

8TH INTERNATIONAL CONFERENCE
ON DIGITAL INNOVATION

ICDI 2025

NAVIGATING THE NEW FRONTIER OF DISRUPTION
HYBRID CONFERENCE

e-Proceedings of the
**8th International Conference on Digital Innovation –
Navigating the New Frontier of Disruption**

Venue: Nguyen Tat Thanh University, Vietnam

Date: 30th – 31st May 2025

Editors

Vimala Devi A/P Martham Muthu
Zunika Samiyah binti Khalib
Masturah binti Mohd Nor

eISSN 2976-2200



9 772 976 2200 03



**THE 8TH INTERNATIONAL CONFERENCE ON DIGITAL INNOVATION –
NAVIGATING THE NEW FRONTIER OF DISRUPTION**

Organized by

Nguyen Tat Thanh University (NTTU)
Malaysia University of Science and Technology (MUST)

**NTTU, VIETNAM
30th to 31st May 2025**

Published 2025

Published by:

*Malaysia University of Science and Technology (MUST)
Block B, Encorp Strand Garden Office
No. 12, Jalan PJU 5/5, Kota
Damansara 47810 Petaling Jaya,
Selangor.*

© Copyright 2025, ICDI2025-NNFD – International Conference

All rights reserved. *No part of this proceedings may be reproduced, stored in retrieval system, transmitted, in any forms or any means, without the prior permission from Nguyen Tat Thanh University (NTTU) and Malaysia University of Science and Technology (MUST).*



DISCLAIMER:

The 8th International Conference on Digital Innovation – Navigating the New Frontier of Disruption (ICDI2025-NNFD) proceedings is an initial academic exercise promoted to provide opportunities to researchers, academicians, and students to disseminate their research findings.

The conference committee has taken all reasonable care in the preparation and compilation of the information contained in this conference proceedings. The editors of the proceedings hereby confirm that the papers published in this conference proceedings were accepted in accordance with requirements stated by the conference committee.

The editors have good faith in all the authors and presenters who have submitted papers to the ICDI2025-NNFD proceedings in terms of the preservation of the accuracy, validity, and reliability of any information published in the ICDI2025-NNFD proceedings.

The editors wish to emphasize that the committee is NOT responsible for any research misconduct leading to plagiarism, manipulation of data, fabrication of data, republication, etc. in the papers. The conference committee has agreed to accept manuscripts up to a certain acceptable similarity index percentage. Hence, the conference committee herein expressly disclaims all and any liability or responsibility to readers for any loss or damage of any kind incurred as a result of the use of these conference proceedings as your reference.

Preface

Nguyen Tat Thanh University (NTTU) and **Malaysia University of Science and Technology (MUST)** are proud to co-organize the 8th International Conference on Digital Innovation – Navigating the New Frontier of Disruption (ICDI2025-NNFD), held at NTTU, Ho Chi Minh City, Vietnam.

As a prestigious annual event, the International Conference on Digital Innovation continues to serve as a global platform for academics, researchers, practitioners, and industry leaders to share knowledge and engage in meaningful discourse surrounding emerging trends in digital innovation. The 2025 edition, themed “Navigating the New Frontier of Disruption,” highlights the pressing need to explore the multifaceted nature of disruption across industries, particularly in finance and accounting, where traditional models are being redefined by technological advancement and evolving consumer expectations.

Conference Objectives

The main objectives of the ICDI2025-NNFD are:

- To explore the role of digital transformation and emerging technologies in driving disruption across industries – The conference aims to examine how digital advancements serve as catalysts for innovation, reshaping business models, market dynamics, and consumer behaviours.
- To foster collaboration and knowledge-sharing for sustainable innovation – The conference provides a platform for scholars, industry leaders, and policymakers to exchange insights on managing disruption through interorganizational collaboration, change management, and future workforce development.

This conference is jointly organized by Malaysia University of Science and Technology (MUST), Malaysia, and Nguyen Tat Thanh University (NTTU), Vietnam, with a shared vision of fostering educational excellence and professional development.

The conference invites submissions in the following thematic clusters:

- Environment, Sustainability, and Social
- Finance, Accounting, and Governance
- Management, Marketing, Business, and Economic
- Supply Chain, Transportation, and Logistics
- Artificial Intelligence (AI) and Digital Transformation

ICDI2025-NNFD aspires to strengthen academic-industry cooperation and cultivate innovation that responds to contemporary disruptions in the global business and technological environment.

Dr. Nguyen Mai Lan
Honourable Chairwoman
Nguyen Tat Thanh
University

Nguyen Tat Thanh University (NTTU) is honored to co-host the 8th International Conference on Digital Innovation (ICDI2025), a platform that continues to strengthen global collaboration in research, education, and applied innovation. As a dynamic institution committed to cultivating forward-thinking professionals, NTTU places strong emphasis on nurturing entrepreneurial spirit, cross-border competence, and research excellence aligned with the demands of Industry 4.0.



In today's ever-shifting global landscape, we are acutely aware of the disruptions affecting international economies, education systems, and organizational structures. From digital finance and AI integration to evolving market dynamics and regulatory frameworks, disruption is no longer an occasional occurrence, it is the new norm. At the same time, it offers profound opportunities to innovate, adapt, and lead with impact.

This year's conference theme, *Navigating the New Frontier of Disruption*, resonates deeply with our mission. It urges us to explore how digital transformation can serve not only as a response to challenges but as a proactive tool for strategic renewal across industries. Whether in sustainability, governance, or business operations, embracing disruptive change with purpose and foresight will be essential to thriving in a world marked by complexity and uncertainty.

We believe that this gathering of scholars, researchers, industry leaders, and students will create an intellectual synergy that pushes boundaries and generates actionable insights. Through thoughtful discourse and knowledge-sharing, we hope to address pressing regional and global issues ranging from technological readiness to policy development and foster academic and professional resilience.

On behalf of Nguyen Tat Thanh University, I would like to express my sincere gratitude to all the partners, contributors, and participants who have made this conference possible. Your presence and engagement affirm the importance of collaboration in paving the way for a smarter, more agile future.

May this conference inspire meaningful connections and spark ideas that will continue to grow long after the final session.

Professor Dr. Premkumar Rajagopal

Program Chair

Malaysia University of Science and Technology

Malaysia University of Science and Technology (MUST), in collaboration with Nguyen Tat Thanh University (NTTU), is honoured to convene the 8th International Conference on Digital Innovation (ICDI2025), scheduled to take place in Ho Chi Minh City, Vietnam, on the 30th and 31st of May 2025. This annual event continues to serve as a distinguished forum for the exchange of ideas, research findings, and applied knowledge, bringing together scholars and industry leaders in a spirit of rigorous academic inquiry and interdisciplinary collaboration.



In the current global context, characterised by volatility, complexity, and accelerated technological change, the notion of disruption has transcended its traditional connotation. It now signifies profound, structural transformations in how industries function, adapt, and innovate. Anchored in the theme Navigating the New Frontier of Disruption, ICDI2025-NNFD provides a platform for critical engagement with the ways in which digital innovation is reshaping core domains including finance, accounting, logistics, marketing, and governance. These disruptions, while presenting multifaceted challenges, also offer unprecedented opportunities for institutional reform, strategic renewal, and societal advancement.

This conference foregrounds not only technological evolution but also the importance of fostering cooperative, cross-disciplinary networks capable of responding adaptively to change. Through robust academic dialogue and international partnership, we aim to collectively interpret the dynamics of disruption and contribute to constructive, forward-looking solutions.

We extend sincere appreciation to all contributors, authors, reviewers, keynote speakers, session chairs, and members of the organising committee, for their scholarly commitment and invaluable efforts in making ICDI2025-NNFD possible. Their dedication ensures the academic integrity and continuity of this significant intellectual endeavour.

As participants engage with diverse thematic streams and scholarly discussions, we encourage active participation, critical reflection, and the pursuit of novel perspectives. It is our hope that this conference will not only disseminate research excellence but also cultivate long-term academic partnerships and thought leadership in the field of digital innovation.

We thank you for your participation in ICDI2025-NNFD and look forward to the rich academic discourse it will undoubtedly inspire.

THE 8TH INTERNATIONAL CONFERENCE ON DIGITAL INNOVATION – NAVIGATING THE NEW FRONTIER OF DISRUPTION

Steering Committee

Honorable Chair	: Dr. Nguyen Mai Lan (NTTU)
Program Chair	: Professor Dr. Premkumar Rajagopal (MUST)
Program Co-Chair	: Dr. Thai Hong Thuy, Khanh (NTTU) Associate Professor Dr. Khashayar Yazdani (MUST)
Conference Advisor	: Professor Emeritus Dr. Barjoyai Bardai
International Advisory Committee	: Professor Dr. Shaheen Mansori Associate Professor Dr. Khairir Khalil Associate Professor Ts. Dr. Asiah binti Lokman Dr. Thanga Thurai Vengadasalam
Secretariat	: Ms. Masturah binti Mohd Nor Ms. Siti Khadijah binti Md Nor Azam

Working Committee

Paper Review Chair	: Dr. Vijayan Ramasamy
Paper Review and Acceptance	: Dr. Kanagaraj Karunanithi : Mr. Chew Yew Chee : Ms. Keertana Naachiyar A/P Ganeson
Proceedings Editorial Board & ISSN	: Ms. Vimala Devi A/P Martham Muthu : Ms. Zunika Samiyah binti Khalib : Ms. Masturah binti Mohd Nor
Proofreading and Editing	: Ms. Paveethra Manian : Ms. Shiva Shanthini A/P Munusamy : Ms. Nurulaini Shariffudin : Ms. Nur Fazlin Akmal Binti Muhammad Faizal : Ms. Sundary N. Rajagopal : Mr. Vincent Paul Jr A/L Aloysious Anthony : Ms. Masturah binti Mohd Nor : Ms. Siti Khadijah binti Md Nor Azam : Ms. Keertana Naachiyar A/P Ganeson
Communications (Internal & External)	: Ms. Siti Khadijah binti Md Nor Azam (Internal) : Ms. Masturah binti Mohd Nor (Internal) : Ms. Nor Syafiqah binti Saman (External - Partners)
Industry Engagement	: Mr. Thiagarajan Marappan : Mr. Balachandar V Rajoo
Internal Proceedings Reviewers	: Professor Emeritus Dr. Barjoyai Bardai : Professor Dr. Shaheen Mansori : Professor Dr. Nor Adnan Yahaya : Professor Dr. David Ngo Chek Ling

Internal Proceedings Reviewers	: Associate Professor Dr. Harwindar Singh : Associate Professor Dr. Khairir Khalil : Associate Professor Dr. Anthony Vaz : Associate Professor Ts. Dr. Asiah Lokman : Associate Professor Dr. Khashayar Yazdani : Associate Professor Dr. Anbalagan Krishnan : Associate Professor Dr. Jerry Chow Tong Wooi : Dr. Thanga Thurai Vengadasalam : Dr. Kamal Abd Razak : Dr. Vijayan Ramasamy : Dr. Ma Tin Cho Mar : Dr. Rajendran Narayanasamy : Ts. Dr. Shanmuga Pillai Murutha Muthu : Dr. Kanagaraj Karunanithi : Dr. Durairaj Veeraiyah : Dr. Ahmad Amru bin Mohamad Zaid : Mr. Sophyan Hakim bin Mohd Hakim : Mr. Chew Yew Chee : Mr. Mohd Faizal Sulaiman : Ms. Nurulaini Shariffudin
External Proceedings Reviewers	: Professor Dr. Sellappan Palaniappan : Dr. Nguyen Thi Thanh Tram : Dr. Phan Bun Gia Thuy : Dr. William Wong Tech Foo : Dr. Nguyen Quang Anh : Dr. Trinh Xuan Hoang : Dr. Nguyen van Tac : Dr. Thai Hong Thuy, Khanh : Dr. Ang Ling Weay
Website Update and Maintenance	: Ms. Gomathi Packiry : Mr. Hafiz Faqrullah bin Mazlan
Conference Kit, Logistics, and Facilities	: Ms. Nurul Sharafina binti Abdul Gani
Program Book, Conference Kit, Certificates & Awards, and Posters	: Ms. Masturah binti Mohd Nor : Ms. Siti Nur Illiyeen binti Zulkiflie : Ms. Nurul Fatinzaiha binti Ahmad : Ms. Aida Wahidah binti Wahid : Ms. Nurhazwani binti Ab. Kauwi
Venue, Plan Execution, and Presentation	: Mr. Nguyen Ngoc Hoa : Dr. Phan Bui Gia Thuy : Ms. Nguyen Thi Thu Hang : Ms. Tran Thi Hao
Finance	: Ms. Khadijah Madinah binti Haris

- Parallel Sessions and Moderators
- : Ms. Gomathi Packiry
 - : Professor Dr. Shaheen Mansori
 - : Associate Professor Dr. Khashayar Yazdani
 - : Ms. Siti Khadijah binti Md Nor Azam

 - : Ms. Sundary N. Rajagopal
 - : Dr. Thanga Thurai Vengadasalam
 - : Dr. Ma Tin Cho Mar
 - : Ms. Nur Fazlin Akmal binti Muhammad Faizal

 - : Ms. Keertana Naachiyar A/P Ganeson
 - : Professor Emeritus Dr. Barjoyai Bardai
 - : Associate Professor Dr. Harwindar Singh
 - : Ms. Aida Wahidah binti Wahid

 - : Mr. Thiagarajan Marappan
 - : Associate Professor Dr. Khairir Khalil
 - : Dr. Rajendran Narayanasamy
 - : Ms. Asallina Daya Anak Charlie

 - : Mr. Vincent Paul Jr A/L Aloysious
 - : Associate Professor Ts. Dr Asiah binti Lokman
 - : Dr. Ahmad Amru bin Mohamad Zaid
 - : Ms. Ereen Nurtiara binti Mohd Azrin
 - : Ms. Siti Nur Illiyeen binti Zulkiflie

Preventing Cyberbullying on Reddit: A Machine Learning Approach

Syeda Zainab Akhtar^a & Asiah Lokman^{*}

^aPhD Informatics, Malaysia University of Science and Technology, Malaysia

^{*}syeda.akhtar@phd.must.edu.my

Abstract: The Internet has become an essential part of human life, enhancing social interactions through networking platforms and discussion forums. However, it also introduces challenges such as cyberbullying, which is increasingly prevalent and often goes unreported. Even when reported, the damage is already done, as victims are affected by harmful comments or posts. While most prior research focuses on post-hoc detection of cyberbullying after harmful content has been posted, this study proposes a real-time pre-screening tool that detects potentially harmful comments *during comment submission* for moderation on Reddit. To achieve this, an integrated machine learning approach is employed, combining outputs from multiple models. The research follows a data science methodology comprising six phases: business understanding, analysis, data collection, data understanding and preparation, modeling, and evaluation. It contributes to the Software Engineering body of knowledge within the Algorithm and Complexity domain. This study evaluates the effectiveness of sentiment analysis libraries (TextBlob and VADER) and classification-based machine learning models, including Logistic Regression, LightGBM, XGBoost Classifier, Linear Support Vector Machine, Random Forest, Stochastic Gradient Descent (SGD), and Universal Language Model Fine-tuning for Text Classification (ULMFiT). The proposed hybrid model consists of three sub-modules: (1) a bad-words detection algorithm, (2) sentiment analysis where VADER achieves 96.3% accuracy for positive sentiments and 72.9% for negative sentiments, and (3) machine learning classifiers, with ULMFiT achieving the highest performance. The final integrated model, combining sentiment and classification outputs through majority voting, achieved an overall accuracy of 95.78%, along with a precision of **92.4%**, recall of **90.8%**, and F1-score of **91.6%**, addressing the common challenge of class imbalance in cyberbullying detection. Our hybrid approach outperforms prior single-feature models by approximately **6.3%**, demonstrating improved robustness in early cyberbullying detection.

Keywords: Cyberbullying Detection, AI in Cybersecurity, Machine Learning, Natural Language Processing (NLP), Sentiment Analysis, Reddit Discussions, Text Classification, Deep Learning

1. Introduction

Social media platforms have become an integral part of daily life, particularly among young people who actively engage in sharing moments, thoughts, and experiences online. These platforms offer a space for communication, entertainment, and information exchange. Some of the most widely used social networking sites include Facebook, Twitter, Instagram, and Reddit. Each platform serves a unique purpose, catering to different types of users. For instance, Facebook is often used for connecting with friends and family, Twitter for real-time discussions and news updates, Instagram for visual storytelling, and Reddit for in-depth discussions and problem-solving.

Reddit, in particular, stands out as a dynamic platform where users can find solutions to a wide range of issues, whether related to personal matters, professional challenges, technology, music, or general life experiences. Unlike other social media platforms, Reddit is structured around topic-based communities known as “subreddits,” where individuals with similar interests engage in discussions, share knowledge, and offer support to one another. The platform thrives on collective participation, as users contribute through posts, comments, and voting mechanisms that determine the visibility of content.

With over 330 million active users, nearly 150,000 communities, and billions of monthly views, Reddit has established itself as one of the most influential discussion portals on the internet. In 2018 alone, the platform recorded 153 million posts, 1.2 billion comments, and a staggering 27 billion votes (Cassis & Bassett, 2018). This level of engagement reflects the significance of online interactions in today's digital landscape, where people rely on virtual communities to seek advice, voice opinions, and share personal experiences.

However, while social media platforms like Reddit foster meaningful discussions, they also come with challenges, one of the most concerning being cyberbullying. As social media continues to grow, so does the risk of online harassment, particularly among teenagers and young adults. Cyberbullying can take various forms, including hurtful comments, targeted attacks, and negative discussions that may harm an individual's mental well-being. Young users, in particular, are highly vulnerable to such experiences, as they may take negative remarks seriously, even when they are minor.

The psychological impact of cyberbullying can be severe. Many young individuals fear discussing their experiences with parents, teachers, or friends, leading to feelings of isolation, anxiety, and depression. In extreme cases, prolonged exposure to cyberbullying can contribute to self-esteem issues, emotional distress, and even suicidal thoughts. Given the increasing prevalence of cyberbullying, there is a growing need for awareness, digital safety measures, and support systems to help individuals navigate the online world responsibly.

As social media continues to evolve, it is crucial for platforms to implement strict policies against cyberbullying and provide resources for users to report harmful content. Additionally, fostering a culture of empathy and responsible online behaviour can play a significant role in making the internet a safer space for everyone.



Figure 1. Research Concept (AI Generated Image)

2. Research Methodology

2.1 Proposed Method

Research methodology for this paper is divided into 6 phases.

2.1.1 Phase 1: Business Understanding

This research focuses on detecting cyberbullying on Reddit, which currently lacks automated moderation. A machine learning model will be developed to block harmful comments before posting.

2.1.2 Phase 2: Analysis

A literature review from IEEE, Scopus, and Web of Science covers bag-of-words, word embedding, sentiment analysis, NLP, and machine learning techniques to enhance detection.

2.1.3 Phase 3: Data Collection

Data is sourced from Pushshift (Reddit corpus) and Carnegie Mellon's offensive words database. Development uses Google Colab (GPU support), Microsoft Azure, and local Python setups.

2.1.4 Phase 4: Data Understanding and Preparation

Preprocessing removes unnecessary symbols, emojis, and stopwords. The offensive words dataset labels Reddit comments as cyberbullying or normal. The dataset was split using an 80:20 stratified split to preserve the proportion of cyberbullying to normal comments across training and testing sets. Different subreddits exhibit distinct linguistic patterns, such as the prevalence of slang, abbreviations, and emojis in casual communities like teenagers, compared to more formal, technical language in r/science. Our preprocessing pipeline was designed to normalize such variations to improve generalization across subreddit contexts.

2.1.5 Phase 5: Modeling

The model consists of:

- a. **Bad Words Filter:** Flags offensive terms.
- b. **Sentiment Analysis:** Uses TextBlob or VADER.
- c. **Classification Model:** Tests various algorithms (Logistic Regression, Random Forest, XGBoost, etc.), integrating the best-performing model.

2.1.6 Phase 6: Evaluation

Performance is analyzed through accuracy metrics, confusion matrices, and gradient descent, selecting the most effective cyberbullying detection model.

2.2 Experimental Setup

The experimental setup for detecting cyberbullying in the proposed model, which follows three main phases. The first phase involves collecting and preprocessing data by removing unnecessary elements. The second phase detects cyberbullying comments using a predefined dictionary of offensive and harmful words.

Sentiment analysis is then applied to determine whether a comment carries a positive or negative tone. In the final phase, a classification model evaluates whether the comment falls under cyberbullying. Based on these assessments, the model provides a recommendation on whether the comment should be allowed or restricted.

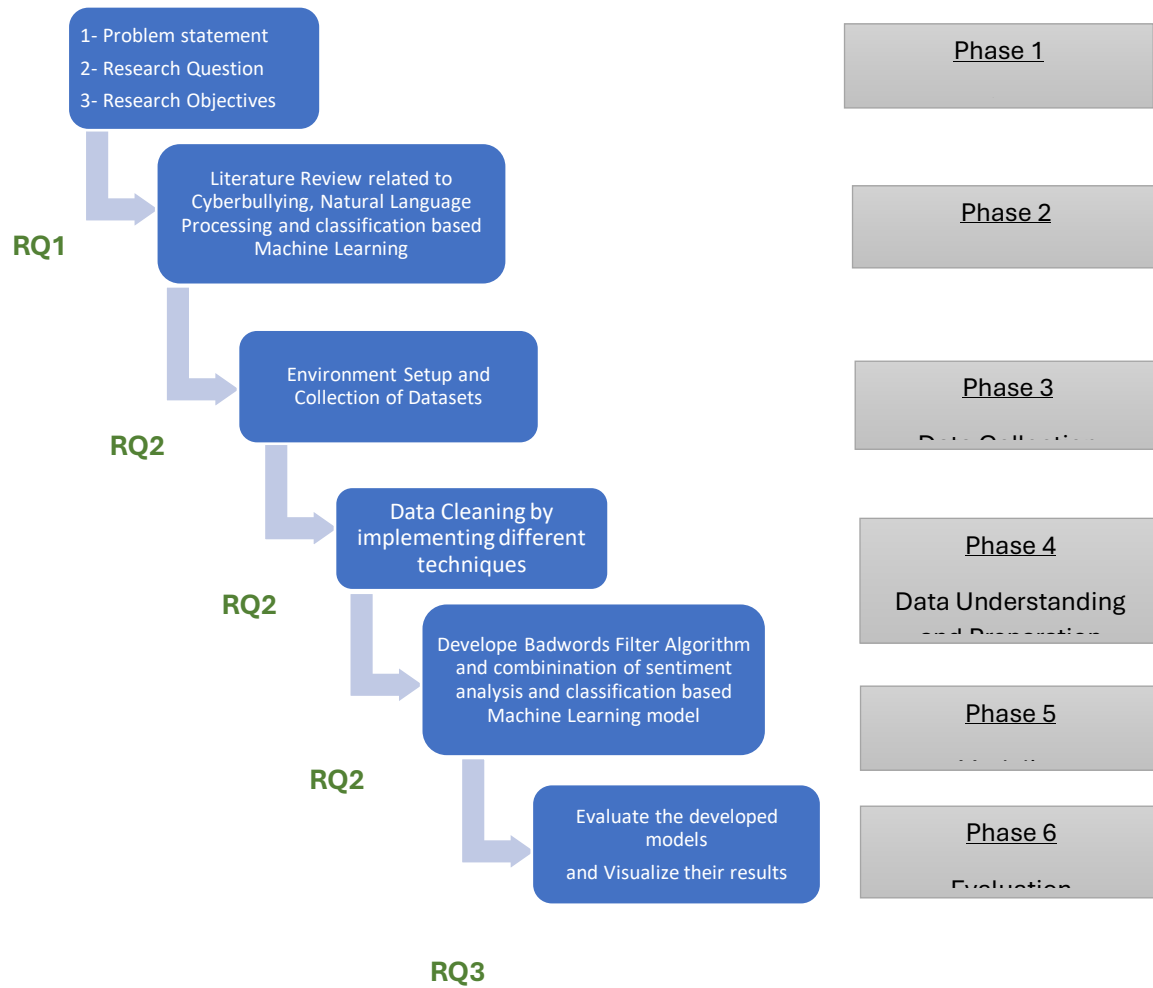


Figure 2. Research Phases

3. Literature Review

This paper reviewed relevant literature, beginning with an overview of cyberbullying, its categories, and detection methods. It then explored social network mining, focusing on network graph mapping in social media. Various word embedding techniques, including One-Hot Encoding, Word2Vec, TF-IDF, FastText, and GloVe (Global Vectors for Word Representation), were examined. The paper also covered Natural Language Processing, Sentiment Analysis, and machine learning approaches, including supervised, semi-supervised, and unsupervised learning. Lastly, different machine learning models such as Logistic Regression, LightGBM, XGBoost, Random Forest, SGD, and ULMFiT were discussed.

3.1 Cyberbullying

Young people increasingly use the Internet, with most having access at home and on mobile devices. For many, the Internet is more than just a tool for learning or entertainment—it is an essential part of daily life. Common online activities among youth include sending emails, chatting with friends, and playing games. However, just as in real-life interactions, some individuals engage in harassment online (National Crime Prevention Council, 2020). Cyberbullying mirrors traditional bullying but occurs through digital platforms, including social media and text messages. Perpetrators may be peers, online acquaintances, or even anonymous users, though they often know their victims. According to Willard (2007) and Nadali et al. (2014), cyberbullying can take various forms, including:

- a. **Flaming:** Engaging in hostile online arguments.
- b. **Masquerading:** Pretending to be someone else to cause harm.
- c. **Denigration:** Spreading false or damaging information to ruin a person's reputation.
- d. **Impersonation:** Assuming another person's identity to create trouble or harm relationships.
- e. **Harassment:** Repeatedly sending offensive and abusive messages.
- f. **Outing:** Publicly sharing private or embarrassing information.
- g. **Trickery:** Manipulating individuals into revealing secrets and then exposing them online.
- h. **Exclusion:** Intentionally isolating someone from an online group.
- i. **Cyberstalking:** Persistent and severe online harassment that includes threats and intimidation, causing significant distress to the victim.

Whereas, recent studies have demonstrated the effectiveness of transformer-based models such as BERT and RoBERTa in cyberbullying detection tasks, leveraging contextualized embeddings to capture nuanced language patterns (Waseem et al., 2020; Dinakar et al., 2021). However, such models are often computationally intensive, limiting their real-time application in comment moderation systems. Our model balances performance and efficiency by integrating lightweight linguistic and sentiment features with classification models.

3.2 Features Used in Cyberbullying Detection

Before exploring details, the features used for cyberbullying detection can be categorized into four main types: content-based, user-based, sentiment-based, and network-based (Salawu, He, & Lumsden, 2017).

A study by Dinakar, Picard, and Lieberman (2015) identified profanity, subtlety, and negativity as key content-based features for detecting cyberbullying. Annotation classified subjects into three categories: race and culture, intelligence, and sexuality. Similarly, Van Hee et al. (2015) used sentiment-based (polarity) and content-based (BoW) features, concluding that a single feature is insufficient for accurate cyberbullying detection. Combining multiple features significantly improves the F-score, leading to better detection accuracy.

Hosseinmardi et al. (2016) incorporated various features such as cyberaggression, profanity, image content, network graphs, and linguistic elements. Their study analyzed network metrics like comments, likes, fans, and followers, collectively termed as a media session. The findings suggest that cyberbullying detection should not rely solely on profanity, as bullying-related media sessions often receive fewer likes despite having a high number of followers.

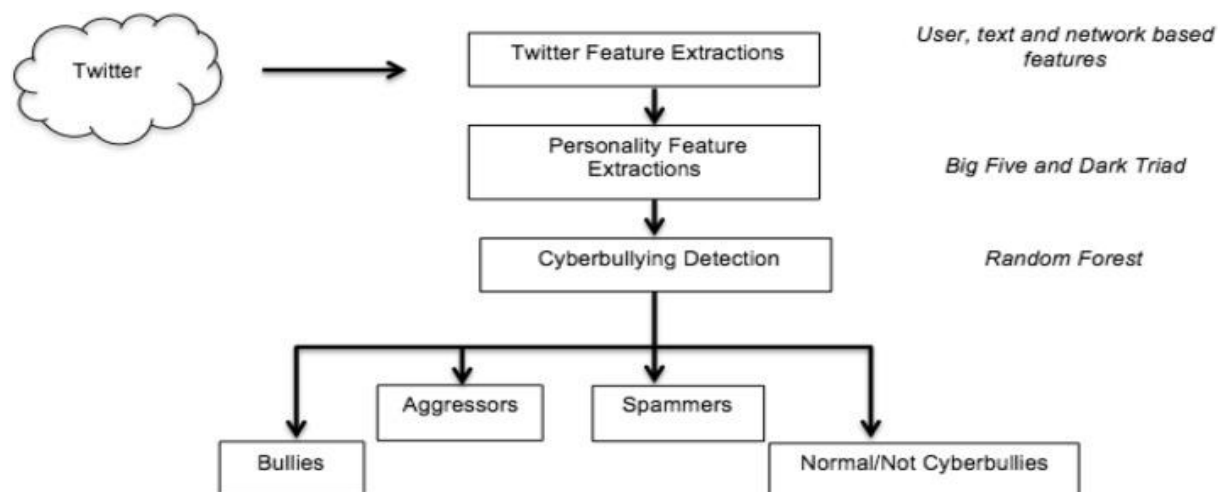


Figure 3. Feature selection technique to classify user types (Balakrishnan et al. 2019)

3.3 Social Networks Mining

Social Networks Mining (SNM) is an emerging research field that combines social media analysis with data mining techniques. It plays a crucial role in analyzing social media data through methods such as Web Mining, Social Network Analysis, Text Mining, Sentiment Analysis, Natural Language Processing, and Opinion Mining. These techniques are widely used on social networking platforms to assess user opinions. Some of these approaches have proven effective in addressing the limitations of current cyberbullying detection methods (Ting et al., 2017).

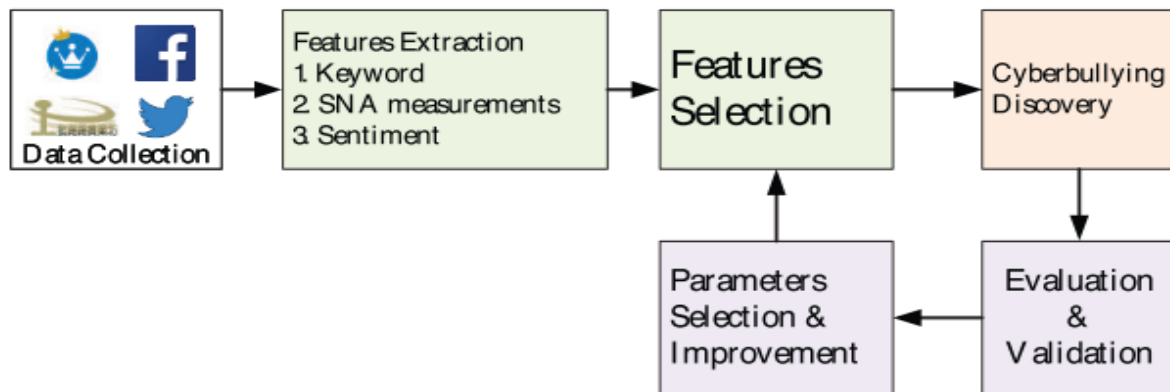


Figure 4. SNM based Cyberbullying detection (Ting et al. 2017)

3.4 Word Embedding

Social Networks Mining (SNM) integrates social media analysis with data mining techniques to enhance cyberbullying detection. Various word embedding methods improve text processing in machine learning. One-hot encoding converts words into frequency-based vectors, while Word2Vec captures contextual meanings through CBOW and multi-word context techniques. TF-IDF filters out common words by assessing their significance in documents. FastText enhances word representation by breaking words into character-based n-grams. GloVe analyzes word co-occurrence to capture semantic relationships. These techniques collectively strengthen natural language processing models for improved text classification and cyberbullying detection.

3.5 Machine Learning Types

Machine learning is divided into supervised, semi-supervised, and unsupervised learning, commonly classified as classification and clustering methods. Supervised learning relies on labelled data to make predictions, whereas unsupervised learning identifies patterns without predefined categories. Since the Reddit dataset in this study is unlabelled, unsupervised learning techniques are applied to process the data.

Supervised Learning includes probabilistic and linear classifiers, with widely used techniques such as k-nearest neighbour (k-NN), support vector machines (SVM), neural networks, linear least squares fit (LLSF), and naive Bayes (NB). Studies indicate that these methods perform optimally when there is a sufficient amount of training data per category (Tang et al., 2014).

Unsupervised Learning is useful for analyzing unlabelled data through models like probabilistic Latent Semantic Analysis (pLSA) a statistical technique for discovering topics in large text corpora based on word co-occurrence patterns and Latent Dirichlet Allocation (LDA) (Madhoushi, Hamdan, & Zainudin, 2015). However, these methods often require large datasets to ensure accuracy and may generate topics that do not always align with human interpretation. Despite these challenges,

unsupervised learning remains a valuable approach for discovering patterns without the need for labelled data.

4. Experimental Results

This paper also discusses the results from the experiment conducted in the previous chapter. The experiment includes Sentiment Analysis and different machine learning model including Logistic Regression, Light GBM, XG Boost Classifier, Linear Support Vector Machine, Random Forest, Stochastic Gradient Descent and Universal Language Model Fine Tuning.

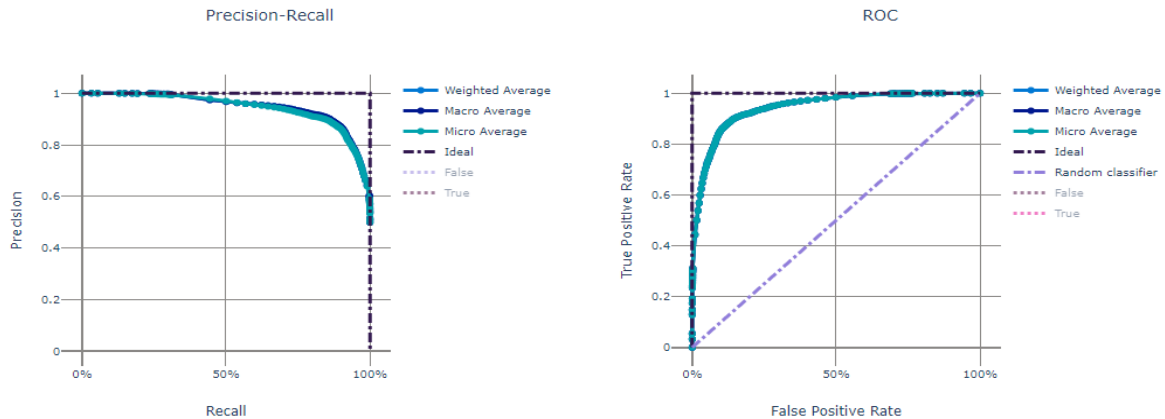


Figure 5. Precision-Recall and ROC for XG Boost Classified

Preprocessing involved converting emojis into text descriptors (e.g., 😊 → 'smiling face') using the emoji Python package. Special symbols and excessive punctuation were removed. All text was lowercased. Lemmatization was performed using the WordNetLemmatizer from the NLTK library to standardize word forms (e.g., 'running' → 'run')

4.1.1 Sentiment Analysis

The results indicate that TextBlob correctly identifies 84% of normal comments as positive or neutral but only detects 37.87% of cyberbullying comments as negative. Figure 5.1 visually represents these findings.

4.1.2 Classification based Machine Learning Models

The model's accuracy improved across training epochs, reaching 91.28% in the first cycle and 94.99% in the final cycle. Training and validation loss remained stable, indicating neither overfitting nor underfitting. The confusion matrix showed that the model correctly identified cyberbullying comments with 92.2% accuracy. Sentiment analysis using VADER outperformed TextBlob, and ULMFiT (Universal Language Model Fine-tuning for Text Classification; a transfer learning method that fine-tunes a pre-trained language model for specific text classification tasks) achieved the highest classification accuracy compared to other models. A majority voting technique was used for final predictions, ensuring the most accurate model determined the outcome.

Table 15. Neural Network Outer Layer Training

epoch	train_loss	valid_loss	accuracy	Time
0	0.572573	0.475073	0.770817	05:15
1	0.539237	0.466670	0.773683	05:13

Table 16. Complete Neural Network Training

epoch	train_loss	valid_loss	accuracy	time
0	0.410436	0.297686	0.869483	05:00
1	0.308408	0.224270	0.905183	05:10
2	0.293449	0.209388	0.913917	04:27
3	0.278269	0.215543	0.912767	04:12

The ULMFiT model was fine-tuned over 3 epochs with a learning rate of 1e-3, batch size of 64, and weight decay of 1e-2. Early stopping was applied based on validation loss.

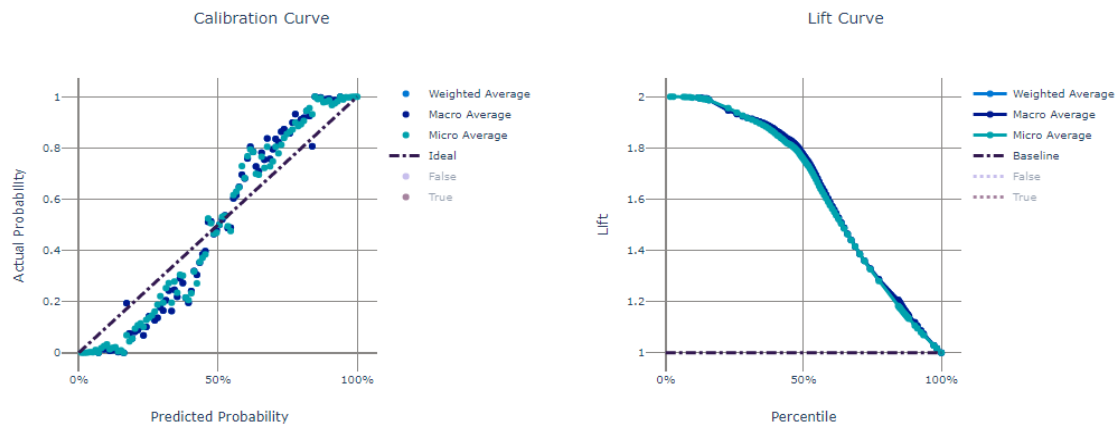


Figure 6. Calibration Curve

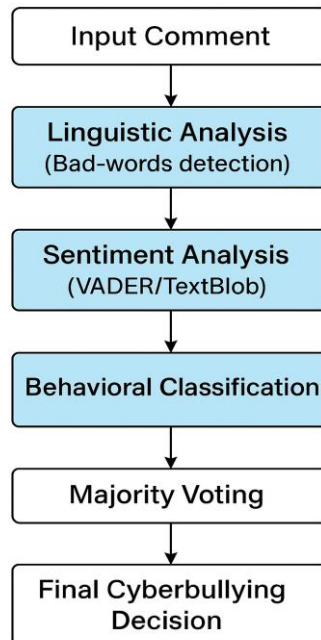


Figure 7. Architecture diagram

4.1.3 Final Analysis

The experimental results for cyberbullying detection. VADER outperformed TextBlob in sentiment analysis, identifying normal comments with 96.3% accuracy and cyberbullying comments with 72.9%, while TextBlob detected only 37.87% of cyberbullying cases. The dataset was split using an 80:20 stratified split to preserve the proportion of cyberbullying to normal comments across training and testing sets. Among classification models, ULMFiT achieved the highest accuracy (94.99%), surpassing LightGBM (93.21%), SGD (90.89%), XGBoost (87.91%), SVM (80.45%), Logistic Regression (77.88%), and Random Forest (56.15%). The final integrated model, combining VADER and ULMFiT using majority voting, achieved 95.78% accuracy, successfully meeting the study's objectives.

5. Limitations

One of the major challenges in cyberbullying detection is the availability of datasets. Extracting data from social media platforms is complex due to strict privacy regulations and the limited accessibility of publicly available data. Since the data were collected from publicly available Reddit posts, formal IRB approval was not required. However, all usernames and any personally identifiable information (PII) were removed to ensure participant anonymity. Furthermore, annotating and labeling data is a labor-intensive task that requires expert involvement to ensure accuracy and reliability. Another significant hurdle is the detection of sarcasm and irony, as these forms of expression often rely on contextual understanding. Accurately identifying sarcastic or ironic remarks remains difficult because it requires analyzing the broader conversation rather than just isolated comments. Due to these limitations, achieving high accuracy in cyberbullying detection continues to be a challenge, highlighting the need for more advanced contextual analysis techniques and improved dataset accessibility.

6. Future Enhancements

Future research in this area can focus on several key improvements to enhance the effectiveness of cyberbullying detection. One important aspect is the identification of sarcasm, as many individuals use sarcastic remarks as a means of bullying. Integrating sarcasm detection into the model would improve its ability to recognize harmful intent hidden within seemingly harmless comments. Similarly, irony detection could be incorporated to further strengthen the system's capability in identifying subtle forms of cyberbullying, where users disguise offensive remarks under ironic statements. Additionally, expanding the dictionary of offensive words and phrases would improve the model's accuracy in detecting harmful content. Training the model on a significantly larger dataset would allow for better generalization across different contexts and user behaviors. However, this would require increased computational resources and funding to support higher processing power and storage capacity. By addressing these areas, the model could evolve into a more robust and comprehensive system for combating cyberbullying on online platforms.

Acknowledgements

We would like to thank all the people who prepared and revised previous versions of this document.

References

- Balakrishnan, Vimala, Shahzaib Khan, Terence Fernandez, and Hamid R. Arabnia. 2019. "Cyberbullying Detection on Twitter Using Big Five and Dark Triad Features." *Personality and Individual Differences*.
- Dinakar et al., "Detecting cyberbullying with contextualized language models," *ACL 2021*.

- Dinakar, Karthik et al. 2012. “Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying.” *ACM Transactions on Interactive Intelligent Systems*.
- Felbo, Bjarke et al. 2017. “Using Millions of Emoji Occurrences to Learn Any-Domain Representations for Detecting Sentiment, Emotion and Sarcasm.” In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*,.
- Görzig, Anke, and Hana Machackova. 2015. “Cyberbullying from a Socio-Ecological Perspective: A Contemporary Synthesis of Findings from EU Kids Online.” *LSE Department of Media and Communications*.
- Hosseinmardi, Homa et al. 2016. “Prediction of Cyberbullying Incidents in a Media-Based Social Network.” In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*,.
- Hutto, C J, and E E Gilbert. 2014. “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).” *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*.
- Kingma, Diederik P., and Jimmy Lei Ba. 2015. “Adam: A Method for Stochastic Optimization.” In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*,
- Lan, Hua, and Yun Pan. 2019. “A Crowdsourcing Quality Prediction Model Based on Random Forests.” *Proceedings - 18th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2019*: 315–19.
- Lin, Tsung Yi et al. 2020. “Focal Loss for Dense Object Detection.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Liu, Bing. 2012. “Sentiment Analysis and Opinion Mining.” *Synthesis Lectures on Human Language Technologies*.
- Long, Jonathan, Evan Shelhamer, and Trevor Darrell. 2015. “Fully Convolutional Networks for Semantic Segmentation.” In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*,.
- Newer works using **RoBERTa** or **DeBERTa** for online toxicity detection.
- Van Hee, Cynthia et al. 2015. “Automatic Detection and Prevention of Cyberbullying.” *International Conference on Human and Social Analytics (HUSO 2015)*.
- Wilson, Ashia C. et al. 2017. “The Marginal Value of Adaptive Gradient Methods in Machine Learning.” In *Advances in Neural Information Processing Systems*.
- Waseem et al., “Leveraging BERT for detecting cyberbullying in social media,” *COLING 2020*.
- Yosinski, Jason, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. “How Transferable Are Features in Deep Neural Networks?” In *Advances in Neural Information Processing Systems*,.
- Zou, Xiaonan, Yong Hu, Zhewen Tian, and Kaiyuan Shen. 2019. “Logistic Regression Model Optimization and Case Analysis.” *Proceedings of IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT 2019*: 135–39.