A Comprehensive Approach to Cyberbullying Classification and Prediction in Social Networks

Jaya suriya A[1], Chrislyn Easter Dafna S[2], Jyothikaa K P[3], Srinidhi S[4], Dr. S. Sivakumari[5]

[1,2,3,4,5]UG Scholars, [5]Head of the Department & Professor, Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, School of Engineering, Coimbatore, India

Abstract

This project aims to address the critical issue of cyberbullying through a comprehensive classification and prediction framework. Leveraging Support Vector Machines (SVM) with Radial Basis Function (RBF), coupled with an Intention Model, the research establishes a robust methodology for identifying and predicting instances of cyberbullying within social networks, particularly focusing on Facebook. The approach consists of the Group Identification Phase (GIP) and Risk Assessment Phase (RAP). In the GIP, the study identifies potential risk groups within the social network, while in the RAP, it assesses the risk of cyberbullying within these groups. By generating training and testing datasets, the SVM-RBF model achieves an impressive accuracy rate of 93%, indicating its efficacy in cyberbullying classification and prediction. Furthermore, with an accuracy of 81%, the Intention Model offers complementary insights into cyberbullying behaviour. This is a significant contribution to social network risk assessment, practical tools for addressing cyberbullying and enhancing online safety.

Keywords: Support Vector Machine, Radial Basis Function

1. Introduction

In recent years, the widespread adoption of social media platforms has revolutionized the way people communicate and interact online. While these platforms offer tremendous opportunities for connectivity and expression, they have also given rise to a dark side - cyberbullying. Cyberbullying refers to the use of digital communication tools, such as social media, text messages, or online forums, to harass, intimidate, or demean individuals. The prevalence of cyberbullying has become a growing concern as it can have severe and lasting effects on the victims, leading to emotional distress, social isolation, and, in some tragic cases, even suicide. Identifying and combating cyberbullying has, therefore, become a critical priority for creating safe and inclusive online spaces. Traditional manual methods for detecting and preventing cyberbullying are often insufficient to address the sheer volume of content generated on social media platforms. Fortunately, advances in Machine Learning (ML) and Natural Language Processing (NLP) have opened new possibilities for automating cyberbullying detection. By leveraging these technologies, we can develop intelligent systems capable of identifying potentially harmful content in real-time, allowing for timely interventions and fostering a healthier online environment.

2. System Analysis

Existing system:

Information and Communication Technologies have significantly changed the way we connect and communicate with one another through social networks. Unfortunately, this progress has also brought about serious issues like cyberbullying, which can have harmful effects on individuals. Currently, the tools provided for users to combat bullying—such as reporting, blocking, and removing harmful posts—are often manual and not very effective. Moreover, relying on a basic text representation method, like the bag-of-words approach, without any additional context, limits our ability to accurately classify cyberbullying content.

In response to these challenges, this research focused on creating an automatic system to detect cyberbullying using two main strategies: Conventional Machine Learning and Transfer Learning. The study employed the AMICA dataset, which includes a wealth of information on cyberbullying and follows a thorough annotation process. In the Conventional Machine Learning approach, we utilized a variety of features such as textual, sentiment, emotional, static and contextual word embeddings, psycholinguistic traits, term lists, and measures of toxicity. This research marks the first time toxicity features have been integrated into the detection of cyberbullying.

Additionally, it is the first to use the most recent psycholinguistic features from the Linguistic Inquiry and Word Count (LIWC) 2022 tool, as well as vocabulary from the Empath lexicon, to enhance cyberbullying detection efforts. The different contextual embeddings studied—like Gilbert, TN BERT, and DistilBERT—showed similar effectiveness, but we ultimately chose DistilBERT due to its superior F-measure performance.

When we looked at individual features, we found that textual features, DistilBERT embeddings, and toxicity measures stood out, setting new performance benchmarks. By combining these features—textual, sentiment, DistilBERT embeddings, psycholinguistics, and toxicity—we were able to enhance the model's performance, achieving an impressive F-measure of 64.8% with the Logistic Regression model. This combination not only performed better than the Linear SVC model but also trained faster and managed high-dimensional features more efficiently.

Proposed system:

The proposed system aims to tackle the serious issue of cyberbullying on social networks, especially focusing on Facebook. It provides an organized framework designed to effectively assess and predict instances of cyberbullying.

The process starts by gathering data from Facebook, which gives us a wealth of user interactions and content to analyze. We use a two-step risk assessment approach. In the first step, called the Group Identification Phase (GIP), we look for groups within the network that may be at risk. After identifying these groups, we move to the second step, the Risk Assessment Phase (RAP), where we evaluate how likely cyberbullying is to occur within them.

To ensure we make accurate predictions and classifications, the system creates training and testing datasets. We then use these datasets in two different prediction models. The first model uses Support Vector Machines (SVM) with a Radial Basis Function (RBF). This method is effective at identifying complex patterns in data, which helps classify and predict potential instances of cyberbullying. Moreover, we also incorporate an Intention Model into the system, which considers psychological and behavioral factors, helping us enhance the detection and prediction of cyberbullying even further.

3. Techniques

Natural Language Processing:

In the era of modern Natural Language Processing (NLP), the development of powerful language models has revolutionized the way machines understand and process human language. Among these ground-breaking models, Distil BERT has emerged as a prominent contender, offering remarkable efficiency and performance in various NLP tasks. Distil BERT is a distilled version of the revolutionary BERT (Bidirectional Encoder Representations from Transformers) model, which was introduced by Google in 2018. BERT's ability to learn context and meaning from both left and right contexts of a word was a significant advancement in NLP. However, its sheer size made it computationally expensive and challenging to deploy in resource-constrained environments. In response to these challenges, Hugging Face, a leading NLP research organization, introduced Distil BERT in 2019. Distil BERT adopts a novel approach called "distillation," which involves compressing the original BERT model while preserving much of its language understanding capabilities. As a result, Distil BERT is significantly smaller and faster, making it more practical for real-world applications without compromising performance.

Machine learning:

Machine Learning (ML) is a cutting-edge field of artificial intelligence that empowers computers to learn from data and improve their performance over time without being explicitly programmed. By mimicking the way humans learn, ML algorithms enable machines to recognize patterns, make decisions, and solve complex problems across various domains. Supervised learning involves training the model on labeled data, where input-output pairs are provided. The model learns to map inputs to corresponding outputs and can then predict outputs for new inputs it has not seen before. This technique is widely used for tasks such as image recognition, natural language processing, and spam detection. Unsupervised learning, on the other hand, deals with unlabelled data, seeking to find patterns, groupings, or hidden structures within the data without explicit guidance. This is particularly useful for tasks like clustering similar data points, dimensionality reduction, and anomaly detection. Reinforcement learning takes inspiration from behavioural psychology, where an agent interacts with an environment and learns to take actions that maximize a cumulative reward signal. This paradigm is employed in scenarios such as training autonomous vehicles, playing complex games, and optimizing resource allocation.

4. Project Description

Our aim is to create a dependable system for detecting cyberbullying on social media platforms using machine learning (ML) techniques. As social media becomes an increasingly integral part of our lives, the issue of cyberbullying has become a serious concern, inflicting emotional pain and distress on its victims. Given the sheer volume of content being generated, it’s simply not feasible to monitor everything manually. This is why an ML solution is essential for automatically spotting instances of cyberbullying.

The objective is to develop a model that can effectively identify posts and comments involving cyberbullying while distinguishing them from non-harmful content. This task presents challenges, such as addressing data bias, minimizing false positives and negatives, and adapting to the evolving language used in online interactions.

Additionally, we must prioritize privacy regulations to ensure users feel comfortable with how their data is handled. Building trust is crucial for the model's successful implementation. Ultimately, our goal is to create a solution that not only detects cyberbullying but also promotes a safer online environment, positively impacting users' well-being.

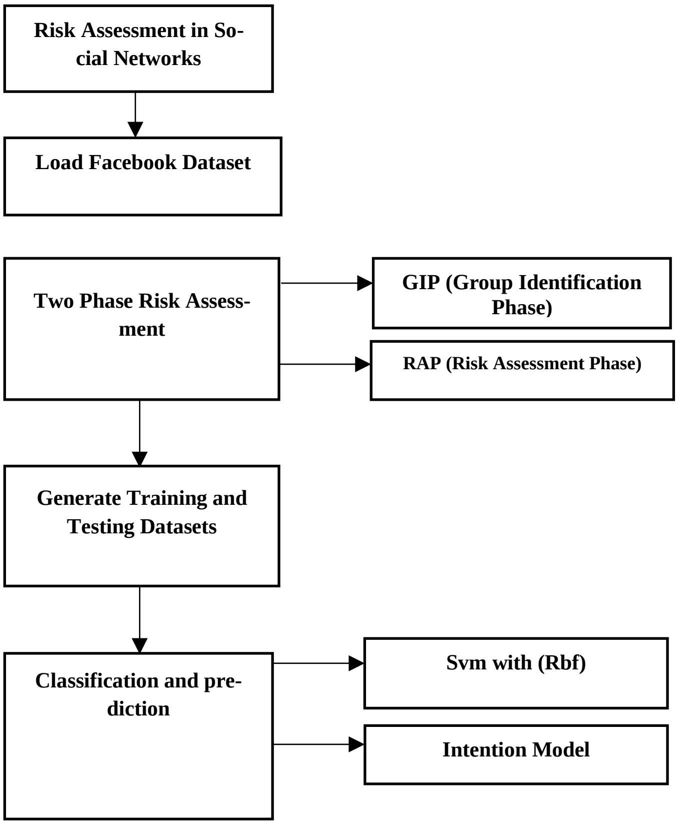
Risk Assessment in Social Networks
Load Facebook Dataset
Two Phase Risk Assessment
GIP (Group Identification Phase)
RAP (Risk Assessment Phase)
Generate Training and Testing Datasets
Classification and prediction
Svm with (Rbf)
Intention Model

Fig 1. Block diagram

A. Risk Assessment in Social Media:

This module is all about understanding the risks of cyberbullying in social networks. It looks into the various ways cyberbullying can happen, identifies groups of people who might be more vulnerable, and examines how these experiences affect individuals and communities. By assessing risks in social networks, we aim to create proactive strategies to prevent cyberbullying and intervene effectively, ultimately fostering a safer and more supportive online environment for everyone.

B. Load Facebook dataset:

This module focuses on loading a dataset extracted from the Facebook platform. The dataset contains relevant information such as for example, user interactions, posts, comments, and other social network activities. Loading the Facebook dataset is essential for conducting comprehensive analysis and modeling of cyberbullying behavior within the platform.

C. Two-phase risk assessment:

The two-phase risk assessment approach involves a structured methodology for evaluating cyberbullying risk within social networks. The first phase, known as the Group Identification Phase (GIP), focuses on identifying potential risk groups based on specific criteria or behavioural patterns. The second phase, the Risk Assessment Phase (RAP), assesses the level of cyberbullying risk within the identified groups through advanced classification and prediction techniques.

First phase - GIP (Group Identification Phase):

GIP is the initial phase of risk assessment, where potential risk groups within the social network are identified and categorized. This phase involves analysing user data to detect patterns of behavior associated with cyberbullying involvement or susceptibility. The primary objective of GIP is to lay the groundwork for targeted risk assessment and intervention efforts in the subsequent phases.

Second phase - RAP (Risk Assessment Phase):

RAP is the second phase of risk assessment, focusing on evaluating the level of cyberbullying risk within the identified groups. In this phase, advanced classification and prediction models are applied to the data to assess the likelihood of cyberbullying occurrences. RAP aims to provide actionable insights for implementing preventive measures and supporting individuals at risk of cyberbullying.

D. Generate Training and Testing Datasets:

This module involves generating datasets for training and testing machine learning models used in the risk assessment process. The training dataset comprises a subset of the data used to train the models, while the testing dataset is used to evaluate the models' performance and generalization capabilities. Generating high-quality training and testing datasets is critical for developing accurate and reliable risk assessment models.

E. SVM with Radial Basis Function based classification and prediction:

This module employs Support Vector Machines (SVM) with Radial Basis Function (RBF) kernels for cyberbullying classification and prediction. SVM with RBF kernels is a machine learning algorithm known for its effectiveness in handling nonlinear classification tasks. It is applied to the data to classify instances of cyberbullying and predict future occurrences within the identified risk groups.

F. Intention Model based classification and prediction:

The Intention Model is utilized for cyberbullying classification and prediction, providing insights into the underlying intentions and motivations behind cyberbullying behavior. This model focuses on understanding the psychological and social factors driving cyberbullying incidents, enhancing the accuracy and depth of cyberbullying risk assessment. The Intention Model complements other classification techniques to provide a comprehensive understanding of cyberbullying dynamics within social networks.

5. Result Analysis

The analysis of our cyberbullying detection framework has produced encouraging results for two different models: the Support Vector Machine with a Radial Basis Function (SVM-RBF) and the Intention Model. The SVM-RBF algorithm shines, achieving an impressive accuracy rate of 93.57%. This means it does an excellent job of identifying and classifying instances of cyberbullying on social networks. Its precision, recall, and F-measure scores are also noteworthy, standing at 93.35%, 93.85%, and 93.60%, respectively. These figures highlight the model's effectiveness in recognising real cases of cyberbullying while keeping false positives and negatives to a minimum. Essentially, it strikes a valuable balance between precision (how many of the predicted cases are actually correct) and recall (how many actual cases the model correctly identifies). Given these strong performance metrics, the SVM-RBF is a reliable tool in tackling online harassment.

On the other hand, the Intention Model performs slightly less effectively, with an accuracy of 81.69%. Its precision, recall, and F-measure scores are 76.64%, 75.05%, and 75.84%, respectively. Although these numbers are lower than those of the SVM-RBF, the Intention Model provides important insights into the motivations and behaviors behind cyberbullying. By examining the intentions behind online interactions, this model brings additional depth to our understanding of this complex issue, helping to inform better strategies for enhancing online safety.

Together, these models highlight our framework's effectiveness in identifying, understanding, and addressing cyberbullying. This multifaceted approach can help foster safer online environments for everyone.

Table 1. Comparison Table

Algorithm	Accuracy	precision	Recall	F-measure
SVM With RF	93.57	93.35	93.85	93.6
Intention Model	81.69	76.64	75.05	75.84

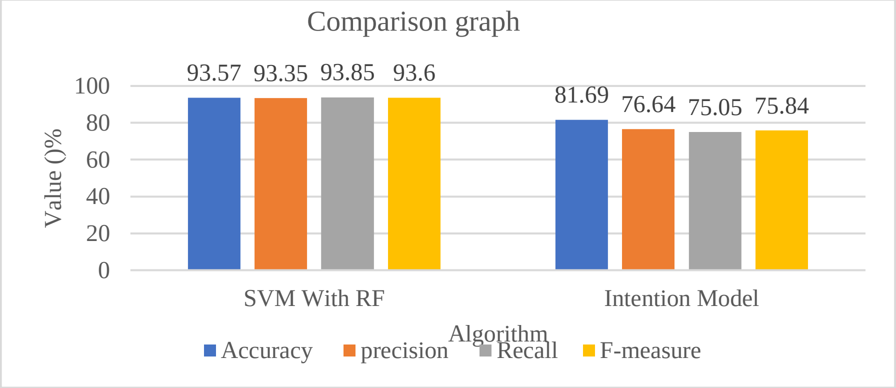

Fig 2. Comparison graph

6. Conclusion and future work

The proposed framework takes a thoughtful and comprehensive approach to tackling the widespread issue of cyberbullying on social networks, especially focusing on Facebook, a platform many people use daily. As online interactions have grown, so too have incidents of cyberbullying, making it crucial to craft effective strategies for both detecting and responding to this troubling behavior. This framework seeks to meet these challenges by utilizing advanced machine learning techniques, like Support Vector Machines with a Radial Basis Function, along with an innovative Intention Model. By harnessing these sophisticated tools, the system offers a reliable way to detect, assess, and predict instances of cyberbullying quickly and effectively.

At its core, the framework identifies potential risk groups by examining user behavior, interactions, and various context factors that may lead to cyberbullying. By accurately evaluating the chances of these incidents occurring, this proactive approach enables timely interventions that can lessen the impact of online harassment. As a result, the framework significantly enhances online safety, fostering a healthier digital environment where everyone can engage without fear of bullying.

Looking forward, there's a wonderful opportunity to broaden this framework beyond just Facebook to include a variety of other social media platforms. This expansion would provide a deeper understanding of cyberbullying dynamics across different online spaces, recognizing that each platform has its unique characteristics that can affect how and why bullying happens. By integrating data and insights from multiple platforms, the framework can become even more effective at addressing this important issue.

Additionally, real-time monitoring capabilities are essential for keeping pace with the ever-evolving nature of cyberbullying behaviors. This feature would allow for the immediate detection of incidents, ensuring that users receive the support they need when they need it most. Furthermore, by incorporating natural language processing techniques, the system's ability to accurately interpret the subtleties of online interactions would improve, helping to spot those more nuanced forms of cyberbullying that might go undetected by traditional methods.

We also see the real value in engaging users in the fight against cyberbullying. By exploring ways to integrate user feedback mechanisms and community-driven moderation strategies, we can empower individuals to take an active role in recognizing and addressing instances of bullying. Creating platforms where users can report incidents and share their experiences fosters a sense of collaboration in tackling online harassment. This community involvement raises awareness and builds a culture of accountability, encouraging users to support one another.

In summary, the proposed framework marks an important step forward in the battle against cyberbullying on social networks. By combining machine learning techniques, real-time monitoring, and community engagement, we can work together to create a safer and more positive online environment for everyone. As this framework evolves and expands, it has the potential to genuinely impact the lives of those affected by cyberbullying, ultimately leading to a healthier and more respectful digital landscape for all.

References

1. B. Cagirkan and G. Bilek, ‘‘Cyberbullying among Turkish high school students,’’ Scandin. J. Psychol., vol. 62, no. 4, pp. 608–616, Aug. 2021, doi: 10.1111/sjop.12720
2. P. T. L. Chi, V. T. H. Lan, N. H. Ngan, and N. T. Linh, ‘‘Online time, experience of cyber bullying and practices to cope with it among high school students in Hanoi,’’ Health Psychol. Open, vol. 7, no. 1, Jan. 2020, Art. no. 205510292093574, doi: 10.1177/2055102920935747
3. A. López-Martínez, J. A. García-Díaz, R. Valencia-García, and A. Ruiz-Martínez, ‘‘CyberDect. A novel approach for cyberbullying detection on Twitter,’’ in Proc. Int. Conf. Technol. Innov., Guayaquil, Ecuador: Springer, 2019, pp. 109–121, doi: 10.1007/978-3-030-34989- 9_9.
4. R. M. Kowalski and S. P. Limber, ‘‘Psychological, physical, and academic correlates of cyberbullying and traditional bullying,’’ J. Adolescent Health, vol. 53, no. 1, pp. S13–S20, Jul. 2020, doi: 10.1016/j.jadohealth.2012.09.018
5. Y.-C. Huang, ‘‘Comparison and contrast of piaget and Vygotsky’s theories,’’ in Proc. Adv. Social Sci., Educ. Humanities Res., 2021, pp. 28–32, doi: 10.2991/assehr.k.210519.007
6. A. Anwar, D. M. H. Kee, and A. Ahmed, ‘‘Workplace cyberbullying and interpersonal deviance: Understanding the mediating effect of silence and emotional exhaustion,’’ Cyberpsychol., Behav., Social Netw., vol. 23, no. 5, pp. 290–296, May 2020, doi: 10.1089/cyber.2019.0407.
7. D. M. H. Kee, M. A. L. Al-Anesi, and S. A. L. Al-Anesi, ‘‘Cyberbullying on social media under the influence of COVID-19,’’ Global Bus. Organizational Excellence, vol. 41, no. 6, pp. 11–22, Sep. 2022, doi: 10.1002/joe.22175
8. I. Kwan, K. Dickson, M. Richardson, W. MacDowall, H. Burchett, C. Stansfield, G. Brunton, K. Sutcliffe, and J. Thomas, ‘‘Cyberbullying and children and young people’s mental health: A systematic map of systematic reviews,’’ Cyberpsychol., Behav., Social Netw., vol. 23, no. 2, pp. 72–82, Feb. 2020, doi: 10.1089/cyber.2019.0370.
9. R. Garett, L. R. Lord, and S. D. Young, ‘‘Associations between social media and cyberbullying: A review of the literature,’’ mHealth, vol. 2, p. 46, Dec. 2016, doi: 10.21037/mhealth.2016.12.01
10. M. O. Raza, M. Memon, S. Bhatti, and R. Bux, ‘‘Detecting cyber bullying in social commentary using supervised machine learning,’’ in Proc. Future Inf. Commun. Conf. Cham, Switzerland: Springer, 2020, pp. 621–630