USING MACHINE LEARNING ALGORITHMS: A LITERATURE SURVEY

1stSrinadh Unnava
(*Associate professor*) IT,
SITE,Tadepalligudem.

2nd Chaithrasri Jasthi
(*Student*) IT,
SITE,Tadepalligudem.

3rd Manikanta Bandla
(*Student*) IT,
SITE,Tadepalligudem.

4thAravind Vemala
(*Student*) IT,
SITE,Tadepalligudem.

5th B.R.Manikanta Garapati
(*Student*) IT,
SITE,Tadepalligudem.

ABSTRACT:

In recent years, the exponential growth of abusive,offensive text has been met by rapid advances in text classification techniques. A newly proposed machine learning algorithm leverages the latest advances in deep learning techniques to enable automatic extraction of expressive features.Bullies use a variety of networking avenues to target victims with offensive comments and posts.There are various algorithms in machine learning that can help detect Cyber Bullying, and some algorithms are better than others.Additionally, Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), Stochastic Gradient Descent (SGD), Random Forest (RF), Adaboost (ADB), Naive Bayes (NB), and Support Vector Machine (SVM). In this paper each of these algorithms had been evaluated using accuracy, precision, recall, and F1 score as performance metrics to determine the detection rate of the classifier applied to the dataset.

Keywords—Cyber bullying, Abusive detection, text classification, deep learning, machine learning, natural language processing.

I. INTRODUCTION

Cyber Bullying is a major cause of concern as it affects people severely. Social Media are safe places to communicate, but they are prone to cyberbullying. Humiliation is more dangerous than traditional bullying because it appears to an unlimited online audience. We don't need the victim's physical appearance, so we can continue without a break. Many of his networking sites do not require a real name to register as a user. Victims of bullying lose self-confidence, become antisocial, and have a negative impact on their mental health. This will make you aware of cyberbullying.A variety of standard definitions for classification of tasks exist in the NLP research area, often used as benchmarks to evaluate new methods. We outline the main representatives, approximately following the taxonomy proposed by M. Ameer Ali et al[1].Social networks are the main perpetrators of cyberbullying. The dynamic nature of these sites has led to an increase in aggressive behavior online. The anonymity of user profiles makes identifying bullies more complicated. Social media are popular because of their connectivity in the form of networks. However, this can be detrimental when rumors and bullying posts spread on networks that are not easily controlled[2].Internet and social media use has obvious benefits for society, but frequent use can also have significant negative effects. This includes unwanted sexual exposure, cybercrime, and Cyber Bullying. We developed a model to detect Cyber Bullying behavior and its severity on Twitter. Feature generation using PMI in the preprocessing phase is an efficient way to handle class imbalance in binary and multiclass classification where minority class misclassification leads to higher cost in terms of recognition model reliability. proven to be a method[3].The government has laws and systems in place to limit Cyber Bullying, but against Bullers.it is important to identify the perpetrators of the bullying, so action should be taken. is difficult. The proposed system focus on detecting the presence of Cyber Bullying activity on social networks and classifying them using the Levenshtein algorithm and his Naive Bayes classifier[4].This paper provided an overview of various Cyber Bullying research contributions, datasets used, findings and research gaps.

II. BASIC CONCEPTS AND TERMS

A. *Cyber Bullying*

First, cyber bullying or cyberharassment is a form of bullying or harassment using electronic means. Cyberbullying and cyber harassment are also called online bullying. as the digital realm expands and technology advances, it is becoming increasingly common, especially among teenagers.bullying or harassing other people in spaces, especially on social media sites.

B. *Abusive*

Several terms have been developed to define cyber abuse (which may or may not be sexual in nature), including cyber harassment, cyber stalking, cyber bullying, digital abuse, and technology-assisted abuse. used. A common factor is the use of technology to establish power and control by inducing fear and intimidation.

C. *Hateraids*

On Twitch and other live streaming services, a hate attack is a situation where a stream is "ambushed" by multiple viewers at the same time, flooding the chat with harassing or hateful messages, and preventing streamers from completing the stream. .

III. LITERATURE SURVEY

Nine different feature types and eight different machine learning classifiers were examined. The most robust result in our analysis was the contribution of the BERT-based model not only in the in-domain but also in the out-of-domain evaluation. Among structural features, key terms perform much better than others[6]. The Support Vector Machine (SVM) achieved the best results in terms of sensitivity, specificity, precision, false positive rate, and precision. We therefore concluded that the machine learning approach (support vector machine) can be reliably used for classifying Twitter data[7].Twitter is an amazing data source used by people from all over the world to share their opinions on various topics. As a result, it provides researchers with a huge platform for obtaining vast amounts of coarse information. A crude processing of this information is used to analyze the user's opinion. Based on this research, I explored different types of the classification algorithms for text analysis. Data mining techniques are used to extract the text from the given pieces of the data. For this reason, text is classified as positive, negative, or neutral[8].The model should be trained using a set of data collected. The approach can be applied to unknown data after training. Approaches such as Naive Bayes, Support Vector Machines, Decision Tree, Random Forest, and Maximum Entropy have been implemented. We mainly focus on the Naive Bayes algorithm and compare the results with other models in terms of accuracy, recall, and F-score. Of all these, decision tree has the highest score[9].For a total of new machine learning-based data/text models, we observed that those trained on relevant data were much more accurate than classifications based on standard dictionaries. This is because the observed text in the tweets is very informal and does not use standard grammatical rules or spelling, so the data here are very unstructured[10].An ensemble classifier that detects hate speech in short texts such as tweets. The input to the base classifier consists not only of standard word unigrams, but also a set of traits representing each user's past propensity to post harassing her messages. More specifically, our scheme can successfully distinguish racist and sexist messages from regular text and achieve higher classification quality than current state-of-the-art algorithms[11].Additionally, it is important to highlight the difficulty of configuring all neural network parameters and how slow the training process can be. And looking to the future, the project could be improved in many ways. By trying the system on larger datasets and implementing more preprocessing techniques such as token normalization, emoji to text conversion, and lemmatization. Run and compare the results with those obtained with the deep learning model[12]Tests also showed that the evaluation of message order strongly affects the final result of the classifier. This can be seen from the high standard derivatives of the tests with different starting values. This is normal process. This is because when a large sequence of messages is input with the same label as the input, the weight adjustment and the learning factor (alpha) adjust the learning with other inputs, causing imbalanced weights. Therefore, the Linear SVM requires balanced inputs for good results. Setting the parameters for this algorithm turned out to be a rather difficult task[13] There are various benchmarks based on demographics, social impact, and cultural factors. We propose a deep learning model based on Transformer context embedding and HateBERT architecture. We preprocessed tweets from the HASOC 2021 dataset, extracted feature embeddings, and trained the system to classify as hate speech with a macro F1 score of 79%. The compiled work demonstrates the extent to which HateBERT is further deployed in experiments and optimized for performance by focusing on novel built-in combinatorial and ensemble approaches[14].The input to the base classifier consists not only of standard word unigrams, but also of a set of traits representing the historical propensity of each user to post harassing messages. Our main innovations are: i) a deep learning architecture that implements the above features using frequency vectorization of words, ii) an experimental evaluation of the above models against a public dataset of tagged tweets, iii) an open-source, racist There are even , which are pretty common in posts like this. Therefore, when aiming to build a language-agnostic solution, word frequency vectorization is a better choice than the pretrained word embeddings used in previous work. We believe that deep learning models can classify high probability texts or analyze general sentiments. In our opinion, the classification algorithm still has room for improvement[15].Here Theyazn H.H. Aldhyani et al. Built to improve the Cyber Bullying Detection System that can be used to analyze and eradicate cases of online bullying by social media users. Cyber BullyingBefore Hate Deep learning classifiers have been developed to detect online tweets and discussion content, and can be applied to the design of Cyber Bullying detection systems for online social media sharing platforms such as Twitter and Facebook. increase.[16].An LTSM-based classification system that distinguishes between hate speech and offensive language. This system describes a modern approach to identifying hate speech on Twitter using word embeddings with LSTM and Bi-LSTM neural networks. The best performing LSTM network classifier achieved 86% accuracy with early stopping criteria based on the loss function during training[17].Along with identifying offensive words. Despite the fact that messages may contain profanity, but are not offensive, it is intuitively clear that combinations of insults within messages occur more frequently. So in this direction. Continued research looks promising to us[18].

IV. NLP TECHNIQUES FOR CLASSIFICATION OF BINERY TEXT

In this section, we analyze the highlights presented in the computationally focused article relevant to discourse discovery of disdain and various research focused on related ideas. Finding good highlights for grouping problems may be one of 's most difficult tasks when working with AI. We then divide this particular segment to show the highlights officially used by other authors. Divide highlights into two categories. A common highlight used in content mining, which occurs regularly in other content mining areas. And the detection of specific aversive discourse was found in the aversive discourse discovery report that was characteristically identified with that topical feature. We present research in this area.

A. *Feature in text analysis*

Most of the articles we found attempt to adapt the methods reliably known in content mining to the specific problem of localization of programmed hate speech. Common highlights are characterized as highlights that are regularly used in content mining. Start in the simplest way with word references and vocabulary.

B. *Lexicons*

One system of content mining is the use of dictionaries. This methodology consists of creating lecture summaries (lexicons). It is displayed and included in the content. The defined numbers can legally be used as highlights or for processing scores. Hate speech was detected, so this was guided using: 414 specific words with abbreviations and contractions. Most of it is descriptive, things are representational. Ortony's vocabulary consists of a list of talks denoting connotations, only one of every strangely reckless remarks contains inherent disrespect and can be equally uncertain Considering the 17 total of words in each comment, forward. Additionally, it is conceivable to use such methodologies in regular articulations.

C. *Distance Metrics*

Some research has highlighted that the hostile words in instant messages may be hidden by intentional misspellings, usually single-letter substitution 18. or homophobic. For example, "joo" for . Lowenstein separation.

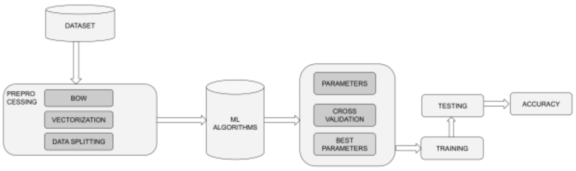
TABLE 1:Survey of papers

Title	Accuracy	Precision	Recall	F-score	Algorithms
Supervisedfeature selection based extreme learning machine (SFS-ELM) classifier for cyber bullying detection in twitter	0.75	0.77	0.80	0.786	CNN model layers
AggressiveTweets, Bully and Bully Target Identification from Multilingual Indian Tweets	0.742	0.7255	0.738	0.73	LSTM
CyberBullying Detection and Classification Using Information Retrieval Algorithm	0.81	0.79	0.77	0.79	NBC
Identification and Detection of Cyber Bullying on Facebook Using MachineLearning Algorithms	0.73 0.72	0.72 0.73	0.99 1.0	0.84 0.843	NV, KNN
Offensive Language Recognition in SocialMedia	-	-	-	0.63, 0.62, 0.68, 0.57	Logistic Regression (LR), NB, SVM, Ensembled Technique
CyberBullying Comment Classification on Indonesian Selebgram Using Support Vector Machine Method	0.7942	-	-	-	SVM

Identification of Good and BadNews on Twitter	-	-	-	0.92 0.55 0.84 0.67	BERT, SVC, LSVC, and LR
Classification of Hate, Offensive and Profane content from Tweets using an Ensemble of Deep Contextualized and Domain SpecificRepresentations	0.80, 0.81 0.81	0.80 0.81 0.812	0.77 0.78 0.79	0.75	ERNIE 2.0, TwitterRobertaOf, hateBERT
Detecting Offensive Language in Tweets Using Deep Learning		0.91 0.92 0.934	0.92 0.931 0.935	0.91 0.924 0.914	single classifier,ensemble LSTM + Random Embedding
Automatic offensive language detection from Twitter data using machine learning and feature selection of meta data	0.900 0.922	0.833 0.899	1.0 0.96	0.92 0.924	NBC, SVM
Identifying ,and Categorizing Offensive Language in tweets using Machine Learning	0.76 0.712 0.75 0.801			0.66 0.71 0.75 0.80	LR, RNN LSTM, CNN, Ensmble
Twitter Data Classification by Applying Multiple Machine Learning Techniques	0.49 0.64 0.80 0.83	0.20 0.56 0.91 0.92		0.25 0.13 0.06 0.05	RF KNN NB LR
Cyber Bullying Detection ,and Classification using Multinomial Naïve Bayes and Fuzzy Logic	0.88	-	-	0.82	MNVB

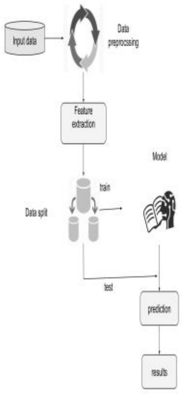
V. METHODOLOGY

A. *Flow chat*



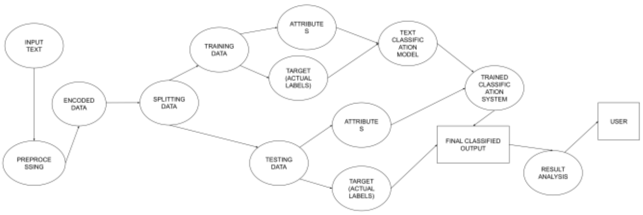
Figure(a)

B. *Life cycle*



Figure(b)

C. Architecture



Figure(c)

VI. CONCLUDING SURVEY

English is the most widely spoken language at and Twitter is the preferred source of information for businesses. We have argued that the author did not use her open records and did not distribute the most recent records collected. This complicates the analysis of results and consequences. Surveys and surveys of relative are very rare in this area. Finally, regarding the highlights used, we found that most studies.

VII. RESOURCES

Outline basic data about the records and collections found. There are now several files and collections on derogatory discourse, so nothing is regulated there. The goal is to monitor whether any of the tasks with access to aversion-oriented information can be used as models or hotspots for annotated information. To do this, I evaluated GitHub using the "Despise Discourse" articulation of the accessible web index. A company search was done on GitHub.

VIII. CONCLUSION

To simplify the way for machine learning researchers we tried to study papers in which contains various machine learning algorithms that contributed to cyber bullying text classification. We found a set of best algorithms from all papers we seen. This work would not bias towards the methodology illustrated in this literature review. We focused only on the algorithms were used and their Accuracy levels, Precision, Recall, F-score. We supposed that the outcome of this study can reduce the time of the practitioner and researcher to choose the best algorithm while predicting the cyber bullying words. Process to detect cyber bullying words are mentioned in which included data collection, data pre-processing, feature extraction, feature selection, and finally classification. We have seen neural networks models give slightly better performance than other models in few papers.

REFERENCES

[1] manage Offensive Text in Social Media - A Text Classification Approach using LSTM-BOOST,Md. Anwar Hussen Waduda , Muhammad Mohsin Kabir a , M.F. Mridha b,∗ , M. Ameer Ali a.

[2] Muskan Patidar1 , Mahak Lathi2 , Manali Jain3 , Monika Dhakad4 , Prof. Yamini Barge,Cyber Bullying Detection for Twitter in international journal for research(ijrs).

[3] Bandeh Ali TalpurID1 *, Declan O'Sullivan2 in Cyber Bullying severity detection: A machine learning approach in ploseone

[4] B. Sri Nandhini ,.I. Sheeba ,Cyber Bullying Detection and Classification Using Information Retrieval Algorithm of ICARCSET '

[5] Nureni Ayofe AZEEZ, Department of Computer Sciences, University of Lagos, Nigeria Sanjay Misra, in Identification and Detection of Cyber Bullying on Facebook Using Machine Learning Algorithms Journal of Cases on Information Technology

[6] Arnisha Akhter, Uzzal K. Acharjee, Md Masbaul A. Polash,Cyber Bullying Detection and Classification using Multinomial Naïve Bayes and Fuzzy Logic inI.J. Mathematical Sciences and Computing

[7] Twitter Data Classification by Applying and Comparing Multiple Machine Learning Techniques Ananya Sarker, Md. Shahid Uz Zaman, Md. Azmain Yakin Srizo International Journal of Innovative Research in Computer Science & Technology (IJIRCST)

[8] pratima Deshpande1 , Purva Joshi2 , Diptee Madekar3 , Pratiksha Pawar4 , Prof. M.D. Salunke5 Classification of Twitter data in Asian Journal of Convergence in Technology

[9] Sentiment Analysis of Twitter Data Using Machine Learning Approaches by Bandaru Mounika & Dr. M.S.V.S Bhadri Raju International Journal of Research

[10] 1N V S Sowjanya, 2K Sunil Kumar, 3N Tejaswi,4V Sesha Chandra Sai,5S V Sai Krishna of TEXT CLASSIFICATION ON TWITTER DATA in IJCRT.

[11] Georgios K. Pitsilis, Heri Ramampiaro and Helge Langseth of Detecting Offensive Language in Tweets Using Deep Learning in arxiv

[12] Identifying and Categorizing Offensive Language in tweets using Machine Learning in Berta Viñas Redondo

[13] Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata in IEEE t Gabriel Araujo De Souza, Marjory Da Costa-Abre

[14] Detecting Offensive Language in Tweets Using Deep Learning,Georgios K. Pitsilis, Heri Ramampiaro and Helge Langseth in arxiv

[15] Basavraj Chinagundi1 , Muskaan Singh2 , Tirthankar Ghosal2 , Prashant Singh Rana1 and Guneet Singh Kohli Classification of Hate, Offensive and Profane content from Tweets using an Ensemble of Deep Contextualized and Domain Specific Representations in Forum for Information Retrieval Evaluation

[16] Theyazn H. H. Aldhyani 1,* , Mosleh Hmoud Al-Adhaileh 2 and Saleh Nagi Alsubari 3 .Cyber Bullying Identification System, Based Deep

[17] Cyber Bullying comment classification on Indonesian Selebgram using support vector machine method November 2017 DOI:10.1109/IAC.2017.8280617 Conference: 2017 Second International Conference on Informatics and Computing (ICIC)

[18] Offensive Language Recognition in Social Media Elena Shushkevich1 , John Cardiff1 , Paolo Rosso2 , Liliya Akhtyamova1

[19] Cyberbully: Aggressive Tweets, Bully and Bully Target Identification from Multilingual Indian Tweets by suman karan

[20] How can we manage Offensive Text in Social Media -A Text Classification Approach using LSTM-BOOST July 2022International Journal of Information Management 2(2):100095 DOI:10.1016/j.jjimei.2022.100095.