International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Cyberbullying and Fake Account Detection on Social Media Using Machine Learning and NLP

B.Amarnath Reddy[1], N.V.Sai Vyshnavi[2], P.Deepthi[3], Y.S.N.Raju[4]

[1]*Cyber Security & Raghu Engineering College*
[2]*Cyber Security & Raghu Engineering College*
[3]*Cyber Security & Raghu Engineering College*
[4]*Associate Professor, CSO & Raghu Engineering College*

Abstract - Now a days the use of social media has grown exponentially with the simultaneous growth of the internet throughout the world and it is mostly attracted by the youth. However, the enhancement of social connectivity by using this social media platform may also lead to a negative impact on society mainly cybercrime, cyberbullying, abuse, etc. This cyberbullying is a major problem in this society which leads to frequent physical and mental stress, particularly for teenagers, transgenders, and women. Cyberbullying is done by using comments on social media platforms commenting vulgar words and also due to the increase of user's various malicious entities like increasing fake accounts increased a lot. Detection of cyberbullying and fake accounts on such large platforms is very difficult and may sometimes lead to false detection. So many incidents have recently occurred throughout the world like suicides, mental illness, etc. So our goal is to identify bullying comments and fake accounts on social media platforms like Facebook, Twitter, etc. By merging natural language processing and machine learning algorithms like xgboost(extreme gradient boosting) to detect fake accounts and logistic regression for cyberbullying.

***Key Words*:** Cyberbullying, Natural language processing, Machine learning, xgboost(extreme gradient boosting), logistic regression, Random forest, etc.

1.INTRODUCTION

Cyberbullying and fake account detection on social media are serious challenges that society faces. In addressing these issues, machine learning and natural language processing (NLP) play crucial roles. Their integration into social media platforms enables automated identification and mitigation of harmful activities. Machine learning algorithms analyze user interactions and content to detect patterns indicative of cyberbullying. Through continuous learning, these systems adapt to evolving forms of online harassment, enhancing their effectiveness over time. NLP techniques allow platforms to understand the context and sentiment behind users' messages, aiding in the identification of potentially harmful content. To combat fake accounts, machine learning models are designed to recognize unusual behavior patterns, such as excessive posting or engagement with specific types of content. NLP contributes by examining the language used in profiles and posts, helping to distinguish between genuine and deceptive accounts. By leveraging these technologies, social media platforms can proactively identify and address instances of cyberbullying and fake accounts, creating a safer online environment for users. Ongoing research and development in the field of machine learning and NLP are essential to stay ahead of emerging threats and to continually enhance the effectiveness of these protective measures.

2. LITERATURE SURVEY

1. **Cyberbullying on Social Media:**
Cyberbullying has become a pervasive issue in the realm of online communication, particularly on social media platforms. Researchers have extensively explored the various forms and manifestations of cyberbullying, ranging from direct harassment to more subtle forms of online aggression. Studies (Patchin & Hinduja, Kowalski) highlight the negative psychological impact on victims and emphasize the need for proactive measures to address this growing concern.

2. **Machine Learning for Cyberbullying Detection:**
The application of Machine Learning (ML) in the identification of cyberbullying has gained significant attention in recent years. Researchers have explored the effectiveness of supervised learning models, including Support Vector Machines (SVM) and deep learning algorithms, in automatically detecting cyberbullying content (Chatzakou, Mishra). These studies emphasize the importance of feature engineering and dataset diversity in improving the performance of ML-based cyberbullying detection systems.

3. **Natural Language Processing in Cyberbullying Detection:**
Natural Language Processing (NLP) techniques play a crucial role in understanding the context and sentiment of textual content, which is essential for accurate cyberbullying detection. Sentiment analysis, semantic parsing, and contextual embedding models have been

explored to enhance the capabilities of NLP in identifying harmful language and malicious intent in online communication (Chen; Nobata).

4. **Fake Account Detection on Social Media:**
Fake accounts pose a serious threat to the integrity of social media platforms. Existing literature has delved into the characteristics and behaviors associated with fake accounts (Yang; Stringhini). Machine learning models, including anomaly detection algorithms and clustering techniques, have been employed to identify patterns indicative of fake account activities (Boshmaf, Ferrara).

5. **Combined Approaches: ML and NLP for Holistic Detection:**
Several studies have highlighted the effectiveness of combining ML and NLP approaches to create more robust and comprehensive systems for online content moderation. Integrating ML algorithms for pattern recognition with NLP techniques for context analysis has shown promising results in addressing the nuanced nature of cyberbullying and fake account detection (Santia; Fortuna).
The literature review underscores the urgency of addressing cyberbullying and fake account issues on social media. ML and NLP techniques offer promising avenues for developing advanced detection systems, but challenges related to bias and ethical considerations must be carefully navigated.

3. PROPOSED SYSTEM METHODOLOGY

In this paper, a system is proposed to detect cyberbullying and fake accounts on social media. The main difference with previous research is that we not only developed a machine learning model to detect cyberbullying content but also developed the detection of fake accounts. It is just like a two-in-one tool for both digital problems.

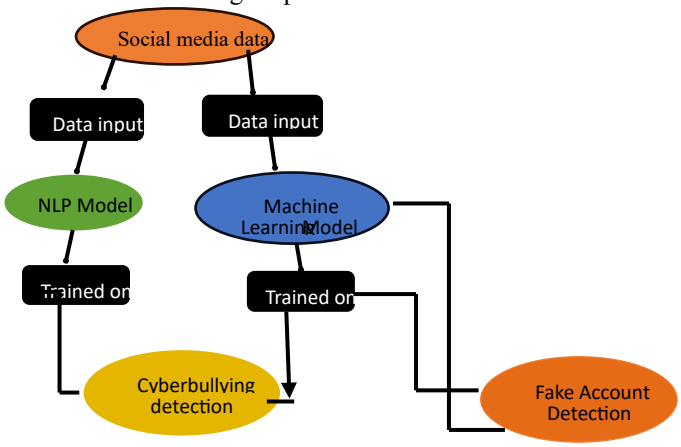

Fig.1 .Cyberbullying and Fake Account Detection Framework

3.1 First let us discuss the steps involved in cyberbullying detection:-

Steps:-

1. Import the latest tweets/comments from different types of social media platforms.
2. The Data Preprocessing and Data Extraction will be performed on the fetched Tweets
3. The pre-processed tweets will be fed into random forest and logistic regression model to calculate the probability of retrieved tweets to check whether a retrieved tweet constitutes bullying or not.
4. If the probability of a tweet being retrieved is between 0 and 0.5, then that tweet will not be considered a bullied tweet. If the probability of a tweet being retrieved is greater than 0.5, it is added to the database
5. So again, the list of users' timeline tweets will be passed to the random forest and logistic regression model to predict the results of the tweets.
6. Eventually the best algorithm is selected by calculating its accuracy.

In this exploration paper, we selected logistic regression with a maximum accuracy of 81%. We order cyberbullying into 4 types grounded on religion, age, gender, and ethnicity. So our model distinguishes the type of the comment/tweet and gives the final input.

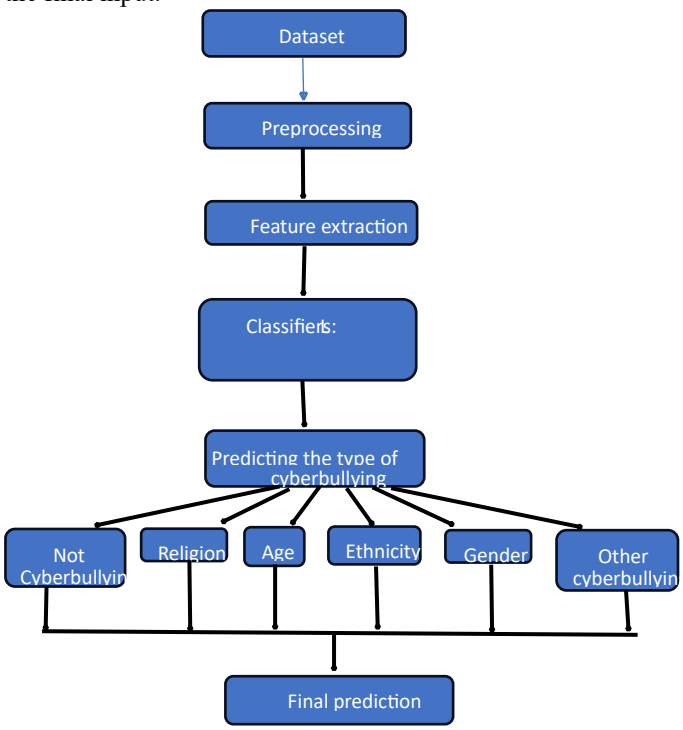

Fig.2 Proposed framework for cyberbullying detection

3.2 Now let us discuss the proposed framework for fake account detection.

Preprocessing:-

The resulting dataset contains raw data that needs pre-processing to obtain useful information. To remove irrelevant and redundant data from the raw data feature selection is done. It is the most important process in machine learning since it improves learning accuracy, reduces computation time, and makes the data more relevant to the learning model.

Feature selection:we have selected 10 features:-

S.no	Features	Description
1	Profile pic	An image that represents the user
2	User name	Name of the user
3	Number of followers	The number of users following this account currently has.
4	Number of following	The number of users following by this account
5	Description length	Length of the bio description given by the user
6	Type of account	Whether the account is in private mode or public mode
7	External URL	External URLs placed in the user bio
8	Users full name	Full name of the user along with the surname
9	No of posts and likes	Number of posts posted by the user and no of likes done by the user
10	Ration of user name and user full name	The ratio of the user name and the user's full name mentioned in the bio

Table 1. Description of Features.

After pre-processing and feature selection, the model is trained for the classification in which the dataset is separated into two distinct sets namely the training set and the test set. The training set is utilized to fit the parameters of the classifier. The test set is used to upgrade the architecture or parameters of the classifier. In the testing phase, the trained classifier classifies the profile as fake and legitimate.

It builds a strong predictive model by combining the predictions of multiple weak models, typically decision trees. By using this algorithm our model has secured 96% accuracy in detecting fake accounts.

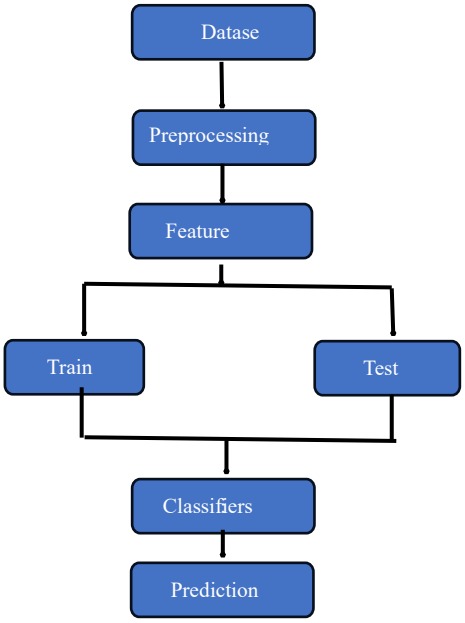

Fig.3 Proposed framework for fake account detection

4. METHODOLOGIES

Natural Language Processing:

Real-world messages or SMS contain many unnecessary characters or text. For example, numbers or punctuation are not relevant to detecting harassment. Before applying machine learning algorithms to comments, we need to clean them and prepare them for the detection phase. During this phase, various processing tasks include removing all extraneous characters such as stop words, punctuation, numbers, tokens, stems, etc.

TF-IDF:

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure that allows you to evaluate the relevance of a word to documents in a collection of documents. TF-IDF should give more importance to frequently occurring words as they are more useful for classification.

Machine Learning:

This module includes applying various machine learning approaches such as Random Forest and logistic regression to detect bullying messages and texts. The classifier with the highest accuracy is determined for a public cyberbullying dataset. Logistic regression secured a remarkable accuracy of 81% but random forest secured 78% only which is a little bit lower accuracy.

XGBoost(eXtreme Gradient Boosting):-

We have used nearly 576 user's account information and in our model, we used the XGBoost algorithm for the prediction of fake accounts. Whereas XGBoost is a powerful machine-learning algorithm due to its performance and effectiveness in

various machine learning competitions and real-world applications. XGBoost is based on the gradient boosting framework, which is a machine-learning technique for regression and classification problems.

5. RESULTS & EVALUATION

In this section, the Logistic regression and Random forest are tested on the dataset collected from various sources like Kaggle, Github, etc. After performing preprocessing and feature extraction on the dataset, for training and testing, and divided the dataset into ratios of 0.70 and 0.30 respectively. Both Logistic regression and Random forest are evaluated to calculate the accuracy, recall, f-score, and precision. Interestingly logistic regression outperformed Random forest in every aspect.

Result for cyberbullying detection:-

Table.2 Represents the accuracy of the Random forest and logistic regression

S.No	Classifiers	Accuracy(in %)
1	Random forest	78%
2	Logistic regression	81%

From the above table, we can conclude that the logistic regression algorithm gives higher accuracy than random forest.

Table.3 Represents the precision, Recall, and F1-Score generated by the types of cyberbullying using Logistic regression.

S.NO	Types of cyberbullying (Logistic regression)	Precision	Recall	F1-Score
1	Age	96%	97%	96%
2	Ethnicity	98%	96%	97%
3	Gender	89%	81%	85%
4	Religion	95%	93%	94%
5	Other cyberbullying	58%	66%	62%
6	Not cyberbullying	56%	55%	56%

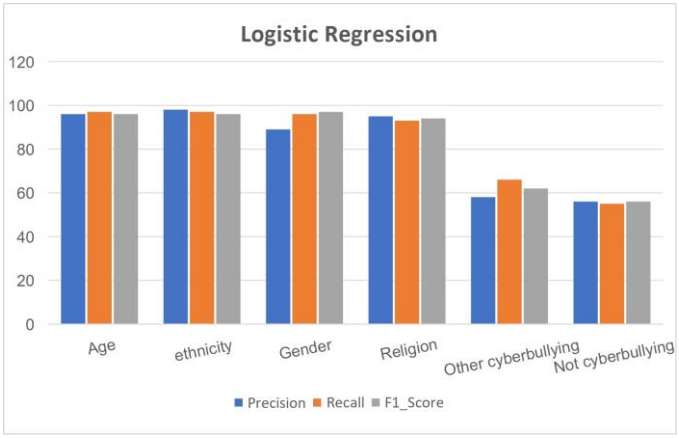

Fig.4 Bar graph for the table.3

Table.4 Represents the precision, Recall, and, F1-Score generated by the types of cyberbullying using Random forest.

S.NO	Types of cyberbullying (Random forest	Precision	Recall	F1-Score
1	Age	96%	98%	97%
2	Ethnicity	98%	97%	97%
3	Gender	88%	79%	83%
4	Religion	95%	94%	94%
5	Other cyberbullying	50%	57%	53%
6	Not cyberbullying	49%	48%	49%

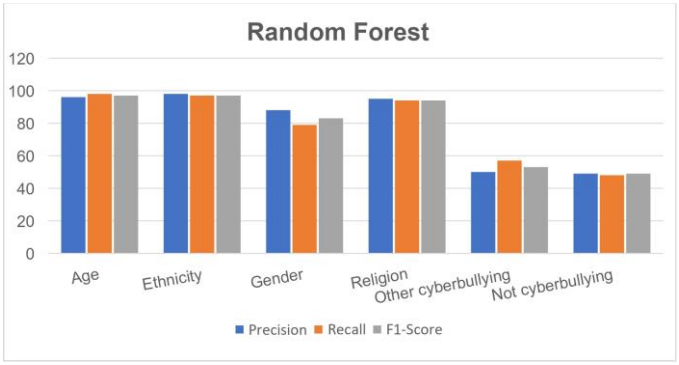

Fig.5 Bar graph for the table.4

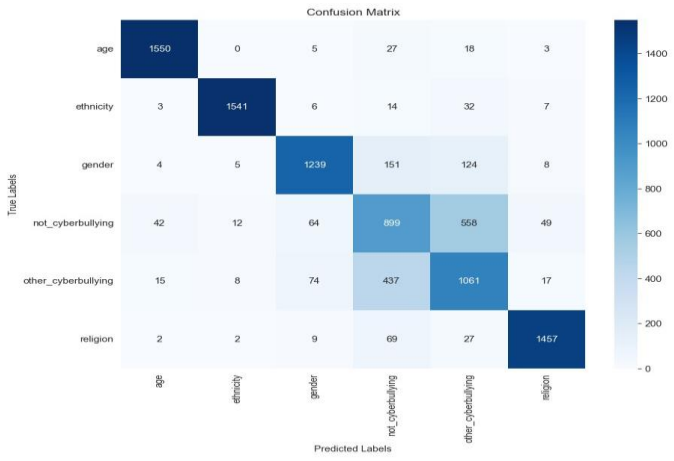

Fig.6 Confusion matrix for the above data

In this xgboost is tested on the collected dataset from various sources like Kaggle, Github, etc for the detection of fake accounts. After performing preprocessing and feature extraction on the dataset, for training and testing, and divided the dataset into ratios of 0.80 and 0.20 respectively. XGBoost is evaluated to calculate the accuracy, recall, f-score, and precision.

Result of fake account detection:-

Table.5 Represents the precision, recall, F1_Score, and accuracy of the xgboost algorithm.

S.NO	Classifier	Precision	Recall	F1-Score	Accuracy (in %)
1	XGBoost(eXtreme Gradient Boosting)	92%	93%	96 %	96%

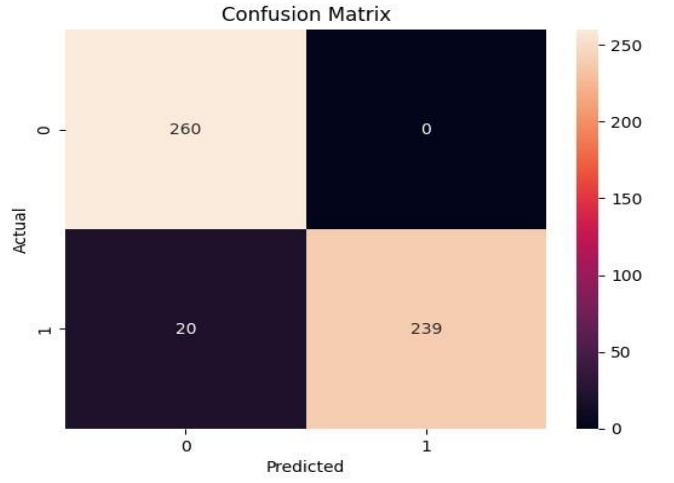

Fig.7 Confusion matrix for the above data.

6. CONCLUSION

Our study aimed to explore cyberbullying and fake account detection using machine learning and (NLP)Natural Language processing. As per our observation, nobody mentioned different types of cyberbullying in the previous works. Hence this study aimed to include that aspect as well. The results depicted that logistic regression performed better than Random forest with 81% average accuracy and Random forest with 78%.In the case of fake account detection, our model retained 96% average accuracy. Surely this result can still be improved better by applying other methods.

7. FUTURE SCOPE

Our model is a text-based model and only works in the English language. Implementing multilingual support would enhance the project's reach and effectiveness across diverse online communities. Enhance the fake account detection model by incorporating advanced behavioural pattern recognition. This could involve analysing user engagement patterns, posting frequency, and interaction styles to identify more sophisticated fake accounts. It is better to use deep learning-based models which outperform the previous traditional models.

8. REFERENCES

- Al-Garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior, 63, 433-443.
- Balakrishnan, V., & Sriram, S. (2020). Fake news detection using machine learning techniques: A systematic literature review. CAAI Transactions on Intelligence Technology, 5(3), 161-172.
- Chandrasekaran, M., & Prabha, S. (2019). Cyberbullying detection using machine learning techniques. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(6), 563-56
- Fortuna, P., Nunes, S., & Rodrigues, F. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4), 1-30.
- Gao, H., & Huang, Y. (2019). A machine learning approach to identifying cyberbullying on Twitter. Telematics and Informatics, 46, 101267
- Golbeck, J., & Hansen, D. L. (2017). Computing and applying trust in Web-based social networks. Synthesis Lectures on Data Management, 9(3), 1-134.
- Khan, L., Siddique, A., & Ahmad, F. (2020). Deep learning-based cyberbullying detection system using semantic and sentiment analysis. IEEE Access, 8, 144165144178

- Kumar, P., & Vyas, O. P. (2019). Cyberbullying detection in social media using machine learning techniques. International Journal of Computer Applications, 183(5), 40-44.
- Mishra, N., & Bandyopadhyay, S. (2018). Cyberbullying detection using machine learning: A review. ACM Computing Surveys (CSUR), 51(6), 1-36.
- Pfeffer, J., Zorbach, T., & Carley, K. M. (2018). Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. Journal of Marketing Communications, 20(1-2), 117-128.
- Qian, J., Wu, Y., Liu, J., & Fang, B. (2019). Cyberbullying detection on social media using machine learning algorithms. IEEE Access, 7, 121914-121922.
- Saha, K., & Mahmood, A. N. (2020). A survey on machine learning techniques for fake news detection. Journal of Intelligent & Fuzzy Systems, 39(4), 5217-5232.
- Teng, Z., Zhang, Y., & Park, J. (2017). Cyberbullying detection on Instagram using machine learning techniques. Telematics and Informatics, 34(7), 1141-1152.
- Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648.
- Wijeratne, S., Balasuriya, L., Sheth, A., Doran, D., & Welton, R. (2018). UWM at SemEval-2018 Task 5: A combined approach for emoji-based irony detection. In Proceedings of The 12th International Workshop on Semantic Evaluation (SemEval2018) (pp. 545-550).