, and 4-gram were also applied. SVM and neural network classifiers are used for training data and their accuracy was evaluated. After evaluating the averages of accuracy, precision, recall, and f-score, the neural network has achieved an average accuracy of 91.76% and for SVM, the average accuracy was 89.87%. When they compared the proposed model with the other related work, they conclude that the proposed approach has the highest accuracy of 91.76% and an f-score of 91.9%. They found that neural network has performed better.

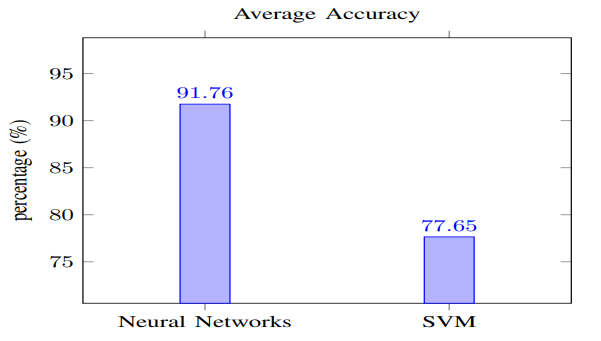


Fig 5: Comparison between the Best Classifiers in Terms of Accuracy

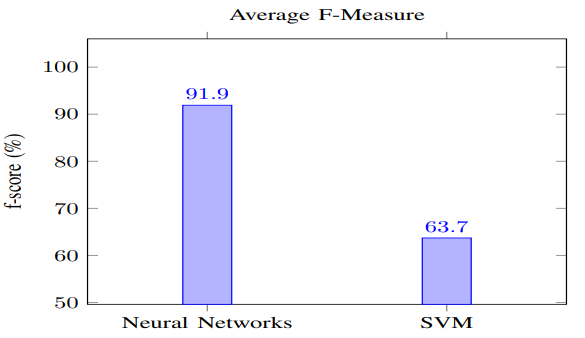


Fig 6: Comparison between the Best Classifiers in Terms of F-Measure

**CHAPTER 4**

**CONCLUSION**

1. **CONCLUSION**

The use of internet and social media has clear advantages for societies, but their frequent use may also have significant adverse consequences. This involves unwanted sexual exposure, cybercrime and cyberbullying. Cyberbullying across internet is dangerous and leads to mishappenings like suicides, depression etc and therefore there is a need to control its spread. In this report we have discussed existing solution for detection of cyberbullying. The purpose of this research is to design and develop an effective technique for the detection. The above experiments were conducted by collecting datasets from different websites. They were evaluated based on various different classifiers such as Support Vector Machine(SVM), Neural Network and used TFIDF, Tokenization etc., for feature extraction.

**CHAPTER 5**

**REFERENCES**

1. **REFERENCES**
2. Bandeh Ali Talpur, Declan OSullivan, Cyberbullying severity detection: A machine learning approach, Research Article: School of Computer Science and Statistics, Trinity College Dublin, Ireland, October 2020
3. Risul Islam Rasel, Nasrin Sultana, Sharna Akhter, Phayung Meesad, Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach, Conference: The 2nd International Conference, September 2018
4. John Hani Mounir, Muhamed Nashat, Mostafaa Ahmed, Zeyad Emad, Ammar Mohammed, Social Media Cyberbullying Detection using Machine Learning, Article: International Journal of Advanced Computer Science and Applications, January 2019
5. <https://www.educba.com/support-vector-machine-in-machine-learning/>
6. <https://www.tutorialspoint.com/what-is-a-neural-network-in-machine-learning>
7. <https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm>