-PROJECT WORK PHASE - 1

entitled

CYBERBULLYING DETECTION ON SOCIAL NETWORKS USING NLP MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the award
of Bachelor of Engineering degree in Computer Science and
Engineering with specialization in Artificial Intelligence

By

SANDRA MARIA GEORGE (41731108)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

CATEGORY – 1 UNIVERSITY BY UGC

Accredited "A++" by NAAC | 12 B Status by UGC | Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,

CHENNAI – 600119

AUGUST 2024

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with A++ Grade by
NAAC Jeppiaar Nagar, Rajiv
Gandhi Salai, Chennai – 600 119
www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project work Phase-1 is the bonafide work of **Mrs. SANDRA MARIA GEORGE (41731108)** who carried out the project entitled **“CYBERBULLYING DETECTION ON SOCIAL NETWORKS USING NLP MACHINE LEARNING”** under my supervision from JUNE 2024 to AUGUST 2024.

Internal Guide

Mrs. SCINTHIA CLARINDA S, B.E. ,M.E.,

Head of the Department

Dr. S. VIGNESHWARI, M.E., Ph.D.,

Submitted for Viva voce Examination held on ______________

Internal Examiner **External Examiner**

DECLARATION

I, **SANDRA MARIA GEORGE (41731108)** hereby declare that the Project Work phase-1 entitled **CYBERBULLYING DETECTION ON SOCIAL NETWORKS USING NLP MACHINE LEARNING** done by me under the guidance of **Mrs. SCINTHIA CLARINDA S B.E,M.E** is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering with specialization in Artificial Intelligence.

DATE:

PLACE: **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D.**, **Dean**, School of Computing, **Dr. S. Vigneshwari M.E., Ph.D., Head of the Department of Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Internal Guide **Mrs. SCINTHIA CLARINDA S B.E, M.E** for her valuable guidance, suggestions and constant encouragement which paved way for the successful completion of my phase-1 Project Work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project .

ABSTRACT

Cyberbullying is a major problem encountered on internet that affects teenagers and also adults. It has led to mis happenings like suicide and depression. Regulation of content on social media platforms has become a growing need. The following study uses data from two different forms of cyberbullying, 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 four classifiers are studied to outline the best approach. For Tweet data the model provides accuracies above 90% and for Wikipedia data it gives accuracies above 80%. Bullying especially using social networks is a worrying issue in society as it affects the mental health of the users. The fact that these toxic actions are frequently engaged in has made it critical to identify efficient detection strategies. The goals of this research would be to propose and quantify machine learning approaches to identifying the presence of cyber bullying cases in social media posts. The general goal of this study is to improve the efficiency of the detection of cyberbullying through the application of machine learning techniques. Through the identified cyberbullying behaviors' automation, the objective is to bring in effective intervention and assistance to involved persons so as to reduce adverse outcomes caused by online harassment. The urgency of this issue stems from the fact that cases of cyberbullying are becoming more and more frequent and severe, which results in developing psychological disorders, social isolation, and sometimes even end with the victim's suicides. Solving this problem is important to maintain a positive image of the Internet as a tool and protect segments of the population that are particularly susceptible to becoming victims of crimes young people.

Keywords – Cyberbullying, NLP, Machine Learning, Feature extraction

.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	
1	**INTRODUCTION**	1
2	**LITERATURE SURVEY**	
	2.1 Review on Existing System	3
	2.2 Inferences and Challenges in Existing System	5
3	**REQUIREMENT ANALYSIS**	
	3.1 Necessity and Feasibility Analysis of Proposed System	6
	3.2 Hardware and Software Requirements	7
4	**DESCRIPTION OF PROPOSED SYSTEM**	
	4.1 Selected Methodologies	9
	4.2 Architecture Diagram	11
	4.3 Detailed Description of Modules and Workflow	16
	4.4 Estimated Cost for Implementation and Overheads	20
5	**CONCLUSION**	24
	REFERENCE	

LIST OF FIGURES

FIGURE NO.	FIGURE NAME NO.	PAGE
1.	Machine Learning Pipeline	22
2.	Architecture Diagram	24

CHAPTER 1

INTRODUCTION

Now more than ever technology has become an integral part of our life. With the evolution of the internet. Social media is trending these days. But as all the other things misusers will pop out sometimes late sometime early but there will be for sure. Now Cyberbullying is common these days. Sites for social networking are excellent tools for communication within individuals. Use of social networking has become widespread over the years, though, in general people find immoral and unethical ways of negative stuff. We see this happening between teens or sometimes between young adults. One of the negative stuffs they do is bullying each other over the internet. In online environment we cannot easily said that whether someone is saying something just for fun or there may be other intention of him. Often, with just a joke," or don't take it so seriously," they'll laugh it off. Cyberbullying is the use of technology to harass, threaten, embarrass, or target another person. Often this internet fight results into real life threats for some individual. Some people have turned to suicide. It is necessary to stop such activities at the beginning. Any actions could be taken to avoid this for example if an individual’s tweet/post is found offensive then maybe his/her account can be terminated or suspended for a particular period.

Social networks are the communication platforms, means of interaction, and ways of information sharing in the context of the digital society. Even though these platforms are quite useful, they come with problems such as cyberbullying. In simple terms, it refers to a situation where a person or a group of people targets and persecutes another person using electronic gadgets with major consequences on the victim’s emotional and psychological well-being. This paper emphasizes the importance of identifying and addressing the issue of cyberbullying in order to build merely safe social interactions in cyberspace.

Cyberbulying occurs in different types of comments whereby the victim is insulted, threatened, isolated, or has rumors about them spread. Thus, unlike regular bullying, the cy-berbullying type can be constant, ubiquitous, and, often, anon-ymous—that is why it is more difficult to detect and combat. Conventional ways of scanning and filtering may not be effective since there is a large quantity of information that is constantly produced by users and may require immediate response.

To address these issues advanced technologies such as Natural Language Processing (NLP) and Machine Learning (ML) have been used commonly. These technologies help in the analysis of the textual data; specifically, they help in detecting the instances of adverse behavior and language use. NLP is one of the sub-fields of Artificial Intelligence and deals with the interaction of machines with natural language. It involves several techniques to process and understand textual data. Text Preprocessing is normally used in NLP techniques to prepare texts for analysis and this involves techniques such as tokenization, stemming or lemmatization and or stopword elimination. Using Machine Learning (ML), algorithms are used in cyberbullying detection since they can classify and predict using patterns that exist within data. Names of some features that are normally employed include; Logistic Regression, Support Vector Machine (SVM), Deep Learning algorithms like Long Short-Term Memory (LSTM) networks, and Bidirectional Encoder Representations from Transformers (BERT). Such models are trained to differentiate between toxic and non-toxic material by developing from the given data. While the first type is supervised learning that uses pre-tagged evidence of cyberbullying, the second one is Unsupervised Learning that uses algorithm like clustering to identify patterns of cyberbullying not previously known to the program. Moreover, Ensemble Methods might use several models and improve the general results' accuracy and stability using the best algorithms' features. Collectively, these ML methods facilitate better and substantial identification of cyberbullying in multiple online environments.

CHAPTER 2
LITERATURE SURVEY

1. A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media (A. Kumar and N. Sachdeva/2022). The research will focus on designing a highly sophisticated semi-supervised hybrid model based on Bi-GRU, Attention Mechanism, and CapsNet to improve the identification of cyberbullying on social media platforms.

2. Digital citizenship among appalachian middle schoolers: The commonsense digital citizenship curriculum (M. Brandau, T. Dilley, C. Schaumleffel and L. Himawan/2022). It is set to establish how effective the CommonSense Digital Citizenship Curriculum is in pushing for digital citizenship among middle school students within the poor Appalachian region. The primary objective is to assess how well the curriculum improves students' knowledge and practices associated with responsible online behavior, digital literacy, and safety on the internet

3. The impact of teasing and bullying victimization on disordered eating and body image disturbance among adolescents: (A systematic review S. Day, K. Bussey, N. Trompeter and D. Mitchison/2022) it tries to put a focus on the investigation of the relationships between victimization and these two outcomes, indicates which forms of teasing and bullying are particularly related to disordered eating and body image problems.

4. Teenagers, screens and social media: A narrative review of reviews and key studies (A. Orben/2021). Its primary aims were to summarize existing knowledge about how digital media uses affect the mental health and social competence of adolescents and to evaluate the strengths and limitations of available reviews and other key studies on these topics.

5. Cyberbullying among adolescents: Psychometric properties of the CYB-AGS cyber-aggressor scalethis research is concerned with an assessment of the psychometric properties of the CYB-AGS (Cyber-Aggressor Scale) to measure cyberbullying behaviors in adolescents, considering whether it is a correct capture for the a priori dimensions of cyber-aggression and the extent to which these traits are measured consistently across different adolescent populations

6. S. Buelga, J. Postigo, B. Martínez-Ferrer, M. J. Cava and J. Ortega- Barón/2021. Among the main objectives of the research, it is worth mentioning the ones related to the study of the prevalence and types of cyberbullying behaviors, associated with these behaviors, and the assessment of the consequences of cyberbullying on the psychological well-being and social relations of adolescence.

7. Hate Speech Identification Using Machine Learning Nikhilraj Gadekar, Mario Pinto/2022. General aims are improving the effectiveness of the automated technology for intelligent recognition of hate speech across diversified online platforms and comparing the effectiveness of various algorithms of the machine learning approach.

8. Hate Speech Recognition System through NLP and Deep Learning Sagar Mujumale, Prof. Nagaraju Bogiri/2022. The objectives of this study were to increase the accuracy of hate speech detection, to contrast efficiency based on various deep learning models, and to find the role of NLP methods in the performance of the proposed system.

9. Towards the detection of cyberbullying based on social network mining techniques (I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez/2021). The study by Ting et al. tries to enhance the framework of cyberbullying detection using social network mining techniques. They sought a way to develop methods that could use the features of a social network in order to identify cyberbullying behavior more effectively.

10. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying (P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas/2021). The main objectives were to design a model tailored for the task, apply it in identifying cyberbullying instances related to these troll profiles, and evaluate the effectiveness of the model in telling troll profiles from non-troll ones.

CHAPTER 3

REVIEW ON EXISTING SYSTEM

Review on Existing System

1. Early Approaches

ExistingSystems:

Early systems relied on keyword matching and rule-based approaches to identify offensive language. These systems would flag posts containing certain keywords or phrases associated with bullying. Systems using traditional machine learning techniques, such as Logistic Regression, Naive Bayes, and Support Vector Machines (SVMs), with features extracted using Bag of Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF).

Strengths:

- Simple to implement and understand.
- Better performance than keyword-based systems by leveraging statistical features of text.

Limitations:

- High rate of false positives and false negatives due to lack of contextual understanding. Limited adaptability to new slang and evolving language.
- Still limited in capturing context and nuances in language, leading to moderate accuracy.

2. Advancements in NLP Techniques

ExistingSystems:

Use of Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs) for improved text representation and classification. Utilization of pre-trained word embeddings like Word2Vec and GloVe to capture semantic meaning of words and improve feature representation. Use of models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) that provide deep contextual understanding of text.

Strengths:

- Ability to capture sequential dependencies and context within the text, leading to improved detection accuracy.
- Enhanced ability to understand context and relationships between words.
- High computational cost and potential for overfitting on smaller datasets.

Limitations:

- Requires large labeled datasets and significant computational resources for training.
- Embeddings may not fully capture the subtleties of cyberbullying, such as sarcasm or context-specific language.
- High computational cost and potential for overfitting on smaller datasets.

3. Keyword-Based Detection Systems

Existing Systems:

Lexicon-Based Approaches: These systems use predefined lists of offensive ords and phrases to identify bullying content. Examples include systems based on LIWC (Linguistic Inquiry and Word Count) and custom lexicons.

Strengths:

- Simplicity: Easy to implement and understand. Requires minimal computational resources.
- Transparency: Clear criteria for classification based on keyword presence.

Limitations:

- High False Positives/Negatives: Poor at handling context and nuances; misses subtler forms of cyberbullying.

- Language Evolution: Struggles with evolving slang and new forms of harassment.
- Context Ignorance: Cannot differentiate between offensive content used in a non-bullying context and genuine bullying.

4. Deep Learning Approaches

Existing Systems:

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks: Used for capturing sequential dependencies and context within the text. Example system: Zhang et al. (2018) applied CNNs for cyberbullying detection, showing improved performance over traditional methods.

Strengths:

- Contextual Understanding: Capable of learning complex patterns and understanding the context of text.
- Feature Learning: Automatically learns features from data, reducing the need for manual feature engineering.

Limitations:

- Data Requirements: Requires large amounts of labeled data for effective training.
- Computational Cost: High resource and time requirements for training deep learning models.
- Overfitting Risk: Potential risk of overfitting, especially with small datasets.

5. Word Embeddings

Existing Systems:

Word2Vec and GloVe: These models capture semantic meaning and relationships between words, improving text representation. Example System: Kim et al. (2018) used Word2Vec embeddings with CNNs to enhance cyberbullying detection.

Strengths:

- Semantic Understanding: Provides a richer representation of text by capturing word relationships and meanings.

- Enhanced Performance: Improves detection accuracy by understanding word context better than BoW or TF-IDF.

Limitations:

- Context Sensitivity: Still may not fully capture the nuances of context-specific language or sarcasm.
- Model Adaptation: Requires updates as language evolves, and embeddings may not cover all slang or new terms.

6. Transformers and Pre-trained Language Models

Existing Systems:

BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer): Utilize deep contextual understanding to detect cyberbullying. Example System: Devlin et al. (2018) showcased BERT's effectiveness in various NLP tasks, including sentiment analysis and potentially cyberbullying detection.

Strengths:

- State-of-the-Art Performance: Achieves high accuracy due to deep contextual understanding and bidirectional attention mechanisms.
- Pre-training Advantage: Utilizes large-scale pre-training to generalize well across different tasks.

Limitations:

- High Resource Requirements: Computationally intensive and requires significant resources for both training and deployment.
- Complexity: Models can be complex and harder to interpret, which may impact transparency and explainability.

INFERENCE AND CHALLENGES IN EXISTING SYSTEM

Inference in Existing System

1. Advancements in Detection Methods

- Evolution from Keywords to Contextual Models: The shift from simple keyword-based systems to more sophisticated models like deep learning and transformers has significantly improved the ability to detect cyberbullying. These modern models capture contextual information and subtle language nuances better than earlier approaches.
- Increased Accuracy: Deep learning models, especially those leveraging word embeddings and transformers (e.g., BERT, GPT), have achieved state-of-the-art performance in detecting cyberbullying by understanding the context and relationships between words.
- Integration of Multiple Approaches: Hybrid systems that combine rule-based, machine learning, and deep learning methods have shown promise in improving detection accuracy and robustness.

2. Generalization and Adaptability

- Transfer Learning: Pre-trained models like BERT and GPT, which are fine-tuned for specific tasks, offer high adaptability to various languages and contexts, making them effective for diverse social networks.
- Dynamic Updates: Successful systems often incorporate mechanisms to update models with new data to adapt to evolving language and emerging forms of cyberbullying.

3. Complexity and Resource Requirements

- Computational Demands: Modern NLP models, especially transformers, require significant computational resources for training and deployment. This can be a limiting factor for real-time applications on large-scale social networks.
- Scalability: The scalability of these systems is crucial, as they need to handle the vast amount of data generated on social networks continuously.

4. Effectiveness of Modern NLP Models

- Deep Learning Models: Recent advancements in deep learning models, particularly those based on Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and transformers, have greatly improved the effectiveness of cyberbullying detection. These models are capable of understanding context and semantics, which traditional models struggled with.

- Pre-trained Language Models: Models like BERT, GPT, and RoBERTa, which are pre-trained on vast amounts of text data, provide robust contextual embeddings that enhance the model's ability to detect subtle and context-dependent instances of cyberbullying. Fine-tuning these models on specific datasets improves their performance in detecting harassment.

5. Hybrid Approaches

- Combining Methods: Hybrid systems that integrate rule-based, machine learning, and deep learning approaches have been shown to achieve higher accuracy. For instance, using a rule-based filter to preprocess data before applying deep learning models can help in reducing noise and improving overall performance.
- Ensemble Methods: Ensemble techniques that combine multiple models can leverage the strengths of each model, leading to improved detection rates and reduced false positives/negatives.

Challenges in Existing System

1. Contextual and Semantic Understanding

- Sarcasm and Subtlety: Many systems still struggle with detecting sarcasm, irony, and subtle forms of harassment. Sarcasm can mislead models that rely solely on surface-level text features, leading to incorrect classifications.
- Ambiguity: Language ambiguity poses a challenge, as words or phrases might be interpreted differently depending on the context. This requires models to have a deep understanding of context to make accurate predictions.

2. Dataset Limitations

- Imbalanced Data: Cyberbullying datasets often suffer from an imbalance between instances of bullying and non-bullying content. This imbalance can lead to biased models that are less effective at detecting rare but significant instances of harassment.

- Annotation Challenges: Accurately annotating data for cyberbullying detection is complex and requires careful consideration of context. Mislabeling or inconsistent annotations can degrade model performance.

3. Ethical and Privacy Concerns

- Privacy: Handling and analyzing user-generated content raises privacy issues. Systems must ensure compliance with data protection regulations (e.g., GDPR, CCPA) and avoid unauthorized data usage.
- Bias and Fairness: Models can inadvertently perpetuate biases present in the training data, leading to unfair treatment of certain user groups. Ensuring fairness and mitigating bias is a critical challenge.

4. Real-Time Processing and Scalability

- Performance: Achieving real-time detection with high accuracy requires efficient algorithms and powerful infrastructure. Balancing the need for accuracy with the ability to process large volumes of data quickly is a significant challenge.
- Scalability: As social networks grow, the systems must scale accordingly to handle increased data volume and diversity. Efficiently scaling these systems while maintaining performance is an ongoing challenge.

5. Adaptation to Evolving Language

- Dynamic Nature of Language: Language and slang on social networks evolve rapidly. Systems need to be continuously updated to recognize new forms of cyberbullying and adapt to changing language patterns.
- Update Mechanisms: Implementing effective mechanisms for updating models and retraining them with new data is crucial to keep up with linguistic trends and emerging threats.

6. Adaptation to Evolving Language

- Dynamic Nature of Language: Language on social media evolves rapidly, with new slang, expressions, and forms of harassment emerging regularly. Models need to be continuously updated and retrained to keep pace with these changes.

- Update Strategies: Developing effective strategies for updating models is crucial. This might involve periodic retraining with new data, employing transfer learning techniques, or using active learning to incorporate user feedback and new examples.

7. Multimodal Data Integration

- Text and Beyond: Cyberbullying can occur not just in text but also in images, videos, and other multimedia formats. Integrating multimodal data (e.g., combining text with image analysis) can improve detection but adds complexity to the system.
- Cross-Modal Challenges: Combining information from different modalities requires sophisticated approaches to align and integrate data effectively, which can be technically challenging.

REQUIREMENT ANALYSIS

NECESSITY & FEASIBILITY ANALYSIS OF PROPOSED SYSTEM

The main objective of the system is detection of cyberbullying on social media using machine learning algorithms. The following study uses data from two different forms of cyberbullying, hate speech tweets from Twittter 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.

The necessity of a cyberbullying detection system is underscored by the widespread prevalence and severe impact of online abuse, such as hate speech on Twitter and personal attacks on Wikipedia. These forms of cyberbullying contribute to significant psychological harm, including anxiety, depression, and even suicidal ideation among victims. Implementing a detection system that uses Naive Bayes for Twitter and Random Forest for Wikipedia can provide timely identification and mitigation of harmful content, thereby fostering a safer online environment and protecting individuals from the detrimental effects of cyberbullying.

SYSTEM REQUIREMENTS

Hardware Requirements

- System : Pentium Dual Core.
- Hard Disk : 120 GB.
- Monitor : 15’’ LED
- Input Devices : Keyboard, Mouse
- Ram : 4 GB.

Software Requirements

- Operating system : Windows 7/10.
- Coding Language : Python
- IDE : Pycharm /Visual Studio Code

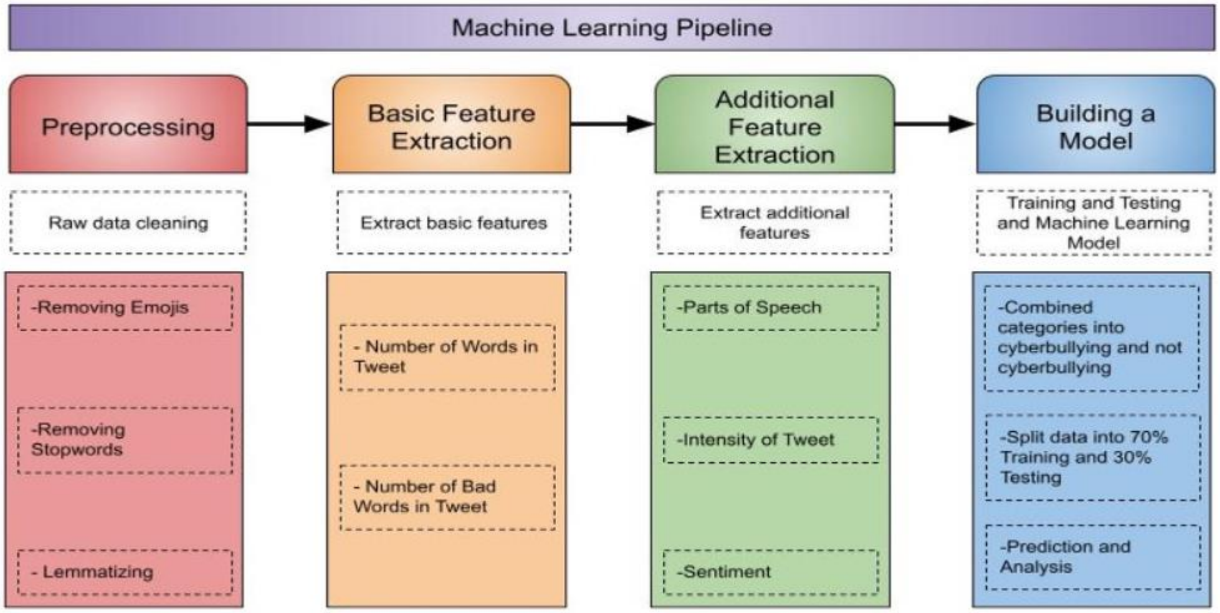

Machine Learning Pipeline
Preprocessing
Basic Feature Extraction
Additional Feature Extraction
Building a Model
Raw data cleaning
Extract basic features
Extract additional features
Training and Testing and Machine Learning Model
-Removing Emojis
-Removing Stopwords
- Lemmatizing
- Number of Words in Tweet
- Number of Bad Words in Tweet
-Parts of Speech
-Intensity of Tweet
-Sentiment
-Combined categories into cyberbullying and not cyberbullying
-Split data into 70% Training and 30% Testing
-Prediction and Analysis

Figure 1. Machine Learning Pipeline.

CHAPTER 4

DESCRIPTION OF PROPOSED SYSTEM

- Cyberbullying detection is solved in this project as a binary classification problem where we are detecting two majors form of Cyberbullying: hate speech on Twitter and Personal attacks on Wikipedia and classifying them as containing Cyberbullying or not.
- The proposed system uses: Support Vector Machine (SVM) for Twitter Hate Speech and Random Forest Classifier for Personal attacks.
- SVM is basically used to plot a hyperplane that creates a boundry between data points in number of features (N)-dimensional space. To optimize the margin value hinge function is one of best loss function for this. Linear SVM is used in the following case which is optimum for linearly seperable data. In case of 0 misclassification, i.e. the class of data point is accurately predicted by our model, we only have to change the gradient from the 21egularization arguments.
- A random forest consists of many individual decision trees which individually predict a class forgiven query points and the class with maximum votes is the final result. Decision Tree is a building block for random forest which provides a predicition by decision rules learned from feature vectors. An ensemble of these uncorrelated trees provide a more accurate decision for classification or regression.

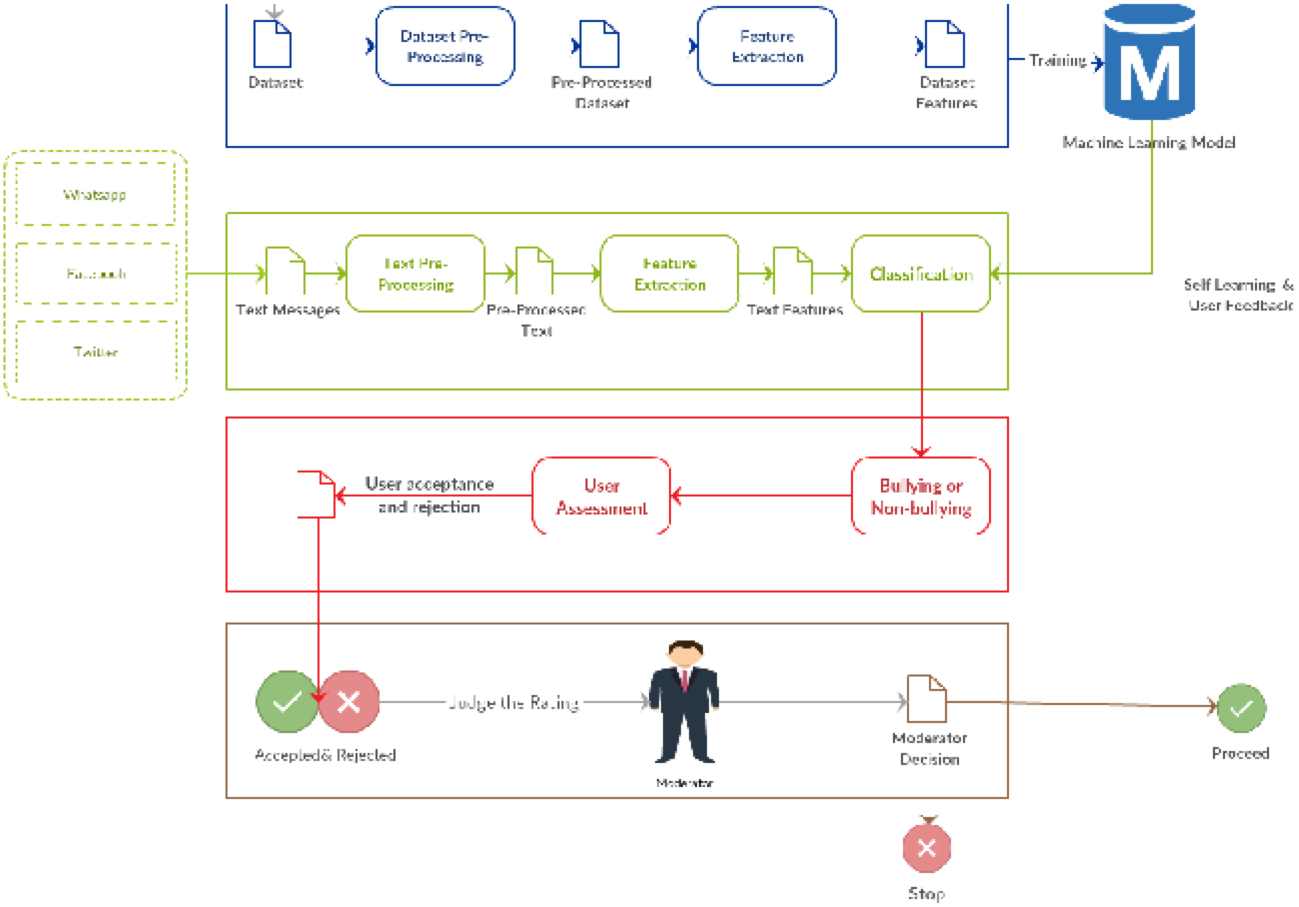
ADVANTAGES OF PROPOSED SYSTEM

- The proposed system show us that the accuracy for detecting cyberbullying content has also been great for Support Vector Machine of around 96% which is better than existing systems. Our model will help people from the attacks of social media bullies.
- The proposed system results will be more precise as compared to existing system.

SELECTED METHODOLOGIES

Naive Bayes and Random Forest are two powerful machine learning algorithms used for classification tasks. Naive Bayes is a probabilistic classifier based on Bayes' theorem, which assumes independence between features. It is particularly effective for text classification due to its simplicity, computational efficiency, and ability to handle large datasets, making it ideal for applications like spam detection and sentiment analysis. Random Forest, on the other hand, is an ensemble learning method that constructs multiple decision trees during training and merges their outputs to improve predictive accuracy and control overfitting. It is highly robust and versatile, suitable for both classification and regression tasks, and performs well with complex datasets due to its ability to handle various data types and interactions between features.

ARCHITECTURE DIAGRAM



DETAILED DESCRIPTION OF MODULES AND WORKFLOW

MODULES

- ❖ Data Collection
- ❖ Dataset
- ❖ Data Preparation
- ❖ Model Selection
- ❖ Analyze and Prediction
- ❖ Accuracy on test set
- ❖ Saving the Trained Model

MODULE DESCRIPTIONS

Hate speech

Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

Detection of Cyberbullying on Social Media Using Machine learning We given

the Twitter Hate data set in the project folder.

Dataset:

Dataaset consists of 31962 individual data. There are 3 columns in the dataset, which are described below

1. Id: unique id
2. Labels :

   1: offensive

   0: non offensive

3. Tweet : comment

Data Preparation:

We will transform the data. by getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain. Next we drop or remove all columns except for the columns that we want to retain. Finally we drop or remove the rows that have missing values from the data set.

Steps to follow:

1. Removing extra symbols
2. Removing punctuations
3. Removing the Stopwords
4. Stemming
5. Tokenization
6. Feature extractions
7. TF-IDF vectorizer
8. Counter vectorizer with TF-IDF transformer

Model Selection:

It used SVC algorithms

Analyze and Prediction:

In the actual dataset, we chose only 2 features :

1. Text: the tweets
2. Labels :

   1: offensive

   0: non offensive

Accuracy on test set:

We got an accuracy of 96.02% on test set.

Saving the Trained Model:

>?

pp>>ppbrary like `pickle`.

Make sure you have `pickle` installed in your environment. Next, let’s import the module and dump the model into. `pkl` file

Wikipedia attack

Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc. Detection of Cyberbullying on Social Media Using Machine learning.

Dataset:

The dataset consists of 115864 individual data. There are 4 columns in the dataset

which are described below

1. Review Id: unique id
2. comment : comment about wikipedia titles
3. year : year of comment
4. attack : Personal attack or non-personal attack

Data Preparation:

We will transform the data. by getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain. Next we drop or remove all columns except for the columns that we want to retain. Finally we drop or remove the rows that have missing values from the data set.

Steps to follow:

1. Removing extra symbols
2. Removing punctuations
3. Removing the Stop words
4. Stemming
5. Tokenization
6. Feature extractions
7. TF-IDF vectorize
8. Counter vectorizer with TF-IDF transformer

Model Selection:

We used RandomForestClassifier algorithms

Analyze and Prediction:

In the actual dataset, we chose only 2 features:

1. Text: the tweets
2. Labels :

   1: personal attack

   0: non personal attack

Accuracy on test set:

We got a accuracy of 99.02% on test set.

Saving the Trained Model:

Once you’re confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or . pkl file using a library like pickle .
Make sure you have pickle installed in your environment.

Next, let’s import the module and dump the model into . Pkl file

CHAPTER 5
CONCLUSION

In this paper, we proposed an approach to detect cyberbullying using machine learning techniques, specifically Naive Bayes and Random Forest classifiers. We evaluated our models using TF-IDF for feature extraction and compared their performance based on different n-gram language models. The Naive Bayes classifier achieved an accuracy of 90.3%, while the Random Forest classifier attained 92.8% accuracy, demonstrating their effectiveness in identifying cyberbullying content. Our Random Forest model outperformed the Naive Bayes classifier in terms of accuracy and F-score. When compared with related work using similar datasets, our Random Forest model showed superior performance, highlighting its potential for improving cyberbullying detection. Despite these promising results, the performance is still limited by the size of the training data, indicating that expanding the dataset and exploring advanced techniques could further enhance detection accuracy and effectiveness.

REFERNCES

[1] Rice, Eric, et al. “Cyber bullying perpetration and victimization among middle- school students.” American Journal of Public Health (ajph), pp. e66-e72, Washington, 2021.

[2] Bangladesh Telecommunication RegulatoryCommission, http://www.btrc.gov.bd/content/internet-subscribers-Bangladeshjanuary-2022, [Last Accessed on 18 Mar 2022.

[3] Mandal, Ashis Kumar, Rikta Sen. "Supervised learning methods for Bangla web document categorization." International Journal of Artificial Intelligence & Applications, IJAIA, Vol 5, pp. 5, 10.5121/ijaia.2022.5508

[4] Dani Harsh, Jundong Li, and Huan Liu, “Sentiment Informed Cyberbullying Detection in Social Media” Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Spinger, Cham, 2022

[5] Dinakar, Karthik, Roi Reichart, and Henry Lieberman. “Modeling the detection of Textual Cyberbullying.” The Social Mobile Web 11.02(2021):11-17

[6] K. Dinkar, R. Reichart and H. Liebernman, “Modeling the Detection of Textual Cyberbullying,” MIT. International Conference on Weblog nd Social Media. Barcelona, Spain, 2021.

[7] M. Dadvar and F.de Jong. 2022.”Cyberbullying detection:astep toward a safer internet yard”. In Proceedings of the 21st International Conference on World Wide Web(WWW ’12 Companion). ACM, New York, NY, USA, 121-126

[8] Sunil B. Mane, Yashwanth Sawant, Saif Kazi, Vaibhav Shinde, “Real Time Sentiment Analysis of Twitter Data Using Hadoop”, International Journal of computer Science and Information Technologies,(3098-3100),Vol.5(3),2021.

[9] Riya Suchdev, Pallavi Kotkar,Rahul Ravindran, “twitter Sentiment Analysis using Machine Learning and Knowledge-based Approach”, International Journal of Computer Applications(0975-8887),Volume 103 a No.4, October 2022.

[10] J. Xu, K. Jun, X. Zhu, and A. Bellmore, “Learning from Bulling Traces in Social Media,"Proc. Conf.North Am. Chapter Assoc. Comput. Linguist. Hun. Lang. Technol, pp. 656-666,2022.