Table of Contents

1.0 Introduction

This research proposal aims to address the pressing issue of cyberbullying within the digital realm by employing machine learning techniques in detection of cyber bullying a case study of X (formerly Twitter). Cyberbullying is a form of online emotional aggression that emerged in the 1990s and has proliferated with personal computers and smartphones. It is imperative to address its gravity, given that an alarming 46% of U.S. teenagers encounter cyberbullying, predominantly through rumors and mean comments(Pew Research Center, 2022).
The research project proposal will comprehensively address the historical evolution of cyberbullying, emphasizing its consequences while highlighting key findings and trends. This will guide the formulation of clear and specific research objectives and align them with the research question and hypothesis, supported by examples from various studies. The proposal will delve into literature review on machine learning techniques and methodologies for data preprocessing, along with a comparative analysis of specific algorithms applicable to cyber bullying detection and classification, data collection techniques and ethical considerations in accessing data from online platforms. The report will also emphasize effective risk management, legal compliance, and professional standards consideration in the context of cyberbullying detection. The ultimate goal of my proposed research is to develop an effective machine-learning model that can accurately detect and classify instances of cyberbullying on X(formerly Twitter), thereby contributing to safer online environments.

2.0 Project Description

2.2 Context of the Problem

According to K. Mahesh et al. (2021), cyberbullying is a growing problem, especially on platforms like Twitter. It involves using electronic technologies to deliberately harass or threaten individuals or groups by sending or posting cruel text and/or graphics. This form of aggression can take various forms, including direct cyberbullying, relational bullying, and the creation of hate groups. Social media platforms present several challenges, including anonymity, legal issues, and the difficulty of filtering offensive content. The dynamic nature of online interactions on Twitter has further contributed to the evolution of cyberbullying, with tactics like hashtag targeting and the rapid spread of damaging misinformation. Despite Twitter's efforts to combat cyberbullying, the problem persists

According to Zhu et al. (2021), an effective cyberbullying detection and classification system can have a significant impact on individuals, online communities, and Twitter as a platform. It can help prevent and reduce the harm caused by cyberbullying in terms of mental health and emotional well-being. An adequate system can also promote a safer and more supportive online environment for all users, enhancing the platform's reputation and attracting more users.

By implementing this system, users can feel safer and more confident expressing themselves online, knowing they are protected from cyberbullying. Twitter as a platform can also benefit from implementing such a system, as it can help enhance its reputation and attract more users.

Zhu et al. (2021) suggest that effective prevention strategies like cyberbullying detection and classification systems should be based on a user-centred and participatory approach that considers user preferences and feedback to ensure effectiveness and relevance

2.2 Project Impact and Beneficiaries

According to Zhu et al. (2021), an effective cyberbullying detection and classification system can have a significant impact on individuals, online communities, and Twitter as a platform. It can help prevent and reduce the harm caused by cyberbullying in terms of mental health and emotional well-being. An adequate system can also promote a safer and more supportive online environment for all users, enhancing the platform's reputation and attracting more users.

By implementing this system, users can feel safer and more confident expressing themselves online, knowing they are protected from cyberbullying. Twitter as a platform can also benefit from implementing such a system, as it can help enhance its reputation and attract more users.

Zhu et al. (2021) suggest that effective prevention strategies like cyberbullying detection and classification systems should be based on a user-centred and participatory approach that considers user preferences and feedback to ensure effectiveness and relevance

2.3 Meeting Requirements and Addressing Specific Problems

To effectively address the specific challenges of cyberbullying on Twitter, it is imperative to align the project with the specific needs and requirements of the client or organization. Emmery et al. (2020) offer valuable insights into the theoretical framing of cyberbullying and the automatic detection of such events. The research considers various aspects of cyberbullying on Twitter, including real-time interactions, diverse user behaviour, and platform-specific features like hashtags and mentions. Unique categories of cyberbullying, such as flaming, outing, and harassment. The research will also consider the risks associated with social media interactions and cyberbullying, which provides insight into the adverse outcomes of such interactions.

2.4 Non Machine Learning and Machine Learning Detection Review

The tangible, practical deliverables of this research project are the identification and evaluation of risk factors and protective factors of cyberbullying, the development of robust machine learning model that classifieds cyber bullying tweets, the ethical and Legal constraint involved in collection of such public data and inference on effective machine learning techniques to employ for such a use case . The intangible practical deliverables include increased awareness and understanding of cyberbullying, its impact on individuals, and the role of social media platforms like Twitter in facilitating or preventing cyberbullying (Zhu et al., 2021).

Social media platforms like Twitter can adopt the strategies and policies outlined in the research to ensure the safety and well-being of users. This aligns with Twitter's mission to "give everyone the power to create and share ideas and information instantly, without barriers" in a safe environment that "promotes healthy conversations."

3.0 Preliminary Literature Review

3.1 Cyberbullying Trend

Cyberbullying, a pervasive issue in the digital age, has driven extensive research into effective detection and prevention mechanisms. Machine learning techniques have emerged as a promising avenue to address this challenge. This preliminary literature review explores critical findings, existing machine learning applications for text analysis and classification, ethical concerns regarding data extraction from social media, and identifies ongoing research gaps to lay the foundation for my proposed project on cyberbullying detection and classification on platform X using machine learning.

Low and Espelage's in 2013 highlights the prevalence of cyberbullying among 11-19-year-olds, revealing that 10-33% become victims, with a higher male involvement. The consequences of cyberbullying span across emotional, academic, and social domains. Ongoing research is imperative to comprehend the evolving nature and scale of this issue in our rapidly changing digital landscape, necessitating adaptable methodologies to keep pace with technological advancements.

The consequences of cyberbullying are profound, affecting both victims and online communities. Notar et al. report that victims often experience depression, anxiety, low self-esteem, and social isolation. The persistent and inconspicuous nature of online harassment amplifies the psychological harm, hindering timely intervention and support. Additionally, cyberbullying is associated with various physical consequences, including substance use, loneliness, and even suicidal ideation. These findings underscore the urgency of developing robust mechanisms for cyberbullying detection and classification.

3.2 Non Machine Learning and Machine Learning Detection Review

The research paper "Development of Computational Linguistic Resources for Automated Detection of Textual Cyberbullying Threats in Roman Urdu Language" in 2021 by Mahesh et al provides an in-depth analysis of cyberbullying detection in the context of the Roman Urdu language. In the literature review segment, the authors delve into various methodologies employed in previous studies concerning cyberbullying detection and classification, extending beyond machine learning.

One of the methodologies explored is the keyword-based approach, which involves identifying specific words and phrases associated with cyberbullying. These identified terms are then utilised to categorise instances of cyberbullying. Another method discussed is the rule-based approach, which entails the development of heuristics to recognize patterns within cyberbullying messages. Such practices include explicit or implied threats, as well as messages that promote self-harm or suicide. Sentiment analysis is another technique reviewed, where the emotional content of a text is analysed to flag messages with a negative tone for further examination. Lastly, the authors investigate graph-based algorithms that employ graph theory to model the interactions and relationships between users on social media platforms.

While the non-machine learning-based methodologies do possess inherent limitations, they remain valuable tools for identifying specific forms of cyberbullying. Additionally, the fusion of these techniques with machine learning strategies holds significant promise for enhancing the accuracy of cyberbullying detection. As such, it is imperative to sustain research efforts in exploring and refining these methodologies, ultimately leading to more precise cyberbullying detection and classification.

The research paper titled "Development of Computational Linguistic Resources for Automated Detection of Textual Cyberbullying Threats in Roman Urdu Language" offers a comprehensive examination of cyberbullying detection within the realm of Roman Urdu. In its literature review section, the authors delve into various methodologies utilized in prior studies on cyberbullying detection and classification, transcending the scope of machine learning.

One of the techniques under scrutiny is the keyword-based approach, which entails the identification of specific words and phrases associated with cyberbullying, subsequently applied to categorize instances of cyberbullying. Another method explored is the rule-based approach, involving the formulation of heuristics to discern patterns within cyberbullying messages. These patterns may encompass explicit or implied threats and statements that advocate self-harm or suicide. Sentiment analysis is also assessed as a technique that examines the emotional content of the text, flagging messages with a negative tone for further review. Additionally, the authors investigate graph-based algorithms that harness graph theory to model interactions and relationships among social media platform users.
While non-machine learning methodologies exhibit inherent limitations, they are valuable tools for identifying specific cyberbullying manifestations. Furthermore, integrating these methodologies with machine learning strategies shows significant potential in augmenting the precision of cyberbullying detection. Hence, it is imperative to sustain research endeavors aimed at exploring and refining these methodologies, ultimately advancing the accuracy of cyberbullying detection and classification.

Machine learning has become a powerful tool in the battle against cyberbullying (Van Hee et al., 2018; Mahesh et al., 2021). This review explores the use of machine learning in previous studies, examining its strengths and limitations and drawing insights from two significant research papers.

Cynthia Van Hee and her research team conducted a pivotal study titled "Automatic detection of cyberbullying in social media text" (Van Hee et al., 2018). Their objective was to automate the detection of cyberbullying in social media text. They introduced an innovative approach involving the analysis of posts from bullies, victims, and bystanders involved in online bullying incidents.

In parallel, K. Mahesh and his collaborators presented a research paper titled "Cyber Bullying Detection on Social Media using Machine Learning" (Mahesh et al., 2021). This study focused on cyberbullying detection on Twitter. They harnessed supervised binary classification machine learning algorithms, utilising data from Twitter hate speech and personal attacks in Wikipedia forums. Their model achieved impressive accuracy levels, exceeding 90% for tweet data and surpassing 80% for Wikipedia data.

3.3 Ongoing Research gaps

Despite their promising results, both approaches have limitations (Van Hee et al., 2018; Mahesh et al., 2021). Van Hee's study grappled with the challenge of discerning the intent behind cyberbullying in online environments where cues like intonation and facial expressions are absent, making online conversations more ambiguous. The power balance between bully and victim, a key bullying criterion, is challenging to conceptualise and measure online. Moreover, the permanence of defamatory information once it's made public on the internet is a persistent issue.

The preliminary literature review identifies critical limitations in existing cyberbullying detection and classification research, emphasizing the need for further investigation. It underscores the growing importance of machine learning techniques in addressing cyberbullying and suggests the potential for innovation by integrating non-machine learning approaches. Ethical considerations in data extraction are acknowledged, positioning my research within an ethical context. The review highlights the necessity of adapting to the ever-evolving digital landscape, making my research forward-looking as I intend to embed streaming data from X. This literature review sets the stage for my research by emphasizing its significance, demonstrating the potential for improvement in cyberbullying detection and classification employing other machine learning techniques.

4.0 Research, Questions, Design and Methodology

4.1 Research Question And Hypothesis

The overarching research aim of this project, "Cyberbullying Detection and Classification with Machine Learning," is to develop a robust machine learning model capable of accurately detecting and classifying instances of cyberbullying on X (formerly Twitter). This research aims to address several specific objectives and answer relative research questions:

Machine Learning Techniques for Detection By Comparing (Objective 1): What machine learning techniques and algorithms are most suitable for accurately detecting and classifying instances of cyberbullying on X? (Research Question 1).

Ethical Integration (Objective 2): How can ethical considerations be effectively integrated into the development and deployment of a cyberbullying detection model on a social media platform like X? (Research Question 2).

Legal and Regulatory Compliance (Objective 3): What legal and regulatory aspects must be addressed when implementing a cyberbullying detection system on a platform like X, and how can compliance be ensured? (Research Question 3).

Professional Responsibilities (Objective 4): What are the professional responsibilities and considerations when developing and deploying a cyberbullying detection model for a public platform, particularly in terms of user privacy and freedom of expression? (Research Question 4).

Challenges and Opportunities with Twitter API (Objective 5): What are the challenges and opportunities associated with utilizing X Twitter API for streaming real-time Twitter data for cyberbullying detection, and how can these be effectively managed? (Research Question 5).

Given the exploratory nature of the proposed research, there is no specific hypothesis to prove or disprove. The research aims to identify suitable machine learning techniques and algorithms for cyberbullying detection and explore the ethical, legal, and professional considerations associated with the development and deployment of a detection model on X. The research further aims to analyze the challenges and opportunities related to utilizing the X Twitter API for data streaming.

4.2 Research Design & Justification

The study adopts a mixed-method research design that aligns with the project’s objectives. The research design consists of two methods: quantitative and qualitative. The quantitative method involves data collection using Twitter’s public API to obtain large-scale tweet datasets for statistical analysis. This method enables the quantitative evaluation of cyberbullying phenomena and dynamics on Twitter while also considering the ethical, legal factors in collection of the data by obtaining ethical approval and informed consent from twitter this Ensures data confidentiality and comply with relevant law for using such a sensitive data.
The qualitative method involves manual content analysis of select tweets to understand the context of cyberbullying occurrences and .
Additionally, a comparative analysis will be performed, contrasting the results and methods of this study with existing research on cyberbullying detection approaches across different social media platforms. This comparative analysis will ensure that the research design is rigorous and informed by the wider scope of cyberbullying detection methods. This design is justified as it facilitates a holistic understanding of the problem, aligning with the project’s aim to develop an effective cyberbullying detection and classification system. Furthermore Random user survey will be carried out on twitter while requesting permission and giving clear context of the use case to validate the machine learning Inferences.

4.3 Methodology

The research will employ a mixed-method approach, incorporating both quantitative and qualitative techniques. Data will be collected using Twitter's public API, with a focus on tweets containing potential cyberbullying content. Preprocessing and feature extraction will be performed to prepare the data for analysis and understanding context as already established and utilized to be efficient. Quantitative analysis will identify trends and patterns in cyberbullying, utilizing 3 different machine learning models such a Logistic regression Bidirection fast gated Recurrent Unit as proven to be efficient for emotion classification classification (Chen 2019), to assess classification performance. Qualitative analysis will involve manual content review for nuanced insights. A comparative analysis will be conducted by reviewing existing research on cyberbullying detection across social media platforms. Ethical and legal considerations will guide data handling, and models will be refined iteratively.

Challenges and opportunities associated with Twitter's API usage will be managed, and user privacy and freedom of expression will be respected throughout the study. The research will be documented transparently, fostering accountability and ethical research practices.

5.0 Resources and Constraint

5.1 Hardware and Software Requirements

The execution of this research project necessitates the utilization of a high-performance computing system equipped with a dedicated GPU. This setup is crucial to facilitate efficient training of machine learning models. The hardware resources will be procured and managed using cloud-based computing services. In terms of software, specialized libraries such as TensorFlow and various NLP toolkits will be employed. These resources form the backbone for the successful implementation and evaluation of the cyberbullying detection model.

5.2 Data Acquisition and Management

A comprehensive and annotated dataset comprising instances of cyberbullying will be acquired from the Twitter platform via the Twitter developer API. This process will strictly adhere to ethical guidelines. Rigorous data preprocessing steps, including cleaning, feature extraction, and validation, will be conducted to maintain the integrity of the dataset. These steps are vital in guaranteeing the accuracy and effectiveness of the machine learning model(Chen 2019).

5.3 Ethical Considerations and Privacy Measures

To uphold ethical standards, all data utilized in this research will be anonymized and aggregated to prevent any potential identification of individuals. Additionally, strict privacy safeguards will be adhered to, and regular consultations with ethics committees will be ensured. These measures guarantee that the research is conducted ethically and responsibly(Floridi 2010).

5.4 Budget and Resource Allocation

Budgetary constraints will be addressed by prioritizing using open-source software and publicly available datasets to minimize costs. Additionally, careful allocation of resources will be made, focusing on cost-effective approaches. This approach ensures that the research remains within budget while maintaining access to the necessary resources for successful execution.

6.0 Social, Ethical, Professional and Legal Consideration

6.1 Social Consideration

When you share your emotions on social media, it's crucial to recognize that systems are constantly trying to interpret your emotional state based on your written words. However, these systems can sometimes misinterpret your feelings, leading to decisions that don't align with your true emotions. Such misinterpretations can result in frustration and, in some cases, even serious harm. Given the widespread use of social media, these misinterpretations can quickly escalate, leading to issues like cyberbullying and breaches of privacy.

This highlights the need to investigate how these systems operate, understand their strengths and vulnerabilities, and work towards improving their accuracy. Furthermore, even when these systems are accurate, they may inadvertently harm individuals in unforeseen ways, necessitating the development of proactive solutions to mitigate such harms.

6.2 Ethical Consideration

When we study how to analyze feelings from words, we face some ethical problems about how to collect and use the data in a fair way. The ACM Code of Ethics and Professional Conduct tells us that computing professionals should respect and protect the privacy rights of people (ACM, 2018). We need to make sure that we hide the identity of the people whose feelings we are analyzing from a lot of data.

This can help us avoid breaking their privacy rights. point out that hiding the identity is important, but not always enough to protect the privacy completely. We need to use other methods, like differential privacy, to make the data more secure. This can also help us respect the rights of the people, build trust with them, and make ethics a key part of our analysis

6.3 Professional Consideration

Professional research should be done with respect and honesty. Researchers need to collect data in a way that is ethical and transparent, and they should tell how they did it. The findings should be tested carefully and reviewed by other experts. Researchers should also use data that matches the population they are studying and avoid any tools that could make the results biased. It is recommended that involving diverse and underrepresented groups and using anonymous evaluations can make research better . It is also important to get funding for less-researched languages and communities and to check the findings on different datasets. These steps can help to make the research reliable, fair, and rigorous

6.4 Legal Consideration

Navigating the legal landscape in the context of cyberbullying and online activities requires careful consideration. Various regulations exist to safeguard individuals and their rights. Notable among these regulations are the European Union General Data Protection Regulation (GDPR) and the US Children's Online Privacy Protection Act (COPPA).

In addition to these privacy-related regulations, it is crucial to be vigilant about other legal aspects, including concerns such as defamation, intellectual property, consumer protection, and advertising endorsements. Different countries may have specific laws and guidelines in place to address these issues. For effective guidance in understanding and complying with these regulations, it is advisable to consult legal experts who can provide tailored advice and detailed information.

7.0 References

Batool, S., Yousaf, R. and Batool, F., (2017). Bullying in Social Media: An Effect Study of Cyber Bullying on the Youth. Pakistan Journal of Criminology, [online] 9(4), pp.119-139

Chen, J.X., Jiang, D.M. and Zhang, Y.N., (2019). A Hierarchical Bidirectional GRU Model With Attention for EEG-Based Emotion Classification. IEEE Access, 7, pp.118530-118540.

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

Floridi, L (ed.) (2010) The Cambridge Handbook of Information and Computer Ethics, Cambridge University Press, Cambridge

Mahesh, K., Gothane, S., Toshniwal, A., Nagarale, V. and Gopu, H., n.d. , 2021 Cyber Bullying Detection on Social Media using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology

Notar, C. E., Padgett, S., & Roden, J. (2013). Cyberbullying: A Review of the Literature. Universal Journal of Educational Research, 1(1), 1-9. https://doi.org/10.13189/ujer.2013.010101

Low, S. and Espelage, D., (2013). Differentiating Cyber Bullying Perpetration From Non-Physical Bullying: Commonalities Across Race, Individual, and Family Predictors. Psychology of Violence, 3,

Research design: Urban, JB, & van, EBM (2017), Designing and Proposing Your Research Project, American Psychological Association, Washington DC.

Pew Research Center. (2022). Teens and Cyberbullying 2022. [online] Available at: https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W. and Hoste, V., (2018). Automatic detection of cyberbullying in social media text. PLoS One, 13(10), p.e0203794. Available at: https://doi.org/10.1371/journal.pone.0203794

Zhu, C., Huang, S., Evans, R., & Zhang, W. (2021). Cyberbullying Among Adolescents and Children: A Comprehensive Review of the Global Situation, Risk Factors, and Preventive Measures. Frontiers in Public Health, 9, 634909. https://doi.org/10.3389/fpubh.2021.634909