Implementing A Scheme That Uses Recurrent Neural Networks (RNNS) And Linear Support Vector Machines (LSVM) To Uncovers Online Bullying on Social Media Platforms

Belonwu,Tochukwu S[1], Okeke, Ogochukwu C[2]
[1]*Lecturer, Department of Computer Science, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria*
[2]*Lecturer, Department of Computer Science, Chukwuemeka Odumegwuojukwu University, Uli, Anambra State, Nigeria*
Corresponding Author: ts.belonwu@unizik.edu.ng

Abstract— **In recent years, one of the main challenges in establishing a safe online environment has been the development of a plan to identify and handle online bullying on social media platforms. The plan outlines a study that used a combination of recurrent neural networks (RNNs) and linear support vector machines (LSVM) to create a cyberbullying detection system. The idea was to develop a system that could discover and identify people who were bullying others online. Using a dataset of text expressions from Twitter and Facebook, the study's algorithm successfully classified instances of cyberbullying; the LSVM achieved an 85% classification accuracy, while the RNN achieved an 81% feature extraction accuracy. This performed better than both deep learning and conventional machine learning techniques. The LSVM and RNN algorithms, according to the authors, have the ability to effectively handle the issue of online cyberbullying. They make a number of suggestions to improve the system even further, including developing more user-friendly interfaces, enhancing dataset diversity, giving data protection first priority, and upgrading the model on a regular basis. The authors also recommend several avenues for further research, such as combining multimodal approaches, putting real-time monitoring and intervention into practice, analyzing data across platforms, and assessing the system's long-term effects. All things considered, the study shows how promising it is to use cutting-edge AI methods like LSVM and RNNs to create practical answers to the urgent problem of cyberbullying.**

Index Terms— **Bulling, Cyber Bulling, Online Bulling, Machine Learning, Prediction, Tracking.**

1. Introduction

In today's society, the internet has had a significant impact on communication and social connections. Children and adolescents are using the internet more regularly, at early ages, and in more varied ways (e.g. smartphones, laptops and tablets). Even though the majority of adolescents' Internet use is inoffensive, and the advantages of digital information exchange are obvious, the liberty and untraceability encountered online makes younger folks vulnerable, with 'cyberbullying' as the most serious threats. Bullying is not a recent concept, and online bullying emerged as digital innovations became major communication aids (Gottschalk, 2019). On the plus side, social communication channels such as blogs, social media sites (such as Facebook, twitter), and online messaging channels (such as WhatsApp) allow you to interact with anybody and at any moment. Furthermore, they are a setting where individuals interact socially, providing opportunities to form new connections and retain social bonds. On the detrimental side, social networking increases the risk of people, especially teenagers and children, being exposed to potentially dangerous situations such as recruitment or sexually subversive conducts, signs of suicidal thoughts, depression and online bullying. Audiences are accessible 24 hours a day, seven days a week, and can frequently stay anonymous if preferred, making digital networks a comfortable option for abusers to identify their victims other than of the school ground (El Asam, Samara & Terry, 2019).

The Internet has also opened up previously unseen avenues for human communication and socialisation. Social networking, specifically, has grown in popularity over the last decade (Lane & Stuart, 2022). People are linking and conversing in ways that were previously unimaginable, thanks to platforms such as MySpace, Facebook, Twitter, whatsapp, Flickr, TikTok and Instagram (Price-Mitchell, 2020).

Manuscript revised June 04, 2024; accepted June 05, 2024. Date of publication June 07, 2024.
This paper available online at www.ijprse.com
ISSN (Online): 2582-7898; SJIF: 5.59

The pervasive use of social media by individuals of all ages generated a massive amount of data for a variety of research themes, including recommender systems, link predictions, visualisation, and social network investigation (Ianni, Masciari & Sperlí, 2021). Whereas the presence of social media has established a great platform for information exchange, it has also opened up a new framework for malicious purposes such as 'spamming, trolling, and online bullying' (Cuncic, 2022). Cyberbullying, as defined by the 'Cyberbullying Research Centre (CRC)', happens whenever anyone uses technology to transmit messages that verbally abuse, victimise, or threaten another individual or a group (Polak & Trottier, 2020). Unlike traditional forms of bullying, where hostility is a brief and momentary face-to-face incident, bullying texts are always active in social media, can be accessed globally, and are frequently irreversible (Unicef, 2020).

Mining social media sites for online bullying presents a number of difficulties and issues. It is difficult to precisely perceive users' motives and connotations in social media based solely on their texts (like posts, tweets, comments), which are usually brief, use colloquialisms, and may contain multimedia such as videos and images (Emmery et al., 2021). Twitter, for instance, restricts its users' texts to 140 characters, which can include text, colloquial language, emoticons, and animations. As a result, determining the emotion of a message can be difficult (Ali & Kurdy, 2022). Twitter, together with Facebook present the best opportunity for mining texts because they present a large textual base for the analysis, and also, they present the opportunity to analyse bystanders in the process. Furthermore, bullying may be difficult to detect if the bully prefers to conceal it with schemes such as sarcasm or silent aggression. Furthermore, the massive volume, changeable, and complicated system of social media platforms.

Make it challenging to pinpoint cyberbullies. Just to illustrate, every day, countless millions of tweets are posted on Twitter (Aroyehun, & Gelbukh, 2018). Cyberbullying is the use of virtual worlds to bully others either known or unknown to the bully. Flooding occurs when the bully repeatedly sends the same remark, nonsense comments, or presses the enter key in order to prevent the victim from contributing to the discussion. Masquerading is when a bully pretends to be someone they are not. A type of online fight is flaming or bashing. The bully sends or posts enticingly insulting, vulgar electronic messages to one or more people, either privately or publicly to an online group. Harassment is a type of conversation in which the bully sends insulting and rude messages to the victim. When a poster sends intimidating or offensive messages, it is considered cyberstalking or cyber threat.

Denigration, also known as dissing, occurs when a cyberbully sends or publishes gossip or false statements about a victim in order to harm the victim's reputation. When a person sends or publishes private or embarrassing information in a public chat room or forum, this is known as outing. This type of cyberbullying is analogous to denigration. However, in an outing, the bully and victim have a close relationship.

2. Aim and Objectives

A. *Aim*

This project aims to model a scheme that uncovers online bullying on social platforms, using Linear Support Vector Machines (LSVM) and Recurrent Neural Networks (RNNs).

B. *Objectives*

- Within the first three months, conduct a thorough literature review on the subject of using Machine learning in detecting social media conversations that are ridden with both explicit and implicit negative conversations like harassment, bullying, swearing words and racism slurs contained in text messages;
- To utilise a combination of machine learning algorithms (LSVM & (RNNs) to classify, cluster, recognise patterns, extract features and predict conversations with bullying cues and flag them up for monitoring in a machine learning training process.
- To develop and implement a software model that detects and predicts cyberbullying texts present in conversation threads with Machine learning Algorithms running in the background.
- To train the new model and generate results, while also running other known ML algorithms to compare the new model's performance to previous ones.
- To present the model's results in a clear and human-readable format, as well as to make recommendations based on them

3. Mathematical Model

The mathematical model formulation for the scheme that uncovers online bullying on social platforms using Linear Support Vector Machines (LSVMs) and Recurrent Neural Networks (RNNs):

A. *Data Collection*

Let $D = \{(x_i, y_i) \mid i = 1, 2, ..., n\}$ be the dataset, where x_i is the feature vector representing the i-th social media post, and $y_i \in \{-1, 1\}$ is the corresponding label (1 for bullying, -1 for non-bullying).

B. *Linear Support Vector Machine (LSVM):*

The LSVM optimization problem can be formulated as:

$$\min (1/2) * \|w\|^2 + C * \Sigma(\max(0, 1 - y_i * (w^T * x_i + b)))$$

where:

w is the weight vector
b is the bias term
y_i is the label of the i-th sample
x_i is the feature vector of the i-th sample
C is the regularization parameter

The decision function for the LSVM model is:

$$f(x) = \text{sign}(w^T * x + b)$$

where $\text{sign}(\cdot)$ is the sign function.

C. Recurrent Neural Network (RNN)

Let the input sequence be $X = (x_1, x_2, ..., x_T)$,

where

x_t is the word embedding of the t-th word in the social media post.

The RNN hidden state at time step t is computed as:

$h_t = f(W_h * h_{t-1} + W_x * x_t + b_h)$

where:

h_t is the hidden state at time step t

h_{t-1} is the hidden state at the previous time step

W_h is the weight matrix for the hidden state

W_x is the weight matrix for the input

b_h is the bias term

$f(\cdot)$ is the activation function (e.g., tanh, LSTM, GRU)

The output of the RNN at the final time step T is used for classification:

$y_{pred} = g(W_o * h_T + b_o)$

where:

y_{pred} is the predicted label

W_o is the weight matrix for the output

b_o is the bias term

$g(\cdot)$ is the activation function for the output (e.g., sigmoid, softmax)

The RNN is trained using backpropagation through time (BPTT) and an appropriate loss function, such as categorical cross-entropy:

$L = -\Sigma\, y_i * \log(y_{pred\_i}) + (1 - y_i) * \log(1 - y_{pred\_i})$

This mathematical model provides the core formulations for the LSVM and RNN components, as well as the ensemble modeling approach. The combination of these techniques can effectively uncover online bullying on social platforms by leveraging the strengths of both traditional machine learning and deep learning methods.

4. Results And Discussion

The proposed system presents a cyber bullying identification system. The system uses a scheme that uncovers online bullying on social platforms, with a focus on Facebook and Twitter, using a Linear Support Vector Machine and a Recurrent Neural Network for the best results.

The system allows users to sign up on the platform and chart with other users. At the back-end, bullying keywords are stored in the knowledge base. This enables the system to filter ongoing bullying in a chat between users.

The software will work both in online and offline modes. It can also work in real time. In the online mode, the system will flag all bullying trigger words as conversations go on in the chosen social media platforms, while in the offline modes, conversation threads will be logged and the software will be deployed to run on the logged conversations in search of bullying keywords, which the system will flag if and when it comes across them.

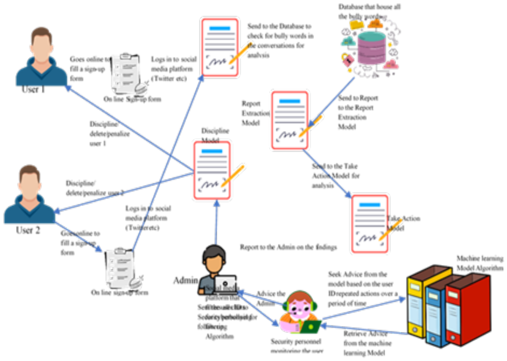
User 1
Goes online to fill a sign-up form
On line Sign-up form
Logs in to social media platform (Twitter etc)
Send to the Database to check for bully words in the conversations for analysis
Database that house all the bully wording
Discipline delete/penalize user 1
Discipline Model
Report Extraction Model
Send to Report to the Report Extraction Model
Send to the Take Action Model for analysis
Take Action Model
User 2
Discipline/ delete/penalize user 2
Logs in to social media platform (Twitter etc)
Goes online to fill a sign-up form
On line sign-up form
Report to the Admin on the findings
Admin
Social media platform that
Algorithm
Advice the Admin
Security personnel monitoring the user
Seek Advice from the model based on the user IDrepeated actions over a period of time
Machine learning Model Algorithm
Retrieve Advice from the machine learning Model

Fig.1. System architecture

The study aimed to tackle the problem of cyberbullying, namely on popular social media sites such as Facebook and Twitter. This study emphasises the notable influence of digital technology on the communication and social interactions of children and adolescents. Although these platforms facilitate convenient and extensive communication, they also subject individuals, particularly young people, to possible hazards such as cyberbullying, solicitation for dangerous activities, and indications of mental health problems.

In order to effectively address cyberbullying, the study suggests creating and applying a software model that utilises machine learning algorithms, specifically Linear Support Vector Machines (LSVM) and Recurrent Neural Networks (RNNs), to identify and anticipate instances of cyberbullying in conversation threads. The objective was to train the model using comprehensive and diverse datasets and evaluate its effectiveness in comparison to existing detection methods, particularly on Facebook and Twitter platforms.

The research highlighted the necessity of a comprehensive approach that integrates technological advancements with human engagement to tackle the dynamic nature of cyberbullying. This encompasses the creation of sophisticated algorithms and machine learning models that can detect nuanced forms of cyberbullying, advocating for digital literacy, and cultivating empathy to promote responsible online conduct.

The work includes a comprehensive evaluation of literature that investigated the application of machine learning methods for identifying cyberbullying on Twitter and Facebook. The review covered a wide range of algorithms, including traditional machine learning techniques as well as deep learning and transfer learning. The Object-Oriented Analysis and Design Methodology (OOADM) is used to design a web application that imports real-time communications from social media platforms, applies a model to detect characteristics of cyberbullying, and presents the results in a user-friendly fashion.

The suggested method provides benefits such as rapid and automated detection, hence minimising the requirement for considerable manual labour. It serves users such as schools, parents, law enforcement agencies, and social media platform users, enabling them to successfully battle cyberbullying. The system incorporates functionalities such as charting and data sharing, email modules, conversation monitoring, and automated tracking systems.

The study highlights the significance of unit and integration testing in guaranteeing the effectiveness, dependability, and usefulness of the programme in diverse settings. The performance evaluation clearly shows that the system attained a high level of accuracy in identifying bullying language. Nevertheless, the efficiency of the system hinges on the adequate training and education of both the staff and users.

The study proposed a thorough strategy for detecting instances of cyberbullying on social media sites, utilising machine learning techniques such as LSVM and RNNs. The suggested system attempted to combat the widespread occurrence of cyberbullying by providing a user-friendly and effective solution for identifying and preventing such behaviour, with a particular emphasis on Facebook and Twitter.

5. Conclusion

The study presented a comprehensive strategy to address cyberbullying on social media platforms, specifically targeting Facebook and Twitter. The software model being suggested utilised machine learning algorithms, specifically Linear Support Vector Machines (LSVM) and Recurrent Neural Networks (RNNs), to identify and forecast instances of cyberbullying in chat threads. The research highlights the necessity of adopting a comprehensive strategy that integrates technological advancements with human engagement in order to effectively address the always changing nature of cyberbullying.

A. Contribution To Knowledge

The study titled " Implementing a Scheme that Uses Recurrent Neural Networks (RNNS) and Linear Support Vector Machines (LSVM) to Uncovers Online Bullying on Social Media Platforms" adds significantly to current knowledge on cyberbullying detection and prevention. The study tackles the growing concern about cyberbullying, especially among children and teenagers, and provides an intelligent framework that uses LSVM and RNNs to recognise and predict cyberbullying texts on social media platforms such as Facebook and Twitter.

The study's evaluation of how digital technology affects social relationships and communication, especially among younger people, is one of its major achievements. It highlights the potential dangers associated with online platforms, including mental health issues, recruiting for illicit activities, and cyberbullying. Acknowledging these issues, the study highlights how crucial it is to implement practical solutions to preserve people's online safety, particularly for vulnerable populations like kids and teenagers.

The study also highlights how text analytics and text mining may be used to better understand the dynamics of cyberbullying. The proposed system builds models that can estimate results on current data based on past interactions using machine learning, languages, and statistics techniques. This approach makes an important contribution to the area by utilising modern algorithms and machine learning models to detect small variations of cyberbullying.

The report also clarifies the grave long-term effects of cyberbullying, such as depression, anxiety, and mental health issues. It highlights how crucial thorough cyberbullying detection techniques are to mitigating the harm done to victims. The proposed approach focuses on Facebook and Twitter, the two most popular social media platforms where cyberbullying happens, in an attempt to close the gap in current models.

The paper acknowledges the challenges associated with acquiring sufficiently labeled information to train algorithms for detecting cyberbullying. It highlights the need for strong training datasets to increase detection efficacy and recognizes the limitations of current models' cross-domain generalization capabilities. This differentiation contributes to the ongoing efforts to raise the precision and potency of cyberbullying detection systems.

Another noteworthy contribution of this research is the creation This research has also produced an intelligent software model that can identify and anticipate cyberbullying texts in real time, which is another important addition. The model's integration of LSVMs and RNNs enhances detection capabilities and enables the automatic identification of cyberbullying occurrences. The study provides information on how the model is implemented, including how many web application languages and the Object-Oriented Analysis and Design Methodology (OOADM) are used to create a user-friendly interface.

The report also highlights the importance of a holistic strategy that combines human involvement with technology breakthroughs. It emphasizes how critical it is to foster empathy and digital literacy in order to encourage acceptable online behavior. By addressing the ever-changing nature of cyberbullying, the study helps to build comprehensive methods that address the root causes and provide support to both victims and perpetrators.

All things considered, this work significantly advances the field of identifying and preventing cyberbullying. It highlights the importance of a complete strategy, provides novel insights into the dynamics of online bullying, and suggests an intelligent framework based on LSVM and RNNs. The study adds to the body of knowledge already in existence and has applications for parents, law enforcement agencies, social media companies, and educational institutions that want to make the internet a safer place for all users.

References

[1]. Ali, R. T., & Kurdy, M. B. (2022). Cyberbullying Detection in Syrian Slang on social media by using Data Mining. International Journal of Engineering (IJERT), 11(4).

[2]. Aroyehun, S. T., & Gelbukh, A. (2018, August). Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labelling. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018) (pp. 90-97).

[3]. Cuncic, A. (2022, February 19). The Psychology of Cyberbullying.

[4]. El Asam, A., Samara, M., & Terry, P. (2019). Problematic internet use and mental health among British children and adolescents. Addictive behaviours, 90, 428-436.

[5]. Emmery, C., Verhoeven, B., De Pauw, G., Jacobs, G., Van Hee, C., Lefever, E., ... & Daelemans, W. (2021). Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. Language Resources and Evaluation, 55(3), 597-633.

[6]. Gottschalk, F. (2019). Impacts of technology use on children: Exploring literature on the brain, cognition and well-being: OECD Education Working Paper No. 195. Paris, Directorate for Education and Skills.

[7]. Ianni, M., Masciari, E., & Sperlí, G. (2021). A survey of Big Data dimensions vs Social Networks analysis. Journal of Intelligent Information Systems, 57(1), 73-100.

[8]. Lane, J., & Stuart, F. (2022). How social media use mitigates urban violence: Communication visibility and third-party intervention processes in digital urban contexts. Qualitative Sociology, 1-19.

[9]. Price-Mitchell, M. (2020, October 02). Teens Discuss Disadvantages of Social Networking.

[10]. Polak, S., & Trottier, D. (2020). Violence and trolling on social media: History, Affect, and Effects of Online Vitriol (p. 227). Amsterdam University Press.

[11]. Unicef. (2020). Cyberbullying: What is it and how to stop it. Retrieved from unicef.