TABLE OF CONTENTS

Page

ABSTRACT

Our system detects cyberbullying on social media by analyzing user's posts and comments. It deals with text posts, memes and videos. The various state-of-art methods used to detect cyberbullying include CNN, LSTM, BLSTM, Bi-GRU, XGBoost and SVM. Our novel system aims to investigate the detection of cyberbullying using hate content identification followed by the level of toxicity of the text.

The research method used by our project is BERT, the novel algorithm developed recently by Google. Our system includes two BERT models.The first model classifies the content to be aggression,attack,non-sarcasm,racism,sarcasm,sexism and toxicity.If the content falls under normal or sarcasm ,no action will be taken otherwise the content will be passed to the second BERT model for toxicity level classification.This is because, for content to be toxic, it should firstly exhibit hate.

The expected result will fall into one of the following toxicity levels: Toxic, Obscene, Insult, Threat, Identify hate and Severely toxic. If the content is toxic,obscene or insulting, the system simply mails and warns the user. Remaining results will be regarded as a considerable form of cyberbullying and thus, users will be mailed a warning and their account is suspended.

LIST OF ABBREVIATIONS

Acronym	Abbreviation
BERT	Bidirectional Encoder Representations from Transformers
WEKA	Waikato Environment for Knowledge Analysi
TF-IDF	Term frequency–inverse document frequency
SVM	Support Vector Machine
DNN	Deep Neural Network
CNN	Convolutional Neural Network
LSTM	Long Short Term Memory
BLSTM	Bidirectional Long Short Term Memory
LIWC	Linguistic Inquiry and Word Count
NN	Neural Network
NB	Naïve Bayes
SA	Sentiment Analysis
API	Application Programming Interface
Bi-GRU	Bi-gated recurrent neural networks
Bi-LSTM	Bidirectional Long Short-Term Memory
SMO	Social Media Optimisation
IBK	Instance Based Learner
MLP	Multi Layer Perceptron
RNN	Recurrent Neural Network
NLP	Natural language Processing

CHAPTER 1 INTRODUCTION

Cyberbullying is an everyday growing problem in social media, and this makes the task of automatic Cyberbullying detection and protecting its victims quite crucial and of immense importance. Cyberbullying or hate speech is the act of posting hateful content on social media which is found abusive due to its repetitive and abusive nature.

Cyberbullying is a type of threat that occurs using digital gadgets such as cellphones, hardware systems, and remote tablets. Cyberbullying can occur offline on social media platforms, or while playing games where people could see, take part in, or share data with one in social media platforms, or. Sending, receiving, posting, or sharing negative, dangerous, false or mean content about someone is considered cyberbullying. It may include humiliating or harassing someone by disclosing personal or private data about them. Some cases of cyberbullying may be considered illegal or criminal.

Cyberbullying, a type of cyber harassment, can take many forms, typically, however, refers to repetitive hostile behavior using digital media (e.g., hurtful comments, videos and images) to intentionally and repeatedly harass or harm individuals. Cyberbullying is permanent (i.e., the content remains accessible online unless removed) and potentially widespread(i.e., online social media provide a wide audience and quick spread of online posts).

1.1. SCOPE

Bullying is a widespread phenomenon, considered to be a serious problem. An increase in the use of electronic media has provided a new platform for bullying. According to a recent study by the York Region Parent Health Connection in Ontario, 60% of all students use chat rooms and instant messaging, 25% report receiving bullying messages, 16% admit to posting threatening messages, 14% have been threatened on the Internet, and 44% possess an email account without their parents' consent. Targets of cyberbullying are often reluctant to report the abuse in case their parents restrict or severely supervise their computer time.

Cyberbullying prevention is the need of the hour. It needs to be monitored and put to an end. Therefore, we introduce our proposed optimization framework that automatically computes the probability of a comment to be indicative of aggression with high accuracy. All comments in a media session are ordered in time. In our approach, we encode the belief of a media session to be indicative of cyberbullying as the number of identified aggressive comments. We measure timeliness as the number of comments "saved" in comparison to a baseline. By taking timely and necessary actions we can curb cyberbullying and help in decreasing the number of cyberbullying cases.

CHAPTER 2 PROBLEM DEFINITION

The objective of our project is to develop a robust prototype that can be integrated with any social media platform to curb cyberbullying. Our system detects cyberbullying on users' posts and comments and further, their level of toxicity. Besides, in case of highly toxic content shared, it also warns the user and/or temporarily suspends their account for the next two hours as a penalty. The system also extends to detecting cyberbullying on three types of content- images (keeping memes in mind),videos as well as text content.

When a user posts an image, its OCR will be done to extract the text. When a user shares or posts text data, all the text in it including the emoticons will be extracted directly.The system uses a modern approach to sentiment analysis. Cyberbullying includes negative or toxic posts or comments. Hence, for content to be toxic, it should firstly exhibit hate content. Thus, our system includes two BERT models to detect cyberbullying.

When a user posts a video,subtitles are extracted from the video keeping youtube videos in mind.The link of the entire playlist itself can be inputted to the proposed

system which will get the subtitles of all the videos of the playlist and further analyze this text as explained previously,by passing to both the BERT models accordingly.

The entire text extracted from the user's post or comment is passed to the first BERT model for the identification of hate in which the content is classified to be one of these eight categories: aggression,attack,non-sarcasm,normal,racism,sarcasm,sexism and toxicity.If the content falls under normal or sarcasm ,no action will be taken otherwise the content will be passed to the second BERT model for toxicity level classification to identify one of the six levels of toxicity: toxic, obscene, insult, threat, identify hate and severely toxic. If the hate content falls into any one of the first three categories (toxic, obscene, insult), the system simply mails and warns the user. But, if it belongs to any of the next three categories (threat, identify hate, severe toxic), it will be regarded as a considerable form of cyberbullying and thus the user will be mailed a warning and their account will also be suspended for the next two hours as a sign of penalty.

CHAPTER 3 LITERATURE SURVEY

Sl. No	Author	Title of the paper	Description	Limitations
1	**Fankar Armash Aslam,** **Hawa Nabeel Mohamme d ,** **P. S. Lokhande**	Efficient Way Of Web Development Using Python And Flask et.al(2015)	Web development is a complex process of structuring content with dynamic data transactions. For maintaining such complexity technologies such as python Jinja Flask are more useful. Python can be used for making web more powerful, fast and efficient with the help of Flask Template Engine.	(a). Python isn't the best for memory intensive tasks. (b). Python not a great choice for a high graphic 3d game that takes up a lot of CPU.

2	**Patrick Vogel** Johann Bernoulli Institute for Mathema tic s and Computer Science University of Groninge n, Netherla nd	A Low-Effort Analytics Platform for Visualizing Evolving Flask Based Python Web Services et.al(2017)	Despite a rich ecosystem of extensions, there is none that supports the developer in gaining insight into the evolving performance of their service. In this paper, they introduce Flask Dashboard, a library that addresses this problem. It presents the ease with which the library can be integrated in an already existing web application, discusses some of the visualization perspectives that the library provides and points to some future challenges for similar libraries.	The disadvantage of Automatically monitoring system evolution is that it will consider on equal ground the smallest of commits, even one that modifies a comment, and the shortest lived of commits, e.g. a commit which was active only for a half an hour before a new version with a bug fix was deployed, with major and minor releases of the software.
3	**Hariani**	Detection Of Cyberbull	Detection of cyberbullying on twitter can be done 1) Analysis preprocessing and data cleansing	Search API is not complete index all tweets but 1500 index tweets. Search API cannot be used to find the tweet was more than a week. Query which is complex may not be successful Search does not support authentication

		ying On social media Using Data Mining Techniqu es et.al(2017)	2) Classification using WEKA, classification is performed on the data that has been clean, TF-IDF weighting and validation data using 10-fold cross validation and then do classification. We can use others different algorithms of data mining such us SVM, K-Means, C45 and others.	Which means all Queries were published anonymously.
4	**Sweta Agrawal**	Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms et.al(2018)	This paper experimented with four DNN based models for cyberbullying detection: CNN, LSTM, BLSTM, and BLSTM with attention. These models are listed in the increasing the complexity of their neural architecture and amount of information used by these models. They proved that ML models Performance is	These models can be further improved with extra data such as information about the profile and social graph of users. Most of the current datasets do not provide any information about the severity of bullying. If such fine-grained information is made available, then cyberbullying detection models

			lower as compared to DNN. All DNN models reported here were Implemented using Keras. We pre-process the data, subjecting it to standard operations of removal of stop words, punctuation marks and lowercasing, before annotating it to assigning respective labels to each comment. For each trained model, we report its performance after doing five-fold cross-validation. They used transfer learning to check if the knowledge gained by DNN models on one dataset can be used to improve cyberbullying detection performance on other datasets.	can be further improved to take a variety of actions depending on the perceived seriousness of the posts.

5	**Mengfan Yao**	Cyberbullying Ends Here:Towards Robust Detection of Cyberbullying in social media et.al (2019)	Reduced time to raise a cyberbullying alert by (i)sequentially examining comments as they become available over time, and Minimizing the number of feature evaluations necessary for a decision to be made for each comment.	Experimental evaluation is limited to a single dataset and therefore the performance of this approach should not be generalized to other platforms. Incompetent to validate the labels in the dataset,

			(ii)Differentiates between cyberbullying (i.e., an act of aggression that is repeated overtime) and cyber aggression (i.e., harassment manifested as one-off profanity).	prevents the user from obtaining a more granular understanding of the effect of false positives on the accuracy of our approach. Comment–level labels, even though imperative for capturing the repetitive nature of cyberbullying as a process over a period of time, can be costly or time consuming to obtain.
6	**Mengfan Yao** University at Albany	Robust Detection of Cyberbullying in social media et.al (2019)	Achieves timely, scalable, and accurate detection of cyberbullying	Focuses only on one online platform i.e. Instagram

7	**Nijia Lu, Guohua W u, Zhen Zhang, Yitao Zheng, Yizhi Ren**	Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. et.al(2019)	Experiments are performed on both the Chinese Weibo dataset and the English Tweet dataset. Characters are used as the smallest unit since working on only characters has the advantage of being able to naturally learn unusual character combinations such as emoticons. Skillfully utilizes shortcuts to stitch different levels of features to learn more granular bullying signals, and we adopt focal loss function to overcome the	High-frequency words that appear in the racism and sexism bullying categories and are likely to be mistaken for bullying in models that do not emphasize semantics. For the corpus of social media, the shallow word CNN's uses a huge embedded table that contains many noise words.

			class imbalance problem in the dataset.	

8	**David Van Bruwaene, Qianjia Huang , Diana Inkpen**	A multi-platform dataset for detecting cyberbullying in social media et.al(2020)	Diversity of examples presentin the dataset (selected based on likelihood) and its good performance enhances the diversity of communication media to which the models can be applied. The comparison between different models demonstrates that the CNN and XGBoost models perform better in general than the SVM models and that adding LIWC features support the models in a positive way	Balancing the training data set negatively impacted performance for both cyber aggression and bullying datasets, with and without added LIWC features
9	**Nabi Rezvani, Amin Beheshti, Alireza Tabebordb a r**	Linking Textual and Contextual Features for Intelligent Cyberbullying Detection in Social Media et.al(2020)	Proposed method has been able to significantly improve on most metrics by adding contextual features	The Twitter dataset was more sparse and while the NN combiner has been able to alleviate the sparseness, the boosting combiner has not been able to perform as good, given that it uses a simple voting scheme

10	**Kirti Kumari, Jyoti Prakash Singh**	Identification of cyberbullying on multi-modal social media posts using genetic algorithm et.al(2020)	The features are optimized using genetic algorithms to increase the efficiency of the whole system. The proposed model is validated with a dataset containing text and image to achieve an F1- score 78% which shows an improvement of 9% over earlier reported	The other components of the posts such as the user's information , network information, the audio, and video content of the post may also be explored for

			results on the same dataset	cyberbullying detection tasks. Multi-lingual comments may be used with images to extract better features. Other optimization algorithms may also be explored for better feature engineering to improve the overall accuracy of cyberbullying post detection.

1 1	**Jalal Omer Atoum** Departme nt of Mathema tic s and Computer Science, East Central University Ada, Oklahoma	Cyberbullying Detection Through Sentiment Analysis(2020)	In this paper,we proposes a SA model for identifying cyberbullying texts in Twitter social media. SVM and NB are used in this model as supervised machine learning classification tools. In this research, we used data sets is a collection of tweets that have been classified into positive, negative, or neutral cyberbullying Before training and testing such machine learning techniques, the collected set of tweets have gone through several phases of cleaning, annotations, normalization, tokenization, named entity recognition, removing stopped words, stemming and	Machine learning methods in cyberbullying suffer from an inability to detect indirect language harassment. Unsupervised ML they may overlap and learn to localize texts with minimal unsupervised algorithms.

			n-gram, and features selection.	

1 2	**Caleb Ziems** Emory University **Ymir Vigfusson** Emory University **Fred Morstatter** USC Information Science Institute	Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification (2020)	In the paper, we characterize the problem of cyberbullying using five explicit factors to represent its social and linguistic aspects. Our classes distinguish cyberbullying from other related behaviors, such as isolated aggression or crude joking. To help annotators infer these distinctions, we provided them with the full context of each message's reply thread, along with a list of the author's most recent mentions. We show that these features improve the performance of standard text-based models. These results demonstrate the relevance of social network and language based measurements to account for the nuanced social characteristics of cyberbullying.	These lexically - trained classifiers will fall short of the more subtle goal of cyberbullying detection. Cyberbullying has been linked to negative mental health outcomes, including depression, anxiety, and other forms of self-harm, suicidal ideation, suicide attempts, and difficulties with social and emotional processing. Cyberbullying might not necessarily hold in all cases since Messages can be otherwise forwarded and publicly viewed without repeated actions. Cyberbullying may be prohibitively expensive to build out social networks for each user due to time constraints and the limitations of API calls.

				In harmful intent ,the target user causes them

				distress or harms their public image. Social network features may not be scalable for real world applications.

13	Ashwin Geet D'Sa *Université de Lorraine, CNRS, Inria, LORIA, F 54000* Nancy, France **Irina Illina** *Université de Lorraine, CNRS, Inria, LORIA, F 54000* Nancy, France **Dominique Fohr** *Université de Lorraine, CNRS, Inria, LORIA, F 54000* Nancy, France	BERT and fastText Embeddings for Automatic Detection of Toxic Speech et.al(2020)	Performs binary and multi-class classification using a Twitter corpus and study two approaches: (i) In feature-based approach, two steps are performed. First, each comment is represented as a sequence of words or word-pieces and for each word or word piece, an embedding is computed using fastText or BERT. Secondly, this sequence of embeddings will form the input to the DNN classifiers, that takes the final decision. CNN and Bi-LSTM models are used as classifiers. (ii) In fine-tuning approach, everything is done in a single step. Each comment is classified by a fine tuned BERT model. Classifies each comment as *non-toxic* or *toxic speech* for binary classification and *offensive*, *hate speech* or *neither* for multi-class classification.	The main confusions occurs between hate speech and offensive speech. This suggests that the model is biased towards classifying tweets as less hateful or offensive than the human annotators.

			Its evident that BERT fine-tuning	

			performed much	
14	**Jaideep Yadav** Computer Engineering Delhi Technological University **Devesh Kumar** Computer Engineering Delhi Technological University **Dheeraj Chauhan** Computer Engineering Delhi Technological University	Cyberbullying Detection using Pre-Trained BERT Model (2020)	Devicing methods for detecting cyberbullying on online platforms is a difficult task. The present study has used a pre trained BERT model , BERT-base model. When this pre trained BERT model was used on databases like Wikipedia, Twitter and Formspring, it was seen that greater accuracy of 96% was achieved in identifying cyberbullying compared to SMO, IBK, J48, JRipLR, MLP,RNN, Glove and CNN. The results show that the performance increases with the increasing number of bully posts as it is able to learn more about the bully sentences.	Better stable results require training of BERT model with larger datasets.
15	**Rupesh Kumar**	Detection of Cyberbullying using Machine Learning et.al (2021)	Detection of cyberbullying using Naive Bayes and SVM They classified the Twitter comments dataset into bullying or non-bullying labels. They have shown better results using SVM with n- grams.	A larger dataset is needed. Deep learning techniques are better than ML techniques in that case.

16	Manuel F. López Vizcaíno	Early detection of cyberbullying on social media networks. et.al(2021)	A lot of features have been considered: -Profile owner features -Media session features -Comment features -Video features	They should explore heterogeneous combinations of different machine learning models on the dual model.

			Two specific machine learning models, threshold and dual have been adopted and their behavior has been verified to conclude that the dual model(using 2 ML algos) works better.	For example, Extra Tree for the positive model, while Random Forest for the negative model. Second, they must extend the features regarding comments, as these concentrate most of the information for the early detection. Third, they should investigate an evaluation based on time, instead of the number of posts, since it may be relevant for the early detection of cyberbullying. They should experiment with other datasets from some other social media platforms to validate our approach and

				generalize the results.
17	**I Hsien Ting**	Towards the Detection of Cyberbullying Based on Social Network Mining Techniques et.al(2021)	In this paper, three techniques will be used, including keyword technique, opinion mining and social networks analysis. These techniques will be combined as an experimental approach for cyberbullying detection. SNM is very helpful for analyzing the data in social media, related	In the future work, we intend to tune the weight and hopefully the accuracy can be improved.

			techniques including Web Mining, Social Networks Analysis, Text Mining, Natural Language Processing, Sentimental Analysis and Opinion Mining. Sentimental Features Extraction more than 70% of cyberbullying posts can be detected correctly by using this approach, which is better than what we expected. In this approach data are collected from four major social networking websites and the evaluation results show an acceptable accuracy for cyberbullying detection.	

18	Mohamad Ehsan Basiri ,Shah la Nemati , Department of Computer Engineering , Shahrekord University, Iran. **Moloud Abdar** , Institute for Intelligent Systems Research and Innovation, Deakin University, Australia. **Erik Cambria,**	ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis(2021)	We propose an attention-based bidirectional CNN RNN deep model for sentiment analysis. We extract both past and future contexts by considering temporal information flow. The attention mechanism allows for putting more or less emphasis on different words. Experiments were conducted on five review and three Twitter datasets. Our model achieves state-of-the-art results on both long and short reviews.	They addressed the main short comings of previous methods which did not yield good performance when employed in different domain from the one they were trained on. Long distance between the sentiment polarity changes words in negative reviews may decrease the performance of the model.

	School of Computer Science and Engineering , Nanyang Technological University, Singapore **U.Rajendra Acharrya** Department of Electronics and Computer Engineering , Ngee Ann Polytechnic, Singapore.			

19	**Hong Fan** State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China **Mohammed A. A.** Alqaness State Key Laboratory for Information Engineering	Social Media Toxicity Classification Using Deep Learning: Real World Application UK Brexit et.al(2021)	Adopts the BERT to classify toxic comments from user generated data in social media, such as tweets. The BERT-base pre trained model was fine tuned on a well-known labeled toxic comment dataset, Kaggle public datasets. Moreover, the proposed model was tested on real-world data, two different tweets datasets, collected in two different periods based on a case study of the UK Brexit. The evaluation outcomes showed that BERT has the ability to classify and to predict toxic comments with a high accuracy rate. Moreover, a	Work could be done to make the model better suited to dealing with specific social media data. These tweets would have to be hand-labeled, which would take a fair amount of time to get enough data to increase the accuracy of the model. Aside from just collecting Twitter data, text from other social media sites like Facebook, YouTube, and Reddit could be added to improve the dataset.

	in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China **Abdelghani Dahou** LDDI Laboratory, Faculty of Science and Technology, University of Ahmed DRAIA, Adrar 01000, Algeria **Ahmed A. Ewees** Department of Computer, Damietta University, Damietta 34511, Egypt **Dalia Yousri** Electrical Engineering Department,		comparison was performed between the BERT-base model with three other models, called Multilingual BERT, RoBERTa, and DistilBERT. The BERT-base model outperformed all compared models and achieved the best results.	

	Faculty of Engineering , Fayoum University, Fayoum 63514, Egypt			
20	**Ashok Kumar J,**	Comment toxicity detection via a	Presents a multichannel convolutional	Multichannel attention

	Abirami S, Tina Esther Trueman Department of Information Science and Technology, Anna University, Chennai 600025, India **Erik Cambria** School of Computer Science and Engineering , Nanyang Technological University, Singapore	multichannel convolutional bidirectional gated recurrent unit et.al(2021)	bidirectional gated recurrent unit to categorize multilabel toxicities in online comments. The proposed model combines CNN and BiGRU in each channel to extract local features and long-term dependencies within comments using many filters and different kernel sizes. Results show that the proposed MCBi GRU model outperforms the existing results.	mechanisms in a distributed environment for multilabel toxic detection was not applied. Better training and testing accuracy was achieved using only n-gram word embeddings than the existing models

21	**Basavaraj N. Hiremath, Malini M. Patil**	Sarcasm Detection using Cognitive Features of Visual Data by Learning Model et.al(2021)	The objective of the paper is to detect sarcasm in human communication. The methodology uses basic cognitive features of human utterances by capturing three modes of data viz., voice, text, and temporal facial features. The Linguistic features of NLP methods help identify sentiment as negative and positive sentences based on polarity using the pre-labelled samples. The multiclass neural network model is used as a soft cognition method for the	The scope of future work is to conduct experiments on typical conversations with context-based, dialogues and benchmark datasets to be able to get fair learning performance and test them on language independent utterances to prove the impact of pragmatic/ cognitive features in human behaviour while uttering sentences.

			detection of sarcasm under cloud resources. The programming language Python and its OpenCV packages and the use of open source Praat have helped to build a unique framework with statistical accuracy checks.	

22	**Yong Fang** College of Cybersecuri ty, Sichuan Universit y, Chengdu 610065, China **Shaoshuai Yang** College of Cybersecuri ty, Sichuan Universit y, Chengdu 610065, China **Bin Zhao** ETC Avionics Co., Ltd., Chengdu 611731, China **Cheng Huang** College of Cybersecuri ty, Sichuan Universit y, Chengdu 610065, China	Cyberbullying Detection in Social Networks Using Bi-GRU with Self Attention Mechanism(et al 2021)	A sequence model was designed to automatically classify cyberbullying posts in social networks based on Bi-GRU deep neural network.The greatest advantage of the proposed approach is that Bi-GRU learns the word embedding sentence in both the directions.Here,GlOV E was used to do the word embedding instead of doing it randomly.To further improve the accuracy of classification,the self-attention mechanism along with GRU was used. Three datasets:Two from Twitter and one from Wikipedia were used to perform the binary classification for a given post:cyberbullying or non cyberbullying.Racism, Sexism,hate,attack,offe nsive were the different categories from these datasets all of which came under the class of cyberbullying.Althoug h the datasets were	Features such as user profile features and textual statistical features could be explored.Relationships between posts can be learnt from a broader perspective by applying graph neural networks.Oversam pling and undersampling could be tried and compared against the proposed model to confirm performance bias.

			imbalanced,no sampling techniques were used to avoid performance bias. The proposed method was compared against baseline approaches including TFIDF and SVM,LR parsers,Bi LSTM and attention.Further,it was also experimented by varying the number of layers in the GRU model from 1 to 3 layers.GPU usage for these variants of the proposed method.It was also experimented if self-attention was indeed necessary.	

CHAPTER 4 PROJECT DESCRIPTION 4.1 Proposed Design

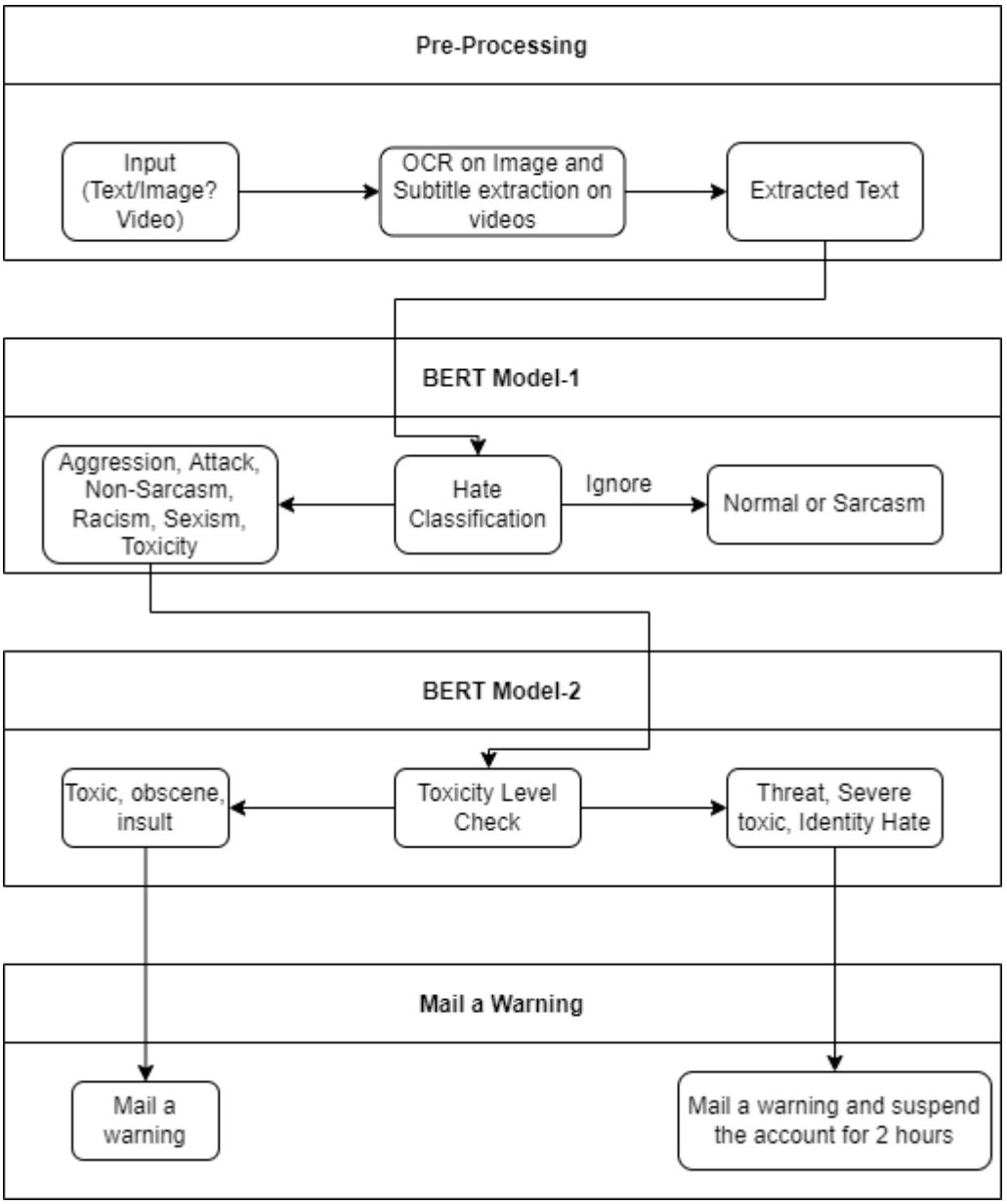

Fig 4.1: Proposed Design

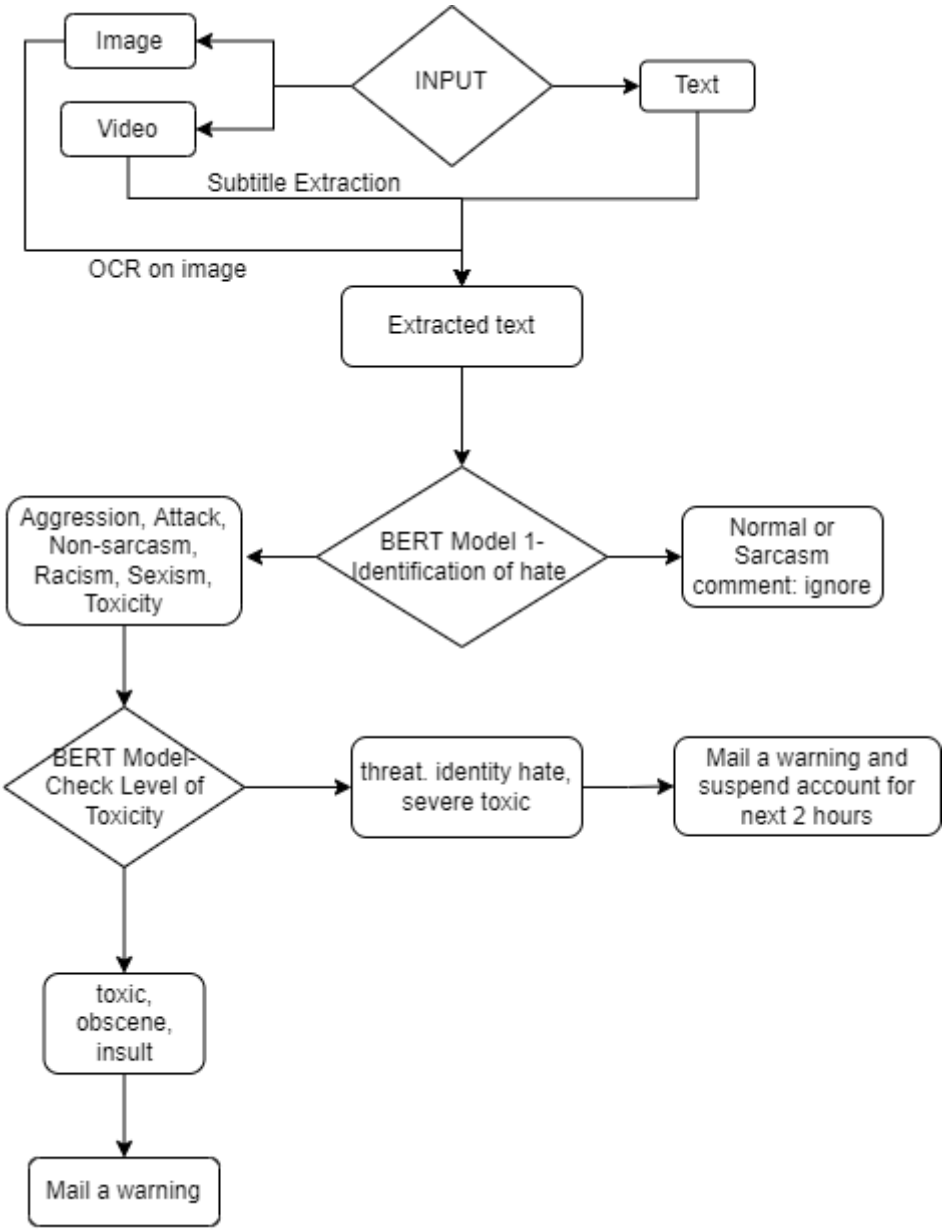

Fig 4.2: Flowchart of the model

4.2 Assumptions and Dependencies

It is assumed that we create and label of few datasets to categories uniqueness without any effect of privacy and license issues for our project. Look for educational

benchmark datasets. The following dataset is ear marked for testing without any license issues.

Dataset-1: Mendeley data for racism,sexism,aggression, toxicity and attack.

Dataset-2:Headache boy,GitHub-Data of multimodal sarcasm detection

Dataset-3: Toxic Comment Classification challenge dataset from Kaggle to classify the level of toxicity of the negative comment

CHAPTER 5 REQUIREMENTS

5.1 Functional Requirements

The functional requirements of Cyberbullying detection system include:

1.Analyzing the level of toxicity of the hate comments, performing hate content identification and identify it to be one of the following eight categories: aggression,attack,non-sarcasm,normal,racism,sarcasm,sexism and toxicity.If the content falls under normal or sarcasm ,no action will be taken otherwise the content will be passed to the second BERT model for toxicity level classification

2.Based on the level of toxicity of the comment required action is taken. on the user.

3.This functionality doesn't hold good only for textual comments, but also for image data which is analyzed by using memes where text is present in image data. Here OCR is used to extract the text from the image.

4.In case of a video,the subtitles of the video are obtained and this text is passed to the system.

5.Later the level of toxicity is estimated by using the BERT model.

6.The Flask frontend will have a login, so that once the user logs in and the comments to keep track of the user credentials.

7.BERT is an open-source machine learning framework for natural language

processing (NLP). BERT is designed to help computers understand the meaning of ambiguous language in the text by using surrounding text to establish context. The BERT framework was pre-trained using text from Wikipedia and can be fine-tuned with a question-and-answer datasets.

8.Flask is a micro web framework written in Python. In this project, a simple front end is developed using Flask. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

5.2 Non-functional requirements

The non-functional requirements of our system include:

Scalability:

1. In this section, we first demonstrate the way our proposed system scales when it has to deal with a substantial number of media sessions.

2. Acceptable responsiveness of the system is our primary goal along with the scalability of the system.

3. In these experiments, we decided that an average alert time under 2 hours is acceptable, which means an alert will be within 2 hours of a cyberbullying instance

Reliability:

1. Our system is reliable as we use a novice approach using BERT to detect cyberbullying.

2. Further, we use two levels of classification to confirm cyberbullying-hate content detection and toxicity detection.

3. The system also warns the user and takes appropriate action for highly toxic content posted.

4. Users who are super-bullies will also be blocked for 2 hours as a sign of penalty.

CHAPTER 6 DELIVERABLES

End to end social media website application developed using React for front-end, Flask for back-end and MongoDB database to save records of users, posts and comments will be delivered. The website will also use BERT models at the backend to detect cyberbullying.It also warns bullies according to the level of toxicity using a warning mail or by blocking their account.

REFERENCES

[1] Ting, I-Hsien, et al. "Towards the detection of cyberbullying based on social network mining techniques." 2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC). IEEE, 2017.

[2] Basiri, Mohammad Ehsan, et al. "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis." Future Generation Computer Systems 115 (2021): 279-294.

[3] Fan, Hong, et al. "Social Media Toxicity Classification Using Deep Learning:

Real-World Application UK Brexit." Electronics 10.11 (2021): 1332. [4] Kumar, Ashok, et al. "Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit." Neurocomputing 441 (2021): 272- 278.

[5] Aslam, Fankar Armash, et al. "Efficient way of web development using python and flask." International Journal of Advanced Research in Computer Science 6.2 (2015): 54-57.

[6] Vogel, Patrick, et al. "A low-effort analytics platform for visualizing evolving Flask-based Python web services." 2017 IEEE Working Conference on Software Visualization (VISSOFT). IEEE, 2017.

[7] Riadi, Imam. "Detection of cyberbullying on social media using data mining techniques." International Journal of Computer Science and Information Security (IJCSIS) 15.3 (2017).

[8] Agrawal, Sweta, and Amit Awekar. "Deep learning for detecting cyberbullying across multiple social media platforms." European conference on information retrieval. Springer, Cham, 2018.

[9] Yao, Mengfan, Charalampos Chelmis, and Daphney? Stavroula Zois. "Cyberbullying ends here: Towards robust detection of cyberbullying in social media." The World Wide Web Conference. 2019.

[10] Yao, Mengfan. "Robust detection of cyberbullying in social media." Companion Proceedings of The 2019 World Wide Web Conference. 2019.

[11] Lu, Nijia, et al. "Cyberbullying detection in social media text based on character‐level convolutional neural network with shortcuts." Concurrency and Computation: Practice and Experience 32.23 (2020): e5627.

[12] Van Bruwaene, David, Qianjia Huang, and Diana Inkpen. "A multi-platform dataset for detecting cyberbullying in social media." Language Resources and Evaluation 54.4 (2020): 851-874.

[13] Rezvani, Nabi, Amin Beheshti, and Alireza Tabebordbar. "Linking textual and contextual features for intelligent cyberbullying detection in social media." Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia. 2020.

[14] Kumari, Kirti, and Jyoti Prakash Singh. "Identification of cyberbullying on multi‐modal social media posts using genetic algorithm." Transactions on Emerging Telecommunications Technologies 32.2 (2021): e3907.

[15] Atoum, Jalal Omer. "Cyberbullying Detection Through Sentiment Analysis." 2020 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, 2020.

[16] Ziems, Caleb, Ymir Vigfusson, and Fred Morstatter. "Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 14. 2020.

[17] d'Sa, Ashwin Geet, Irina Illina, and Dominique Fohr. "Bert and fasttext embeddings for automatic detection of toxic speech." 2020 International Multi Conference on:“Organization of Knowledge and Advanced Technologies”(OCTA). IEEE, 2020.

[18] Yadav, Jaideep, Devesh Kumar, and Dheeraj Chauhan. "Cyberbullying Detection using Pre-Trained BERT Model." 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2020.

[19] Kumar, Rupesh. "Detection of Cyberbullying using Machine Learning." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12.9 (2021): 656-661.

[20] López-Vizcaíno, Manuel F., et al. "Early detection of cyberbullying on social media networks." Future Generation Computer Systems 118 (2021): 219-229. [21] Hiremath, Basavaraj N., and Malini M. Patil. "Sarcasm Detection using Cognitive Features of Visual Data by Learning Model." Expert Systems with Applications 184 (2021): 115476.
[22] Fang, Yong, et al. "Cyberbullying detection in social networks using Bi-gru with self-attention mechanism." Information 12.4 (2021): 171.