Unmasking Anti-social comments from social media platform using Natural Language Processing and Machine Learning

Zihadul Karim Xenon
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Raisa Rahman Rodela
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Jimmati Arbi
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Md. Farhadul Islam
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.mustakin.alam@g.bracu.ac.bd

Md Humaion Kabir Mehedi
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

Annajiat Alim Rasel
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—Online comments that breach social standards and hurt people's feelings or communities are known as "antisocial comments." The explosion of social media and online communication platforms has given individuals a medium for free expression. However, it has also given birth to antisocial speech in the form of online harassment and cyberbullying. Although much work has been done to detect hate or abusive speech on social networking websites in English, research on the Bangla language still needs to be done due to the lack of standard-labeled datasets for precise and adequate Natural Language Processing in the Bangla language. In this work, we investigated many data-driven strategies for early identification and control of antisocial behavior through NLP and Machine Learning. This research shows that antisocial behavior may be detected using similar methods on social media sites like Twitter. The research continues by emphasizing the difficulty in training models to identify and discriminate between personality and behavior disorders that have diagnostic overlap with antisocial behavior and the consequent need to expand our understanding of these conditions. We effectively identified antisocial comments from the dataset using machine learning methods, including Logistic Regression, Support Vector Machine, Naive Bayes, Random Forest, and XGBoost. Additionally, we compute scores for completeness and sufficiency to evaluate the authenticity of explanations. Our findings show an accuracy of 72% for XGBoost, which outperformed the other algorithms.**

***Index Terms*—Anti Social Comments, Social Media Comments, NLP in Bangla Language, ML.**

I. Introduction

People have become heavily dependent on social media platforms like Facebook, Twitter, YouTube, Reedit, etc., to interact, exchange information, and share their sentiments in the present generation. However, the prevalence of antisocial comments has also raised a severe issue on social media platforms. Online activities that flout social standards and harm others or society are called antisocial behavior in social media comments. Various antisocial comments predominate social media which include trolling, harassment, hate speech, and cyberbullies. In addition to harming people's feelings, these comments escalate the risk of developing into more severe occurrences like online mobs, harassing, and other types of cybercrime. Therefore, identifying and responding to antisocial comments is essential to preserving a positive and secure online environment.

Since hate speech, cyberbullies, and other types of antisocial behavior are not limited to any particular language or culture, creating models to recognize and categorize such remarks in various languages is essential. Therefore, generating a model to identify antisocial Bengali comments on social media platforms is crucial for enabling healthy and respectful online communication in Bengali-speaking communities. Due to the exquisiteness and complications of the Bengali language, recognizing antisocial behavior poses particular tribulations. Detecting offensive Bengali speech in social media is influential in visualizing various abusive Bengali texts, comments or speech.

Machine learning (ML) algorithms have gained more and more traction in recent years to identify antisocial comments on social networking sites. These algorithms scan massive volumes of text data and look for patterns that point to

the presence of antisocial comments using Natural Language Processing (NLP) techniques. However, it is challenging to identify antisocial statements due to the intricacy of natural language and the exhaustive range of settings in which these comments are made.

The effectiveness of several machine-learning algorithms in identifying disruptive comments on social media sites is examined in this research. The paper will specifically look at topic modelling, sentiment analysis, and ML methods, including supervised and unsupervised learning algorithms. The advanced ML algorithms are operated in the research, including Support Vector Machines (SVMs), Naive Bayes, and Random Forests. The paper also concentrates on the interoperability of the model. These techniques enable interpretive insights and the decision-making process, which can help detect the most significant features contributing to the decisions of the models. The study will also provide the degree of severity of the comments to assist individuals in determining toxicity categories.

II. Previous Works

The negative comments aimed at people and organizations hosted on social media platforms are alarming and threaten to the victim's mental and physical well-being [7]. Traditional methods of combating online hatred concentrate on the characteristics of offenders and their perspectives on the individual they target.

The research worked by Rashid (2022) describes the development of a new dataset of toxic language in Bengali collected from user-generated Facebook material [9]. The dataset is regarded as an extensive, curated, value-added, and evaluated dataset that may be used as classifier material to detect offensive comments in social media. The dataset is helpful because it sheds light on Bengali users' harmful language on social media, especially in Bangladesh and how it reflects racial, gender, and sexist sentiments. Using bigrams as toxic tokens and thematic categories for classification, machine learning and NLP practitioners may utilize the dataset to detect hate speech. The collection of poisonous language used on Facebook by Bengali users includes 1959 distinct bigrams that have been painstakingly gathered and annotated. Each bigram has been analyzed to include details like its transcription in IPA, peculiar spelling, and level of toxicity. The dataset is divided into eight categories according to themes, such as" Misogynist bully," Sexist and Patriarchal bully," Vulgar, incivility sarcasm," Political hate words," Religion/communal hate words," Racism on Body, Gender Color," and" Moral Policing Sewer, Name trolling." The most bigrams and occurrences are found in the" Sexist Patriarchal Bully" class, while the least is found in the" Racism on Body, Gender Color" class. Also, the dataset contains data about the degree.

Another research worked by Sarker et al. (2022) presents a ML-based process to categorize antisocial Bengali comments on Facebook and YouTube [10]. The study provides examples of antisocial comments in the paper, such as comments that retain demeaning language, personal attacks, and ambushing content. A dataset containing 2000 comments was constructed. They utilize various natural language processing techniques to preprocess the data, such as tokenization, stop-word removal, and stemming. They then trained different machine learning models on the preprocessed and engineered data, including logistic regression (LR), random forest (RF), Multinomial Naive Bayes (MNB), support vector machine (SVM), and GRU. The research results show that the MNB and GRU models perform the best among all the models, respectively. The study identified the critical features that the machine learning model used to classify antisocial comments, such as the presence of profanity, sarcasm, and negative sentiment.

A detailed study on personality disorder shows that antisocial behavior is prevalent offline and online and is a widespread problem twitter, a popular social media platform [11]. According to the study, online platforms require a mechanism that can automatically identify and stop antisocial behavior without impacting other users. The authors provided a methodology for identifying cyberbullying, inflammatory language, and hate speech on Twitter using natural language processing techniques. The detecting early antisocial behavior proposal involves collecting and labelling tweets, verifying the labels by a qualified person, using natural language processing techniques to clean and preprocess the data, and training a machine learning model. They describe the process of vectorizing (Word Frequency and TF-IDF) text data to create a machine-learning model to analyze antisocial behavior in tweets. The dataset was divided into subsets using K-fold Cross Validation to train and validate the machine learning model. Five machine learning algorithms were implemented with both vectorization methods: Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, and Naïve Bayes. The results show high accuracy and precision for all five algorithms. The research proposes a data-driven approach using natural language processing and machine learning techniques to detect and prevent antisocial behavior on online platforms, particularly Twitter. However, the current measures are not effective enough, and the proposed framework can proactively detect and restrict such behavior with high accuracy. The model can be used into an online solution to take appropriate action, such as delete and block users. The diagnostic criteria for various conditions can make it difficult to effectively train a model to categorize and differentiate them, though.

Another study represents that deep learning model could accurately diagnose depression, anxiety, bipolar, borderline personality disorder, schizophrenia, and autism by evaluating and learning user posting information [6]. They utilized Python's NLTK to tokenize users' posts and filter common words. They created six separate binary classification models for each symptom to increase performance. They accurately

identified a user's mental condition by constructing six models for each mental disorder, each taking data from users with only one mental disorder. They used synthetically to address the class imbalance in the gathered data. Their dataset had training (80%) and testing (20%) sets. XBoost and CNN followed. They speculate that the unbalanced data problem can be solved if additional data is gathered, leading to improved results.

Chanda and his team members used 1709 depressive Twitter comments as their dataset for the research where four classifiers: SVM, KNN, Decision Tree, and Random Forest were used [2]. This method uses natural language processing (NLP) to evaluate tweets regarding linguistic processing and mental state. LIWC psycholinguistic lexicon collection was used. If the particular comment has a threshold number of depressive words, they perform four classifiers to predict the outcomes. It is concluded that SVM outperforms other classifiers with 71% accuracy (2022).

In new research done by Duwairi and Halloush (2022), a multi-view fusion model that uses deep learning algorithms was proposed to identify common Personality Disorder (PD) from social media posts in a professional-driven manner utilizing descriptions from the DSM-5 [4]. The research was done on the Arab dataset model comprised of 8000 textual tweets and 8000 images describing the mental states of 150 users. Using image detection, they first detected the images that represent PD. Also, they detected the expressive posts, and then, after analyzing, they found out that those are the symptoms of two different personality disorders, which are Schizotypal and Bipolar Disorder (BPD).

From the study of Bae et al. (2021), we learned that ML techniques could evaluate text carrying the keywords that indicate schizophrenic behavior of social media postings made by people with schizophrenia [1]. To identify the themes that reflect the main symptoms of schizophrenia, such as hallucinations, delusions, and negative symptoms, they used unsupervised LDA clustering. Based on topic distributions and LIWC characteristics, classifying the schizophrenia and non-schizophrenia groups was successful, with the highest accuracy of 96%. Four different algorithms, Logistic Regression, SVM, Random Forest and Naive Bayes, were implemented. However, the themes addressed in online schizophrenia forums, as well as the language characteristics linked with those who have schizophrenia, are not perfectly accurate.

Kabir et al. (2023) detected depression severity from Bengali social media texts using natural language processing techniques [5]. They took Bengali text-based data from blogs and open posts to develop a process for annotated corpus development and textual information extraction from Bengali literature for predictive modelling. To evaluate the severity of depression from texts, the author applied machine learning and deep learning models. They scraped the data from social media using Selenium to diagnose accurately used DSM-5. They used models like the random forest, SVM, logistic regression, k-nearest neighbor, and naive Bayes for preprocessing and data modelling. According to the authors, the recurrent neural network model accurately detected the severity compared to other models.

In another research, depressed users were identified through the tweets they shared on Twitter [12]. In order to do that, they had to fetch the users' data from the posts using various keywords which indicate depression and aggressive behavior in social media. For the sentiment calculation, three lexicons were used AFINN, BING, and NRC gave the scores which were later normalized from -1 to +1 and made an average to get the final score. Moreover, for the weighted calculation, there was an assigned weight for each of the eight emotions, which used to differ according to the level of emotions and give the final level of depression in particular tweets.

Priya et al. (2020) discusses predictions of anxiety, depression, and stress made using a machine learning algorithm for the research data collected using by DASS-21 questionnaire and analyzing texts from social media [8]. The used CNN algorithm which gave approximately 79% accuracy, whereas the SVM algorithm was just 58%.

In another research, a unique method was developed for detecting users who have or are at risk of developing depression through using assessments of eight fundamental emotions as characteristics from Twitter tweets [3]. Using Ekman's core emotion model, they used the emotional system to measure emotions. The data set was separated into temporal and non-temporal feature sets for differentiation. Mathematical, statistical methods like mean, standard deviation, entropy, mean momentum and mean differencing were used. They compared the results of the proposed systems' temporal and non-temporal data sets (Chen et al., 2018).

Due to the lack of knowledge, negative comments are spreading on social media daily [13]. Proper knowledge and the machine learning approach can develop an outcome where the detection of antisocial comments by Bengali people will be established.

III. Dataset

Online social media and streaming site comments were gathered for the dataset used in this study. We created a list of contentious events that happened in Bangladesh after 2017 and excluded blogs and online news portals because of the low user activity in the comment sections in order to ensure thorough coverage of hate speech (HS) remarks. To get ideas for controversial topics and keywords for our HS dataset, we conducted a brief survey of 20 undergraduate students at Shahjalal University of Science and Technology

(12 males and 8 females), who frequently used online social media and video streaming platforms. The students represented 17 different majors.Participants were asked to suggest themes or keywords from a variety of fields, resulting in six major categories: sports, entertainment, crime, politics, religion, and controversial statements by scholars (CCS). This was done in order to maintain linguistic diversity.

Using an open source program called Facepager, we looked for publicly accessible content on social networking sites like Facebook and internet streaming services like Youtube based on these keywords. We also looked for videos related to roasting, Bangla TikTok, and other topics because we noticed that the comment areas on these videos frequently feature poisonous language. Scraped comments from these videos were labeled as "other." All commentator names, affiliations, and other identifying information were taken out of the dataset to ensure anonymity.

Over 100k comments were gathered in total, and to ensure a varied vocabulary, duplicate and extremely similar remarks were eliminated using a Jaccard Index cutoff of 0.8. We finally had a dataset of 50,281 comments for analysis after the annotation procedure. The dataset was tokenized using lexicons to handle stop words and preprocessed to eliminate unused symbols, emojis, and superfluous spaces. Three separate deep neural network (DNN) models, CNN, Bi-LSTM, and Conv-LSTM, were trained for hate speech identification using the preprocessed data. The test data was used, and results were forecast based on that.

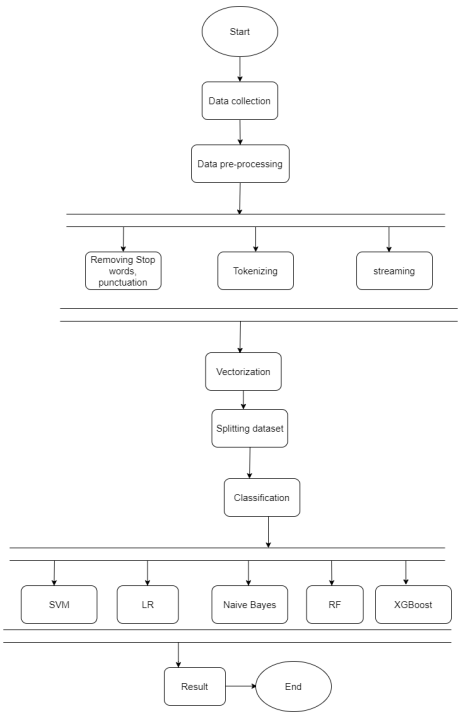

Fig. 1. Workflow Diagram

IV. Methodology

To identify anti-social comments from social networking platforms, we used some machine learning algorithms like : TF-IDF vectorizer, Support Vector Machines (SVM), Random Forrest,XG Boost, Naive Bayes and Logistic Regression. This algorithms were selected because of their ability to classify text based data in a quite success rate. Some of the algorithms are really good at classification and regression task.

[1] TF-IDF: TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic that recollects the implication of a word in a document or a corpus. TF-IDF is commonly used for text analysis and feature selection in natural language processing and machine learning. In our research on labeling antisocial comments on social media platforms, TF-IDF was used for selecting essential terms in the text data most relevant to the classification job.

[2] SVM : The supervised machine learning algorithm Support Vector Machines (SVM) is frequently employed for classification and regression tasks. It operates by identifying the most effective hyperplane for classifying the data. The hyperplane is a line that divides the two classes in a binary classification problem. The hyperplane is a multidimensional surface that separates the classes in a problem of multiclass classification. The SVM method develops a more reliable and accurate model by locating the hyperplane that optimizes the margin between the classes.

[3] Random Forrest : Several decision trees are combined using the ensemble learning algorithm Random Forest to get a more precise model. It functions by creating decision trees based on subsets of attributes that are randomly chosen. The program then combines all of the trees' predictions to create the final categorization. Random Forest's key benefit is its capacity to handle high-dimensional data with numerous attributes. Additionally, it lowers over fitting and offers a gauge for feature relevance.

[4] XG Boost : A popular gradient boosting technique for classification and regression applications is called XG Boost (Extreme Gradient Boosting). Weak learners are gradually added to the model, increasing the model's accuracy. Beginning with a straightforward model, the method gradually builds up the ensemble with more complex models. The performance of the model is enhanced by using a gradient descent approach to optimize the loss function. In terms of speed, scalability, and accuracy, XG Boost excels. It can also manage missing values and has built-in regularization.

[5] Naive Bayes : An efficient probabilistic approach for classifying texts is called Naive Bayes. It operates by computing the likelihood of each feature given the class and by computing the probability of each class given the features using Bayes' theorem. Because of this presumption, the

algorithm is referred to as "naive," which stands for "naive." Naive Bayes can handle big datasets with high-dimensional characteristics and is simple to implement. It is frequently used for document classification, sentiment analysis, and spam filtering.

[6] Logistic Regression : An efficient statistical method for binary classification tasks is logistic regression. In order to predict the probability of each class, it fits a logistic function to the data. The logistic function converts the input parameters into a range of output values between 0 and 1, which stands for the likelihood of the positive class. Both continuous and categorical features can be handled by logistic regression, which also offers interpretable coefficients that show the relative weight of each feature. It is frequently employed in marketing, credit scoring, and medical research.

V. Results and Discussion

To unmask Anti Social comments from social media platform, five different nlp and machine learning algorithm was implemented. All of the algorithms generated nearly similar form of result. Over 40,000 comments from social media was used to train the models in the hope of getting the best possible result.

From the tables, we can see that all the approaches give us and average accuracy of 70 percent. The results would be far better if the data set could be pre-processed more precisely. Since the data set is based on Bangla Language, we could not classify some hate/negetive words in the pre processing phase as there is not enough resources or library for that.

SVM and XGB gave 70 percent accuracy where RF,NB and LR gave us 71 percent accuracy. F1 score, Precision and Recall were .70 for SVM. For RF, it was .71. XGB has .70 f1 score and precsion where the recall was .69 . For NB and LR , F1 score is .70 and Precision and Recall is .71 and .70. With some more pre processing and more data this score can be increased.

Furthermore, the models were not optimized to train on bangla data set rather to train on large text based task. This is another reason of this reluts. Training on custom models optimized for this kind of dataset will increase the accuracry more.

TABLE I
Comparison of 5 different Approaches

Metric	**SVM**	**RF**	**XGB**
Accuracy	70%	71%	70%
F1 Score	0.70	0.71	0.70
Precision	0.70	0.71	0.70
Recall	0.70	0.71	0.69

Metric	**NB**	**LR**
Accuracy	71%	71%
F1 Score	0.70	0.70
Precision	0.71	0.70
Recall	0.70	0.71

VI. Conclusion and Future Work

The paper emphasized identifying and stopping antisocial behavior on social media sites. The detection and categorization of antisocial comments have shown encouraging results when using data-driven methods based on Neural Networks, NLP, and ML techniques. It is still difficult to precisely categorize certain forms of behavioral problems and guarantee that these screening and prevention approaches are working. The evaluations have also highlighted the need for implementing a specific library for antisocial Bengali keywords. Our purpose was to alert social media companies to guard against the spread of antisocial behavior while balancing the protection of free expression. In order to increase the precision and accuracy of the detection models and to investigate the usage of similar approaches on additional online platforms, further study is required. Overall, these assessments of academic papers provide helpful information about the status of research on identifying antisocial behavior on social media platforms and the promise of data-driven solutions to this problem. A safer and more helpful online community may result from the creation and use of efficient detection and prevention techniques. We will continue our research as the resources need more accuracy in the Bangla language, and we have just utilized the bottom liners. In the future, we want to combine machine learning and deep learning models to create a significant ensemble model that will help us improve our current findings.

References

[1] Yi Ji Bae, Midan Shim, and Won Hee Lee. Schizophrenia detection using machine learning approach from social media content. *Sensors*, 21(17):5924, 2021.

[2] KAUSHIK CHANDA. Evaluating mental health issues as collected from social media by machine learning techniques. 2022.

[3] Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion proceedings of the the web conference 2018*, pages 1653–1660, 2018.

[4] Rehab Duwairi and Zain Halloush. A multi-view learning approach for detecting personality disorders among arab social media users. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–19, 2023.

[5] Muhammad Khubayeeb Kabir, Maisha Islam, Anika Nahian Binte Kabir, Adiba Haque, and Md Khalilur Rhaman. Detection of depression severity using bengali social media posts on mental health: Study using natural language processing techniques. *JMIR Formative Research*, 6(9):e36118, 2022.

[6] Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, 10(1):1–6, 2020.

[7] WB Lian, SKY Ho, CL Yeo, and LY Ho. General practitioners' knowledge on childhood developmental and behavioural disorders. *Singapore Medical Journal*, 44(8):397–403, 2003.

[8] Anu Priya, Shruti Garg, and Neha Prerna Tigga. Predicting anxiety, depression and stress in modern life using machine learning algorithms. *Procedia Computer Science*, 167:1258–1267, 2020.

[9] Mohammad Mamun Or Rashid. Toxlex_bn: A curated dataset of bangla toxic language derived from facebook comment. *Data in Brief*, 43:108416, 2022.

[10] Manash Sarker, Md Forhad Hossain, Fahmida Rahman Liza, Syed Nazmus Sakib, and Abdullah Al Farooq. A machine learning approach to classify anti-social bengali comments on social media. In *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, pages 1–6. IEEE, 2022.

[11] Ravinder Singh, Jiahua Du, Yanchun Zhang, Hua Wang, Yuan Miao, Omid Ameri Sianaki, and Anwaar Ulhaq. A framework for early detection of antisocial behavior on twitter using natural language processing. In *Complex, Intelligent, and Software Intensive Systems: Proceedings of the 13th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS-2019)*, pages 484–495. Springer, 2020.

[12] Jini Jojo Stephen and P Prabu. Detecting the magnitude of depression in twitter users using sentiment analysis. *International Journal of Electrical and Computer Engineering*, 9(4):3247, 2019.

[13] Tie Hua Zhou, Gong Liang Hu, and Ling Wang. Psychological disorder identifying method based on emotion perception over social networks. *International journal of environmental research and public health*, 16(6):953, 2019.