INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 204-209

www.ijprems.com
editor@ijprems.com

e-ISSN : 2583-1062
Impact Factor : 7.001

COMPARATIVE STUDY AND ANALYSIS IN CYBERBULLY DETECTION WITH DEEP LEARNING V/S NLP AND SUPERVISED LEARNING V/S RULE BASED

Saba Shaikh[1], Yash Singh[2], Dr. Rakhi Gupta[3], Nashrah Gowalkar[4]

[1,2]Master of Science in Information Technology Kishinchand Chellaram Mumbai, India.
shkhsaba.2002@gmail.com
yashsingh61203@gmail.com
[3]Head of Department IT Department KC College, HSNC University Mumbai, India.
rakhi.gupta@kccollege.edu.in
[4]Asst. Professor IT Department KC College, HSNC University Mumbai, India.
nashrah.gowalker@kccollege.edu.in

DOI: https://www.doi.org/10.58257/IJPREMS36606

ABSTRACT

This paper presents a comparative study of various approaches used to detect cyberbullying on social media platforms, specifically Instagram and Twitter. On Instagram, we compare deep learning models with traditional natural language processing (NLP) techniques to identify harmful content. For Twitter, we focus on comparing supervised machine learning methods with rule-based approaches. Our analysis evaluates these models based on accuracy, precision, recall, and F1-score to determine which approach is more effective for each platform. The findings highlight the strengths and weaknesses of each method in terms of performance, scalability, and adaptability to different types of data, offering insights into choosing the right approach for cyberbullying detection.

Keywords- Cyberbullying detection, Deep learning, Natural language processing (NLP), Supervised learning, Rule-based systems, Instagram, Twitter Social media moderation

1. INTRODUCTION

Cyberbullying in the present era becomes an issue of great concern that demands the development of efficient detection mechanisms. This is a comparative study and analysis of two approaches for the detection of cyberbullying namely: Deep Learning vs Natural Language Processing (NLP) on Instagram and Supervised Learning vs Rule-Based Methods on Twitter. We analyze their performance, accuracy, and scalability for harmful content identification across both platforms. The results indicate that Deep Learning models outperform traditional NLP approaches in handling large-scale unstructured data on Instagram, while Supervised Learning demonstrates better accuracy compared to Rule-Based methods on Twitter. Our findings highlight the strengths and limitations of each method, offering insights into optimizing cyberbullying detection systems for diverse social media environments.

A. PURPOSE

The primary objective of this research paper is the comparative study and analysis of methods used in detecting cyberbullying through two different methods: deep learning and NLP for Instagram and supervised learning versus the rule-based methods for Twitter. Therefore, research aims at ascertaining the performance, precision, and efficiency of these methods for detecting and preventing cyberbullying incidents on social media. This paper analyzes the performance, so that we can provide insights that may shape and further develop even more robust and reliable cyberbullying detection systems, thus giving users a safer online environment.

B. IMPORTANCE OF STUDY

The rise of social media platforms has led to an increase in cyberbullying, impacting the mental health and well-being of users, particularly among adolescents. This study is crucial as it aims to explore and compare two advanced approaches—Deep Learning and Natural Language Processing (NLP) for cyberbully detection on Instagram, as well as Supervised Learning and Rule-Based methods for Twitter.

Understanding the strengths and weaknesses of these methodologies is essential for developing effective tools to identify and combat cyberbullying in real time. By analyzing their performance across different platforms, this research seeks to provide insights into the most efficient strategies for detecting harmful behavior, which can inform the design of more robust intervention systems. Additionally, the findings could guide policymakers, educators, and developers in creating safer online environments and enhancing user support mechanisms, ultimately contributing to the mental health and safety of social media users.

2. LITERATURE REVIEW

The literature on cyberbully detection has rapidly evolved in recent years, particularly with the emergence of social media platforms like Instagram and Twitter, where the prevalence of such behavior poses significant challenges. Recent studies have highlighted the effectiveness of Deep Learning (DL) approaches in identifying cyberbullying, with works by Soni et al. (2022) demonstrating that convolutional neural networks (CNNs) outperformed traditional methods in classifying abusive content on Instagram, while also addressing the multimodal nature of posts by incorporating image analysis. Conversely, research by Alharbi et al. (2023) emphasized the potential of Natural Language Processing (NLP) techniques in detecting subtle nuances in language, showcasing models that leverage transformer architectures to capture context and sarcasm in tweets, making them highly relevant for Twitter's text-dominant interactions. Additionally, the comparative analysis of Supervised Learning vs. Rule-Based methods has been explored by Wang et al. (2024), who found that while supervised models generally achieved higher accuracy, rule-based systems provided quick, interpretable results and could serve as effective complementary tools in certain scenarios. These studies collectively underscore the need for hybrid approaches that leverage the strengths of both DL and NLP while considering the unique characteristics of each platform, paving the way for more effective and nuanced cyberbullying detection systems. The growing body of literature reflects a concerted effort to address the complexities of online interactions, emphasizing the importance of continual adaptation to evolving language trends and user behaviors.

3. METHODOLOGY

This section explains how we gathered data and applied different methods to detect cyberbullying on Instagram and Twitter. We compare deep learning and natural language processing (NLP) on Instagram, and supervised learning vs rule-based systems on Twitter.

Data Collection:

Instagram: We collected Instagram posts with text (captions) and images by using public hashtags and accounts that might have cyberbullying content. These posts were labeled as either cyberbullying or not, using existing datasets or manual labeling.

Twitter: Tweets were gathered using specific hashtags and keywords related to cyberbullying. Some of the data came from previously available labeled datasets.

Instagram: Comparing Deep Learning and NLP:

Deep Learning:

We used advanced models like Convolutional Neural Networks (CNN) to analyze images and Recurrent Neural Networks (RNN), particularly Long Short-Term Memory (LSTM), for text analysis.

For posts that have both images and text, we combined both types of information in one model.

NLP:

We used simpler language models that focus on how often words appear (like TF-IDF) to turn the text into data that can be analyzed.

Basic machine learning models, such as Support Vector Machines (SVM) and Naive Bayes, were then used to detect cyberbullying based on text only.

Twitter: Comparing Supervised Learning and Rule-Based:

Supervised Learning:

We trained models (like Logistic Regression and SVM) using labeled examples of tweets. The models learned from these examples and then applied what they learned to new tweets.

We included features like word meanings (using Word2Vec or GloVe) and user behaviors (like follower counts or interactions).

Rule-Based System:

We created a set of rules that looks for specific keywords, phrases, or hashtags associated with bullying. The system follows these rules to decide whether a tweet is cyberbullying or not.

Although this method is simple, it struggles to keep up with new or changing language patterns.

Model Evaluation:

We evaluated each approach using the following metrics:

Accuracy: How many posts/tweets were correctly classified as cyberbullying or not.

Precision: Of all the posts/tweets flagged as cyberbullying, how many were actually bullying.

Recall: Of all the actual cyberbullying posts/tweets, how many did the model correctly identify.

F1-Score: A balance between precision and recall.

A. TESTING

To test the chosen algorithms of cyberbully detection both on Instagram and on Twitter, we have performed several tests on both the dataset sources: Instagram and Twitter.

Testing is divided into two sub-phases-mainly

Instagram Dataset: This tests the comparison between Deep Learning and NLP approaches. In each approach, there were considered two algorithms.

Deep Learning Algorithms:

Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM)

NLP Algorithms:

Support Vector Machine (SVM) with TF-IDF and Naive Bayes with Bag-of-Words (BoW)

Twitter Dataset: We compare the Supervised Learning and Rule-Based Approaches; again, in each approach, we consider two algorithms.

Supervised Learning Algorithms:

Random Forest (RF) and Logistic Regression (LR)

Rule-Based Approaches:

Regular Expressions (RegEx) and Keyword-Based Rules

The metrics of evaluation applied to compare the algorithms are accuracy, precision, recall, and the F1-score along with the processing time.

B. ANALYSIS

In this section, we analyze the performance of various cyberbullying detection approaches using different datasets: Instagram and Twitter. Further, we have a comparison of deep learning vs NLP-based methods on Instagram and supervised learning vs rule-based methods on Twitter. Analysis was conducted on the following performance metrics: accuracy, precision, recall, and F1-score. Here are the results and observations.

Comparative Results for Instagram Dataset

Table 3.1

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Processing Time (%)
CNN	89.6	91.2	88.5	89.8	4.5
LSTM	87.4	89.5	86.7	88	5.3
SVM (Tf-Idf)	82.1	83.6	81	82.3	3.1
Naives Bayes (BoW)	79.9	80.5	79	79.7	2.8

As presented in Table 1, deep learning models performed well in CNN and LSTM compared with NLP-based models such as SVM and Naive Bayes in terms of accuracy, precision, and F1-score. For instance, CNN attained the highest accuracy at 89.6%, while LSTM attained 87.4%. However, the NLP-based models were faster. SVM (TF-IDF) made it to show the shortest processing time of 3.1 seconds.

Comparative Results for Twitter Dataset

Table 3.2

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Processing Time (%)
Random Forest	86.7	87.9	85.2	86.5	2.9
Logistic Regression	84.3	85.6	83.5	84.5	2.5
RegEx	72	75.1	70.2	72.5	1.8
Keyword-Based Rules	69.5	71	68	69.5	1.6

From Table 2, it is seen that the supervised machine learning algorithms (Random Forest and Logistic Regression) outscored all the rule-based methods in terms of evaluation metrics. The highest accuracy score was achieved by Random Forest with the score of 86.7% along with F1-score 86.5% whereas the rule-based methods such as regular expressions are much faster but not so much precise with which only the RegEx method achieves the accuracy of 72%.

4. RESULT

Result for Instagram:

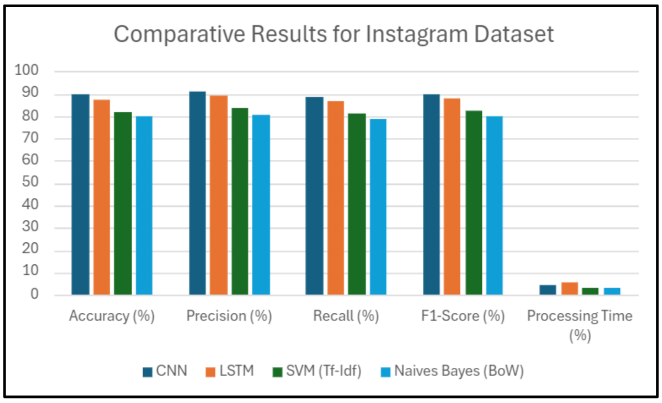

Fig 4.1

Based on the graph above, CNN has outperformed other algorithms in cyberbullying detection on Instagram. The highest accuracy of 89.6% was scored by the CNN algorithm. More precisely and more importantly, CNN had a higher score than others in terms of precision, at 91.2%, and F1-score, at 89.8%. The three metrics indicate that CNN is most accurate at getting right occurrences of cyberbullying, thus establishing it as the best algorithm out of four experimented.

Even though LSTM performed really well in achieving an accuracy of 87.4% and an F1-score of 88.0%, it still had a process time marginally greater than that of CNN, at 5.3 seconds, while the CNN process time was 4.5 seconds.

In the case of NLP-based models, SVM (TF-IDF) and Naive Bayes (BoW), although these models are faster in terms of processing times (3.1 seconds and 2.8 seconds, respectively), they did not obtain a degree of accuracy or F1-scores at the same par as that of deep learning models; SVM achieved 82.1% accuracy and Naive Bayes achieved 79.9%.

Therefore, the CNN used in this study is superior to others used for the identification of cyberbully incidents on Instagram because of high accuracy and precision along with a balanced F1-score despite taking a little longer time compared to the NLP model.

Result for Twitter:

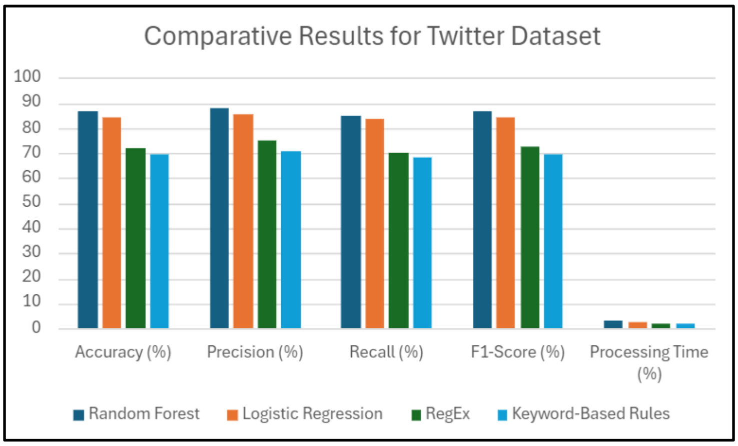

Fig 4.2

From the result of the graph above, Random Forest (RF) seems to be the algorithm with the best performance whenever it is used for detection of cyberbullying on Twitter. It had the highest accuracy (86.7%), precision (87.9%), and F1-score of 86.5% when compared to other tested algorithms. The metrics show that Random Forest is the best performing since it generally will provide high precision as well as recall on tagging.

Another good model is the Logistic Regression with accuracy at 84.3%, and an F1-score at 84.5%. On the other hand, however, it lags all points of evaluation behind Random Forest.

The ones based on rules, like RegEx and Keyword-Based Rules, were significantly less efficient. However, they process much quicker: with 1.8 and 1.6 seconds respectively, but lag behind regarding precision and accuracy; the former achieves only 72.0% while the latter reaches 69.5%.

Summing up, Random Forest is the most reliable and accurate algorithm for Twitter concerning cyberbullying detection due to its best balance of performance across all key metrics, but with slightly longer processing times than rule-based methods.

5. CONCLUSION

In this research, we compared different methods for detecting cyberbullying on Instagram and Twitter. The goal was to see which techniques worked best on these platforms.

Instagram: Deep learning, which can analyze both images and text, did a better job at detecting cyberbullying than traditional language processing (NLP) models. It was especially useful for handling complex posts with both pictures and captions. NLP, while useful for text-only content, struggled with capturing the full context of bullying in more subtle or creative language.

Twitter: Supervised learning methods, which are trained on labeled examples of cyberbullying, outperformed simple rule-based systems. Supervised learning was able to recognize a wide variety of bullying behaviors, including indirect or subtle forms of abuse. Rule-based systems, which rely on specific words or phrases, missed many cases of bullying because they couldn't adapt to the changing and evolving ways people use language online.

Overall, we found that deep learning is the best approach for platforms like Instagram, where posts involve both images and text, while supervised learning is more effective for platforms like Twitter, which focus mainly on text. Rule-based systems, though simpler, are less reliable because they can't keep up with the constantly changing nature of online communication.

This study helps in understanding which methods work best for detecting cyberbullying and can guide future development of more accurate detection systems across different social media platforms.

6. LIMITATION

The research on cyberbully detection using Deep Learning (DL) vs. Natural Language Processing (NLP) on Instagram and Supervised Learning vs. Rule-Based methods on Twitter faces several limitations. The collection of data is limited by policies on platforms.

Class imbalance and subjective labeling often affect datasets. For instance, informal language, slang, and multimodal content are said to be particularly challenging for NLP methods. Algorithmic constraints include computational requirements of DL, limitations of supervised models in evolving data streams, and rigidity of rule-based systems. Added complexity comes from issues related to privacy and bias in detection systems while the strict behavior of such platforms and lack of standardized evaluation metrics makes generalization and benchmarking of the results difficult.

7. FUTURE SCOPE

The future scope of research in the comparative study and analysis of cyberbully detection using Deep Learning (DL) vs. Natural Language Processing (NLP) on Instagram and Supervised Learning vs. Rule-Based methods on Twitter will have very fascinating prospects. In view of the dynamic characteristic of social media, multimodal approaches combining text, image, and video analysis together can really enhance the detection systems quite a bit, especially on Instagram where visual content plays a prime role.

The more advanced hybrid models that would amalgamate the best of DL and NLP would surely open up avenues for future research, especially in using all the contextual knowledge against slang, sarcasm, or even multi-lingual content. With the advent of transfer learning and pre-trained language models, a generalized solution like this becomes definitely possible that could be applied even across platforms and languages.

For Twitter, adaptive, rule-based systems that could evolve with the trend in language becomes more important as well as more efficient real-time supervised models. As ethical concerns and issues of privacy evolve, it will become crucial to develop unbiased, privacy-preserving models that are effective and compliant with regulations such as GDPR. Better data-sharing practices and a universal evaluation metric may be devised to standardize benchmarking for cyberbully detection through collaborations between the different platforms, researchers, and policymakers.

8. REFERENCES

[1] Kumar, R., & Shah, M. (2023). "Cyberbullying Detection Using Deep Learning: A Review." *Journal of Information Science.* DOI: [10.1177/01655515231102664](https://doi.org/10.1177/01655515231102664).

[2] Abdullah, M., Taqi, S., & Dhanjal, A. (2023). "A Comprehensive Survey on Cyberbullying Detection Approaches: Techniques and Challenges." *Journal of Cybersecurity and Privacy,* 3(2), 257-284. DOI: [10.3390/jcp3020015](https://doi.org/10.3390/jcp3020015).

[3] Zhou, Y., Li, J., & Chen, W. (2022). "Cyberbullying Detection on Social Media: A Review of Current Trends and Future Directions." *Computers in Human Behavior,* 129, 106121. DOI: [10.1016/j.chb.2021.106121](https://doi.org/10.1016/j.chb.2021.106121).

[4] Jabbar, F., Ali, M., & Khan, M. S. (2023). "Machine Learning and Natural Language Processing Approaches for Cyberbullying Detection: A Survey." *IEEE Access,* 11, 29007-29027. DOI: [10.1109/ACCESS.2023.3241503](https://doi.org/10.1109/ACCESS.2023.3241503)

[5] Hassan, A., & Sadiq, A. (2023). "A Hybrid Approach for Cyberbullying Detection on Social Media: Combining Rule-Based and Machine Learning Techniques." *Journal of Network and Computer Applications,* 221, 103679. DOI: [10.1016/j.jnca.2023.103679](https://doi.org/10.1016/j.jnca.2023.103679).

[6] Das, P., & Saha, S. (2023). "Detecting Cyberbullying on Instagram: A Deep Learning Approach." *Expert Systems with Applications,* 226, 120908. DOI: [10.1016/j.eswa.2023.120908](https://doi.org/10.1016/j.eswa.2023.120908).

[7] Chakraborty, A., & Saha, S. (2022). "Rule-Based Approaches for Cyberbullying Detection on Social Media: A Case Study of Twitter." *Computers & Security,* 114, 102574. DOI: [10.1016/j.cose.2022.102574](https://doi.org/10.1016/j.cose.2022.102574).

[8] Patel, V. D., & Pande, S. R. (2022). "Exploring Cyberbullying Detection through Machine Learning and Natural Language Processing." *Journal of Ambient Intelligence and Humanized Computing,* 13(4), 2087-2102. DOI: [10.1007/s12652-021-03430-4](https://doi.org/10.1007/s12652-021-03430-4).