An Unsupervised Machine Learning Approach to Unfold Online Toxicity Pattern: A Look at the Bangladeshi Community

Ayesha Hossain

Student ID: 1708022

Mohammad Mustasin Moin

Student ID: 1708042

Fahim Faisal Raunaq

Student ID: 1608009

Department of Industrial and Production Engineering

Bangladesh University of Engineering and Technology (BUET)

Dhaka-1000, Bangladesh

May, 2023

Certificate of Approval

The thesis titled ”An Unsupervised Machine Learning Approach to Unfold Online Toxicity Pattern: A Look at The Bangladeshi Community” submitted by Ayesha Hossain, Student ID: 1708022, Mohammad Mustasin Moin, Student ID: 1708042, Fahim Faisal Raunaq, Student ID: 1608009, has been accepted as satisfactory in fulfillment of the requirements for the degree of Bachelor of Science in Industrial & Production Engineering on May 27, 2023.

Dr. Nafisa Mahbub

Assistant Professor

Department of Industrial and Production Engineering

Bangladesh University of Engineering and Technology (BUET)

Acknowledgments

We would like to express our deepest gratitude to our supervisor Dr. Nafisa Mahbub for her guidance on this thesis, for showing us the path of conducting successful research, and above all, for always being our mentor. She shared her wisdom with us in analyzing subject matters and, at the same time, valued our thinking approach to synthesize those topics. Her suggestions drove us towards better ways of thinking, her reviews enriched us in solving problems, and her support strengthened us during our disappointment. We shall forever cherish the memories of working with her. We deeply thank our friends and families for always believing in us even when we were losing our confidence.

Declaration

The thesis is submitted as a six-credit course for a Bachelor of Science degree at Bangladesh University of Engineering & Technology, Dhaka, Bangladesh. The authors granted permission to state that the library can use it, and we support its use for legitimate library purposes.

Ayesha Hossain

Student ID: 1708022

Mohammad Mustasin Moin

Student ID: 1708042

Fahim Faisal Raunaq

Student ID: 1608009

Abstract

The rise of social media platforms has given rise to a significant issue of online toxicity, which involves disseminating harmful content, fostering division among communities, amplifying feelings of hopelessness, and disseminating self-harmful behavior. While considerable research has been done on automatically detecting and classifying online toxicity using data-driven techniques, there is a notable lack of studies focusing on low-resource languages like Bangla. Specifically, there is limited research on analyzing temporal and spatial patterns and understanding social network dynamics using unsupervised machine learning techniques. Furthermore, comprehensive datasets that encompass essential attributes such as timestamps, demographic information of commenters, and diverse reactions are lacking. This study aims to address these research gaps by investigating the variations of online toxicity over time and space, examining common reactions to toxic activities, and understanding the social dynamics associated with online toxicity. A dataset comprising 25,000 toxic comments extracted from popular Bangladeshi news channels on Facebook has been developed to accomplish this. The dataset includes annotations for six classes: Toxic, Insult, Profanity, Identity attack, Neutral, and Threat, utilizing manual and automatic techniques. Subsequently, the dataset is analyzed using unsupervised machine learning techniques, specifically K-Prototype clustering, to uncover temporal and spatial patterns of toxicity and explore social network dynamics. This research makes a valuable contribution to the existing body of knowledge by providing a comprehensive dataset that incorporates commenter metadata such as gender, location, reactions to each comment, and the time elapsed between the post and each comment. The findings of this study provide insightful observations regarding the prevalence and locations of online toxicity in Bangladesh, shed light on demographic characteristics associated with toxic behavior, unveil temporal dynamics, and analyze individual responses.

Keywords: Online toxicity; social media analytic; facebook comments; spatio-temporal patterns; unsupervised learning; network dynamics.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the rapid progression of Web 2.0, characterized by interactive and communicative aspects, digital platforms have become instrumental for individuals worldwide to share opinions, consume online news, access real-time updates, and even stream their favorite television channels. Over the past decade, various social media platforms, including Facebook, Twitter, Sina Weibo, and others, have gained immense popularity due to their ability to facilitate interactive two-way communication and information sharing. Key features such as sharing, commenting, and retweeting enable users to disseminate information rapidly across social networks and friend groups, spanning various topics such as politics, business, entertainment, and health. Especially during crisis events like natural disasters, terrorist attacks, political upheavals, and pandemics, these social media platforms play a crucial role in disseminating timely and accurate breaking news updates, safety guidelines, situational awareness, and critical health information to those affected (Zhu et al., 2011; Oh et al., 2013; Saroj and Pal, 2020; Li et al., 2021). They can serve as a vital source of information and help connect individuals to the resources they need in challenging circumstances. Despite the tremendous advantages social media platforms offer, it is essential to recognize the challenges they present. One of the key concerns is the unmoderated nature of these platforms, which often leads to the diffusion of online toxicity (Castaño-Pulgarín et al., 2021). Online toxicity refers to the presence and spread of harmful, offensive, and aggressive behavior exhibited by individuals or groups in online spaces such as social media platforms, forums, comment sections, or online gaming communities. It involves profane language, personal attacks, threats, hate speech, discrimination, or the dissemination of false information to cause harm, provoke negative reactions, or incite conflict.

Online toxicity knows no boundaries and has global consequences that impact individuals, communities, and society (Ayo et al., 2020; Chetty and Alathur, 2018). Bangladesh is not immune to this issue with its substantial online presence and high internet penetration. Within the Bengali community, online toxicity has far-reaching effects, causing mental health issues, such as anxiety, depression, and a diminished sense of self-worth. During the COVID-19 pandemic, the Bengali community, like many others, faced the challenge of online toxicity exacerbating the situation. Online toxicity fueled stigmatization, discrimination, and xenophobia, further deepening social divisions. It also fuels division and polarization, creating hostility and animosity between groups. The spread of false information further exacerbates the problem, eroding trust and contributing to social and political tensions. One distressing example of online toxicity in Bangladesh is the prevalence of threats and identity attacks on social media platforms. Public figures and minority groups often face vicious attacks and threats, which harm their well-being and hinder constructive discussions and social cohesion (Rezvi and Hossain, 2021). Using derogatory language and personal attacks during political debates further exacerbates tensions and prevents meaningful dialogue. Religious and ethnic minorities are also targeted by online toxicity, leading to discrimination, fear, and the perpetuation of intolerance. One notable example is the proliferation of hate crimes against Muslim communities in the United Kingdom resulting from Woolwich attack in May 2013 (Awan, 2014). Such toxic behavior marginalizes these communities and hampers social unity and understanding efforts. This dissemination of hate speech eroded trust, contributed to social and political tensions, fostered self-harming behavior, and hindered progress.

There is a clear and urgent need to establish a scalable and automated framework that can uncover patterns of online toxicity to foster a positive and respectful online environment. This framework should protect individuals from harm, promote well-being, and encourage productivity. Machine learning techniques provide viable solutions to tackle online toxicity, and researchers have focused on studying this issue across various social media platforms and microblogging sites (Ayo et al., 2020; Mullah and Zainon, 2021). For example, Boishakhi et al. (2021) and William et al. (2022) have explored the application of machine learning approaches in studying online toxicity. They have proposed automatic systems for hate speech detection using machine learning methods.

While significant research has been conducted on automatically detecting and classifying online toxicity using data-driven techniques, there is a noticeable lack of studies specifically addressing low-resource languages like Bangla. However, it is essential to recognize that the impact of toxicity caused by hate speech extends beyond language barriers. Recently, several researchers have employed machine learning and deep learning techniques to address the online Bangla hate speech issue (Karim et al., 2021; Aporna et al., 2022; Karim et al., 2023). However, most existing research primarily focuses on detecting and classifying online toxicity, leaving other aspects, such as temporal and spatial pattern analysis, social network dynamics, etc., largely unexplored. Additionally, there is a scarcity of comprehensive datasets that include vital attributes like timestamps, demographic information of commenters, and diverse reactions. Motivated by the current situation, this study intends to fill the research gap by investigating the following research questions (RQs):

RQ1: How does online toxicity vary over time and space in the context of the Bangladeshi community?

RQ2: What are individuals' reactions to online toxic activities?

RQ3: What social dynamics involve online toxicity within the Bangladeshi community?

RQ4: How can a predictive model be developed to automatically uncover patterns of online toxicity of social media users?

To address these RQs, the study involves the creation of a comprehensive dataset comprising around 25,000 toxic comments from popular Bangladeshi news channels and newspapers on Facebook. This dataset includes attributes such as toxic comments, commenter reactions, demographic information, and comment timestamps. The level of toxicity is assessed using Human Coders and the pre-trined model, BanglaBERT (Bidirectional Encoder Representations from Transformers). K-Prototype clustering, an unsupervised machine learning algorithm, is applied to determine the optimal number of clusters. This clustering approach enables the exploration of temporal and spatial toxicity patterns and analyzes social network dynamics within the Bangladeshi community.

The subsequent sections of this study are structured as follows: Section 2 offers an extensive review of existing literature, examining prior research in the field and research gaps with the study's contribution. Section 3 outlines the data collection, processing, and analysis procedures, outlining the methodology used to create the comprehensive dataset tailored to the Bengali community. This section also encompasses the utilization of K-Prototype clustering to determine the optimal number of clusters. Section 4 presents the findings and subsequent discussions. Lastly, Section 5 offers a comprehensive summary of the study, highlighting its limitations and suggesting potential avenues for future research.

Chapter 2

Literature Review

This section aims to explore the present level of research on applying data-driven techniques in detecting online toxicity with a specific focus on the Bangladeshi community. Detailed research gaps and the study's contribution are also discussed.

2.1 Related Works

Online toxicity has become a major concern in the digital era, with offensive and harmful content spreading across various platforms. To address this issue, it is crucial to develop scalable and automated techniques to identify and assess such toxic behavior. Data-driven approaches such as machine learning or deep learning algorithms have emerged as promising approaches for automatically detecting and classifying online toxicity. However, while significant research has been conducted on toxicity detection in high-resource languages across different social media platforms, there is limited research focusing on low-resource languages like Bangla, particularly within Bangladeshi online communities.

In recent years, researchers have started paying attention to online hate speech detection and classification within the Bangladeshi community, utilizing machine learning algorithms (Emon et al., 2019; Chakraborty and Seddiqui, 2019; Das et al., 2021). For example, Islam et al. (2022) collected 3006 pure Bengali comments from social media pages and employed different machine learning classifiers to detect whether the comment is abusive or non-abusive. Ahammed et al. (2019) collected data from Facebook and labeled comments by two classes: Hate speech or not. They applied different machine learning classifiers for Bangla hate speech classification. Deep learning approaches have also been employed to accurately identify toxic language, hate speech, and other online toxicity prevalent in the Bangladeshi context (Emon et al., 2019; Chakraborty and Seddiqui, 2019; Das et al., 2021). For instance, Islam et al. (2022) collected 3006 pure Bengali comments from social media pages and employed different machine learning classifiers to detect whether the comment is abusive or non-abusive. Ahammed et al. (2019) collected data from Facebook and labeled comments by two classes: Hate speech or not. They applied different machine learning classifiers for Bangla hate speech classification.

Integrating machine learning and deep learning techniques has shown promising results in hate speech detection. For instance, Ghosh et al. (2022) employed a hybrid deep learning approach on the publicly available dataset and achieved satisfactory accuracy compared to previous works on the same dataset. Integration of machine learning and deep learning techniques for hate speech detection is also prevalent in Bangladesh. Banik and Rahman (2019) investigated the toxicity of the Bengali language in social media from an annotated publicly available using supervised machine and deep-learning models. Their findings highlighted the outperformance of the deep learning model over other classifiers. Hossain Junaid et al. (2021) applied both machine learning and deep learning models to classify Bangla hate speech detection in videos from Youtube.

Researchers have explored the use of annotated datasets comprising Bangla text from different online platforms, such as social media, online news portals, and community forums. These datasets have been instrumental in training and evaluating machine learning and deep learning models for hateful speech detection. Various studies have developed benchmark datasets by collecting comments from platforms like YouTube and Facebook, annotating them into categories, and employing different classification algorithms. For example, Ishmam and Sharmin (2019) developed the first corpus of hateful speech detection in the Bengali language, extracting 5,126 comments from social media.

They annotated the comments into six categories: Hate Speech, Communal Attack, Inciteful, Religious Hatred, Political Comments, and Religious Comments, and applied different machine learning and deep learning algorithms for hateful speech detection. Another initial work is (Awal et al., 2018), where they extracted 2665 Bangla comments from Youtube and applied Naive Bayes classifier to automatically detect abusive comments. Romim et al. (2021) developed a gold standard labeled dataset with 30,000 user comments collected from YouTube and Facebook. All comments were classified into seven categories: sports, entertainment, religion, politics, crime, celebrity, TikTok, and meme, and labeled by 50 annotators. The study's contribution is the development of a benchmark-labeled dataset. Sarker et al. (2022) developed a dataset of 2000 Bangla comments from two prominent platforms in Bangladesh, Facebook, and YouTube. They employed different machine learning classifiers and an artificial neural network model to detect whether the comments are social or anti-social quickly and efficiently. These datasets have contributed significantly to the literature and provided valuable resources for studying and combating online toxicity in the Bangladeshi context.

Recently, Romim et al. (2022) has developed a large Bangla hate speech dataset with 50,281 comments. To our knowledge, it is the most extensive publicly available repository for studying hate speech in Bangla. This dataset has made notable contributions by providing a substantial volume of data, identifying hate speech targets, and categorizing different types of hate speech.

2.2 Research Gaps, and Contributions

While research in detecting and addressing online toxicity in the Bangladeshi context is still in its early stages, limited studies have focused on leveraging machine learning and deep learning techniques and developing annotated datasets of harmful content (Romim et al., 2021; Ishmam and Sharmin, 2019). However, there is a significant absence of a comprehensive dataset incorporating essential attributes such as comment timestamps, demographic information of commenters, and traits of reactions within the Bengali community. Additionally, the unfolding of temporal and spatial patterns, social dynamic analysis, and analysis of different reactions exhibited by individuals in each comment using an unsupervised machine learning model have not been explored yet in the Bangladeshi community. To fill the research gap, this study intends to unfold the online toxicity in Bengali hate speech using the K-Prototype algorithm.

While research on detecting and addressing online toxicity in the Bangladeshi context is still in its early stages, there is a progressive focus on leveraging machine learning and deep learning techniques and developing comprehensive datasets for harmful content (Romim et al., 2021; Ishmam and Sharmin, 2019). Previous studies have focused on detecting and classifying online toxicity, leaving other aspects, such as temporal and spatial pattern analysis, social network dynamics, etc., and employing unsupervised learning techniques, largely unexplored. Additionally, there is a scarcity of comprehensive datasets that include vital attributes like timestamps, demographic information of commenters, and diverse reactions. To address these gaps, this study aims to develop a comprehensive repository including different metadata and uncover online toxicity in Bengali hate speech using the K-Prototype algorithm.

The proposed framework has several significant contributions. Firstly, it focuses on creating a comprehensive database that collects and categorizes toxic comments on Bangla Facebook, including information about when and where they occur. This fills a gap in the current research and provides valuable insights into the prevalence and locations of online toxicity in Bangladesh. These insights can guide targeted interventions and educational initiatives to address the issue effectively.

Secondly, the framework investigates how demographic characteristics relate to engagement in toxic commenting behavior. Understanding which demographic groups are more likely to engage in online toxicity makes it possible to develop targeted prevention and intervention strategies that cater to the specific needs of different population segments in Bangladesh.

Moreover, analyzing the temporal dynamics of online toxicity can offer valuable insights into how and when such behavior escalates or fluctuates over time. This analysis considers specific events, cultural and social influences, and external factors contributing to online toxicity. Identifying these underlying dynamics makes it possible to develop a deeper understanding of the problem and implement more effective interventions.

Thirdly, studying individuals' responses to online toxicity can provide insights into the psychological impact of harmful content and facilitate the development of targeted support systems and interventions. This research is particularly important for the Bangladeshi community as it uncovers culturally specific coping mechanisms and helps design educational programs that promote digital resilience and responsible online behavior.

Lastly, analyzing the social network can identify influential users who spread negative comments, enabling the development of strategies to disrupt the cycle of harmful content and create a healthier online environment. By understanding the dynamics of how toxic behavior spreads through social networks, it becomes possible to implement interventions that target these influential users and promote positive online interactions.

Chapter 3

Research Methodology

The framework for investigating online toxicity patterns in the Bangladeshi community adopts a data-driven approach and consists of five stages. The first stage involves collecting data by manually gathering Facebook comments from various online news channels and newspapers in Bangladesh. The collected data includes attributes such as comments, user engagement metrics (such as different reactions), demographic information, and comment timestamps. In the second stage, relevant features for analysis are selected through necessary feature selection. This involves identifying and choosing features essential for understanding online toxicity patterns, such as the types of toxicity present, the sentiment of reactions to comments, and the demographic information of the commenters. The third stage focuses on data preprocessing techniques. These techniques are applied to the collected dataset to ensure data quality and integrity. Tasks such as data cleaning, handling missing values, and removing duplicate comments are performed to prepare the dataset for analysis. In the fourth stage, data analysis is conducted using descriptive techniques. This involves visualizing and analyzing the dataset to identify patterns, trends, and relationships between the selected features. By examining the data, researchers can gain insights into the characteristics and dynamics of online toxicity in the Bangladeshi community. Lastly, an unsupervised machine learning technique called K-Prototype clustering is employed in the fifth stage. This technique clusters commenters based on their characteristics, such as demographic information and toxic behavior patterns. By applying this clustering method, the study aims to identify distinct groups of commenters who exhibit similar patterns of online toxicity. This analysis can help uncover common traits, spatiotemporal patterns, and social dynamics associated with online toxicity in the Bangladeshi community. A block diagram of the proposed methodology is depicted in Figure 3.1.

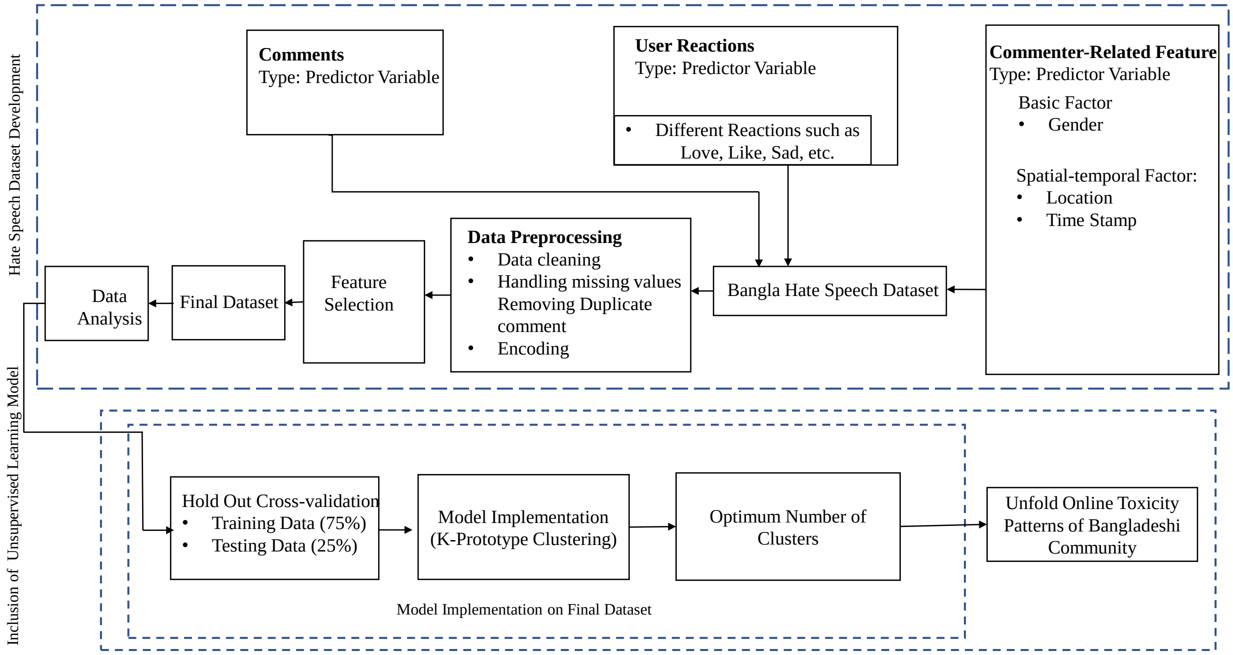

Hate Speech Dataset Development
Comments
Type: Predictor Variable
User Reactions
Type: Predictor Variable
• Different Reactions such as Love, Like, Sad, etc.
Commenter-Related Feature
Type: Predictor Variable
Basic Factor
• Gender
Spatial-temporal Factor:
• Location
• Time Stamp
Bangla Hate Speech Dataset
Data Preprocessing
• Data cleaning
• Handling missing values Removing Duplicate comment
• Encoding
Feature Selection
Final Dataset
Data Analysis
Inclusion of Unsupervised Learning Model
Hold Out Cross-validation
• Training Data (75%)
• Testing Data (25%)
Model Implementation (K-Prototype Clustering)
Optimum Number of Clusters
Unfold Online Toxicity Patterns of Bangladeshi Community
Model Implementation on Final Dataset

Figure 3.1: A block diagram of proposed research methodology

The following sections provide more detailed descriptions of the five stages of the proposed methodology for unfolding online toxicity patterns of the Bangladeshi Community.

3.1 Data Collection

In the first stage of the framework, data is extracted from various popular Facebook newspapers and news channels in Bangladesh, including Prothom Alo, Daily Star, Rumor Scanner, Bangladesh Protidin, Channel 24, Shomoy News, ATN News, and Independent TV. These sources are selected based on their popularity and the high user engagement they receive on Facebook. By analyzing the usage patterns and customs of social media users in Bangladesh through these sources, researchers can gain insights into the online behavior of the Bangladeshi community.

However, collecting data from Facebook presents challenges. The Facebook API no longer provides data to users (Mike Clark, 2021), so researchers have had to rely on third-party tools to gather the necessary information. Despite extensive searching, a perfect third-party tool that provides all the required data from Facebook could not be found. Therefore, a combination of manual collection and third-party tools was used to gather the data. While the commenter's gender and location were manually collected, other data, such as comments, timestamps, and engagement metrics, were obtained through third-party tools.

Although popular social media platforms like Twitter, YouTube, Reddit, and Instagram are widely used in Bangladesh, the data collection efforts focused solely on Facebook (Napoleon Cat Stats, 2023). This decision was influenced by the high engagement of Bangladeshi social media users on Facebook, as it remains the dominant platform in the country.

The collected dataset includes comments from five categories: National, International, Sports, Entertainment, and Miscellaneous. For each post within these categories, comments and associated information such as the commenter's gender, location (specifically Upazila and District), timestamps, and engagement metrics are extracted. This comprehensive hate speech dataset represents the Bangladeshi online community and provides a valuable resource for analyzing online toxicity patterns.

To ensure consistency in the dataset, data transformation is necessary. Commenters' profiles often mention their locations in various formats, such as districts, unions, or upazilas. However, only the Upazila and District fields are retained for commenters residing in Bangladesh for the dataset. For commenters located outside Bangladesh, only the associated country name is mentioned, providing a standardized format for location data in the dataset.

3.2 Feature Selection

The focus of the second stage is on feature selection, where the aim is to choose a subset of pertinent features from the available data. These selected features are crucial in enhancing the model's predictive power. In this study, the required attributes are categorized into four main categories, each with its own subcategories.

1. ***Types of Toxicity:*** This category aims to assess the level of offensiveness in a comment and determine the specific subcategory it falls under. The attributes considered within this category include Toxic, Insult, Profanity, Identity attack, Threat, and Neutral.

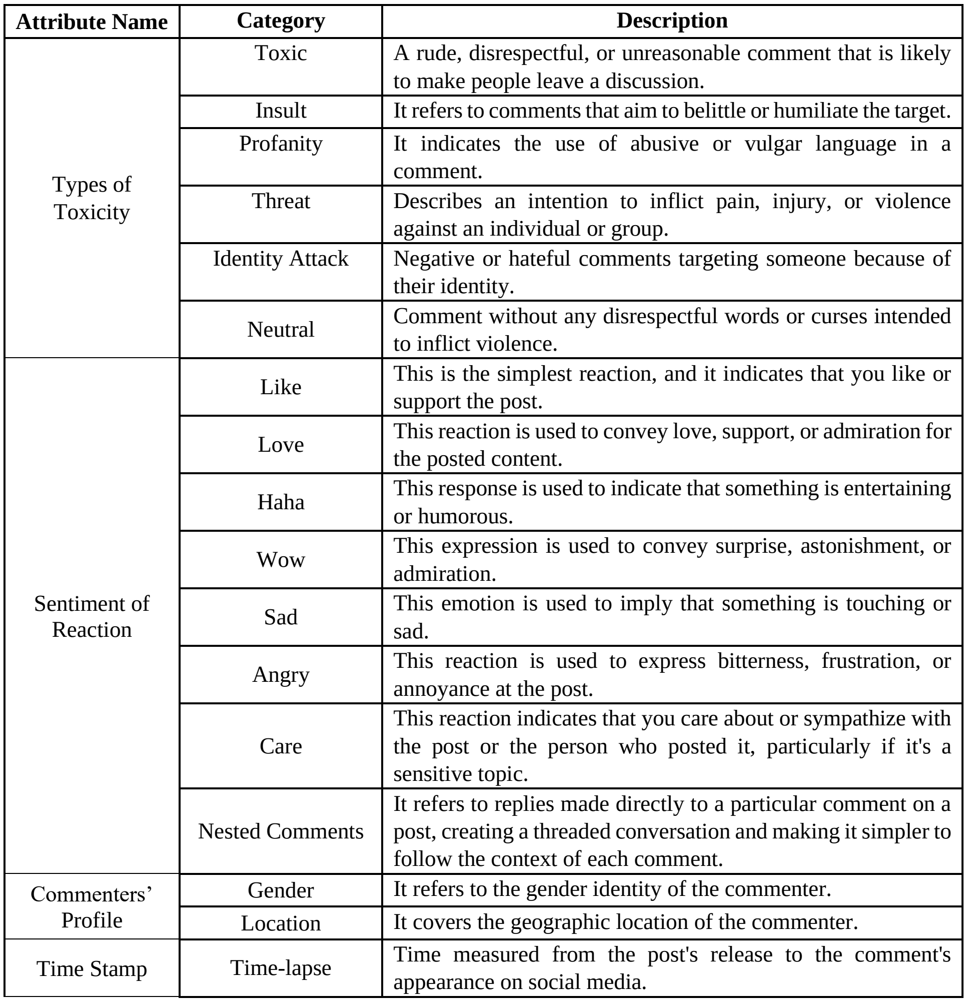
2. ***Sentiment of Reaction:*** Here, the objective is to identify the reaction and response to offensive comments. This analysis examines whether these reactions exhibit positive energy or contribute to escalating future hostility. The category encompasses seven types of reactions and also considers nested comments.

3. ***Commenters' Profile:*** This category focuses on the demographic information of commenters. Specifically, it includes their gender and location. It should be noted that the availability of age information for many commenters was limited during the manual data compilation process, despite being initially considered for inclusion in the research.

4. ***Time Stamp:*** This attribute indicates the duration between the release of a post and a particular comment. It provides insights into the temporal aspect of the comments.

All features with their description are presented in Table 3.1.

Table 3.1: Details of all features and their description.



Attribute Name	Category	Description
Types of Toxicity	Toxic	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.
	Insult	It refers to comments that aim to belittle or humiliate the target.
	Profanity	It indicates the use of abusive or vulgar language in a comment.
	Threat	Describes an intention to inflict pain, injury, or violence against an individual or group.
	Identity Attack	Negative or hateful comments targeting someone because of their identity.
	Neutral	Comment without any disrespectful words or curses intended to inflict violence.
Sentiment of Reaction	Like	This is the simplest reaction, and it indicates that you like or support the post.
	Love	This reaction is used to convey love, support, or admiration for the posted content.
	Haha	This response is used to indicate that something is entertaining or humorous.
	Wow	This expression is used to convey surprise, astonishment, or admiration.
	Sad	This emotion is used to imply that something is touching or sad.
	Angry	This reaction is used to express bitterness, frustration, or annoyance at the post.
	Care	This reaction indicates that you care about or sympathize with the post or the person who posted it, particularly if it's a sensitive topic.
	Nested Comments	It refers to replies made directly to a particular comment on a post, creating a threaded conversation and making it simpler to follow the context of each comment.
Commenters' Profile	Gender	It refers to the gender identity of the commenter.
	Location	It covers the geographic location of the commenter.
Time Stamp	Time-lapse	Time measured from the post's release to the comment's appearance on social media.

3.3 Data Preprocessing

Data preprocessing is a critical step in improving the accuracy of machine learning models by cleaning, transforming, and organizing raw data into a format that can be effectively analyzed. In this study, several data preprocessing techniques are applied. Firstly, duplicate comments are removed from the raw dataset to address potential issues caused by data entry errors, system errors, or data integration from multiple sources. Then, we check the missing values and delete them. Finally, the dataset repository is ready for further analysis.

3.4 Data Analysis

After the hate speech dataset is cleaned thoroughly, descriptive techniques are used to analyze the dataset. This analysis involves visualizing and exploring patterns, trends, and relationships among the selected features to better understand online toxicity within the Bangladeshi community. Initially, we focus on examining how comments are distributed across different categories. In Figure 3.2, the visual representation shows the distribution of comments gathered from various news categories on Facebook. It is worth noting that the *National* category receives the highest number of comments among the chosen categories. This can be attributed to the active participation of Bangladeshi Facebook users in discussions concerning national matters.

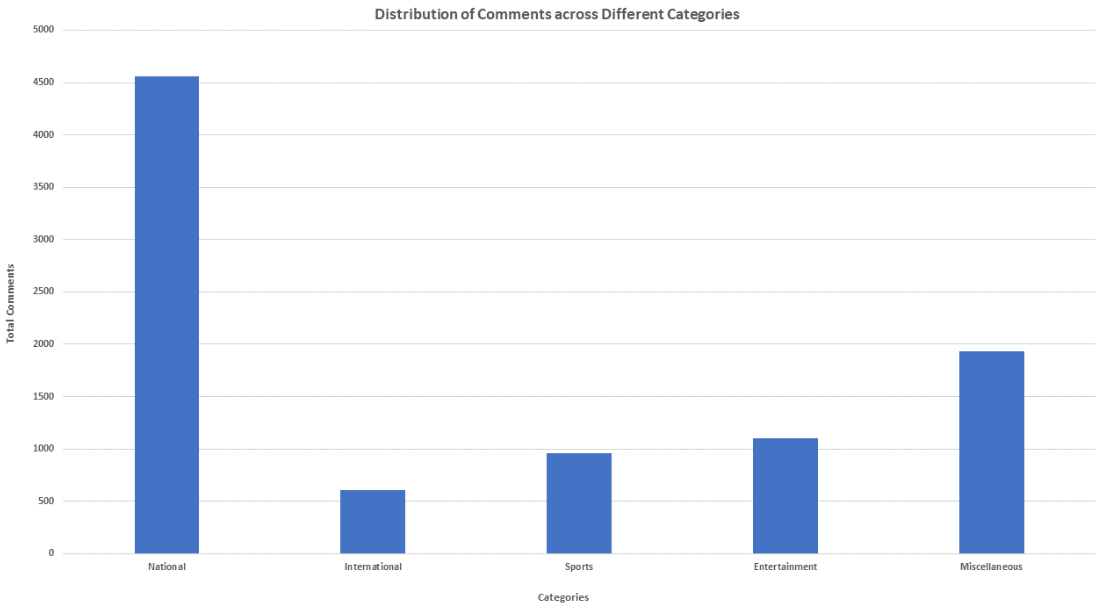


Figure 3.2: Distribution of comments across different categories

To further analyze the hate speech dataset, our attention turns to the distribution of nested comments. As mentioned, comments are categorized into six classes using Human Coders and the BanglaBERT model. Figure 3.3 illustrates the distribution of nested comments based on the assessments from Human Coders and BanglaBERT. It is important to highlight that in both approaches, the most common toxicity type of nested comment is classified as *Neutral*, with the *Identity attack* category closely following behind.

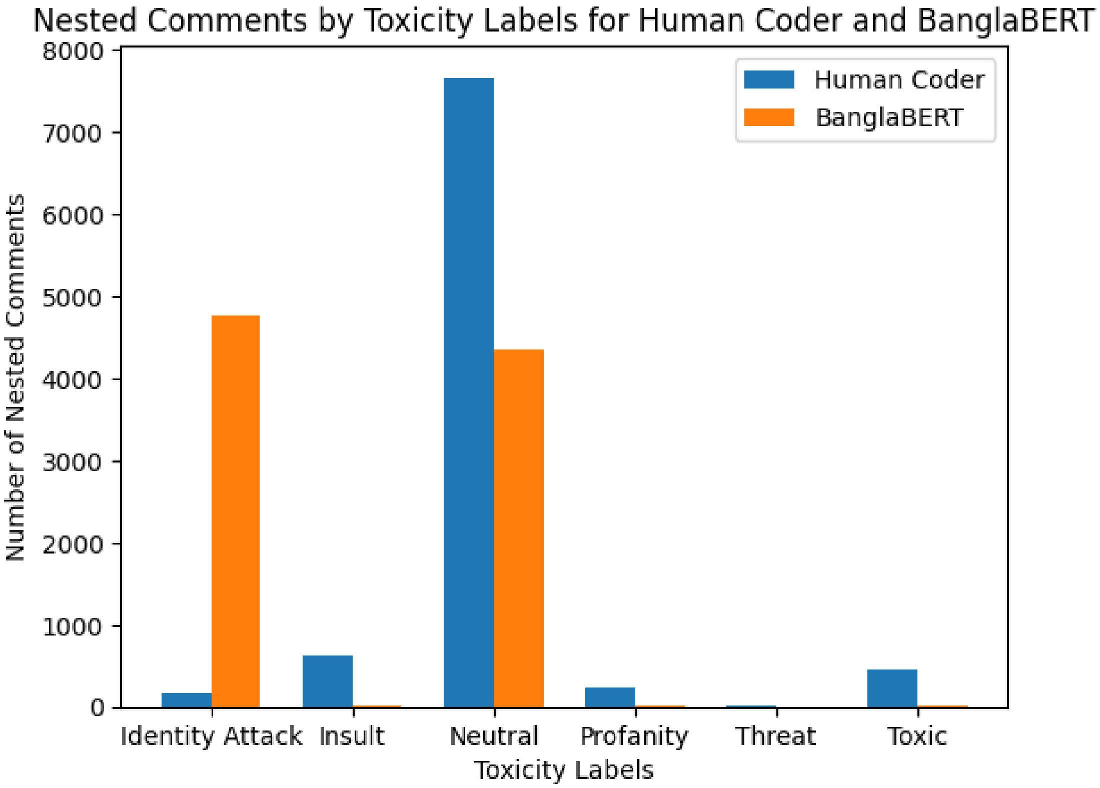


Figure 3.3: Distribution of nested comments across different categories

We proceed with analyzing toxicity distribution in comments across different categories and as a whole. Figure 3.4 presents a visual representation of each category and the overall dataset.

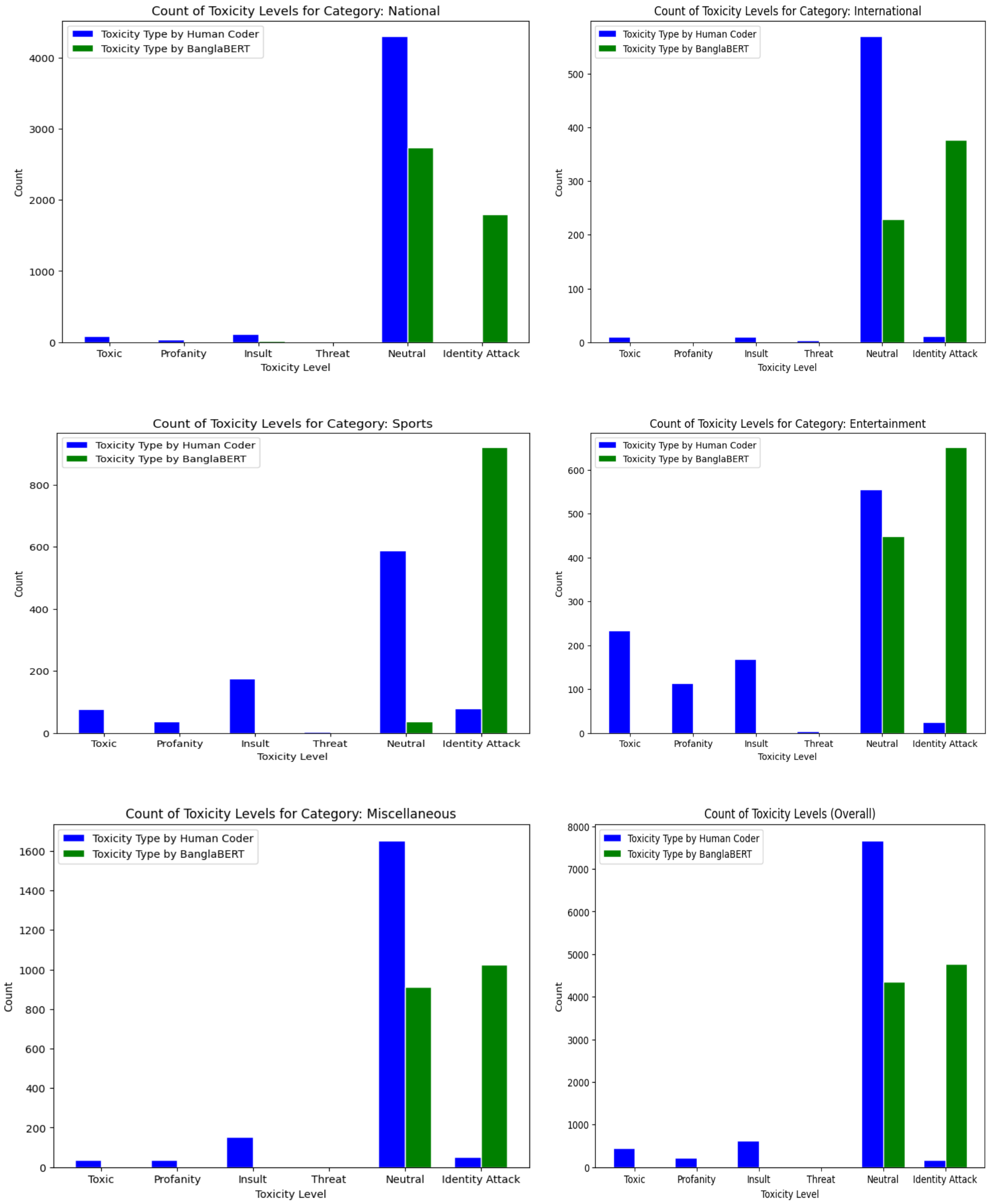


Figure 3.4: Toxicity distribution of comments for different categories and overall

The figure showcases the distribution of toxicity types in the comments, allowing us to observe the prevalence of each category across different categories and the dataset as a whole. According to the Human Coder approach, the most common toxicity type for all categories and as a whole are labeled as *Neutral*. However, when using the BanglaBERT model, except for the National category, the *Identity attack* category emerges as the most frequent toxicity type. One possible reason is that the Human Coders and the BanglaBERT model have different biases when annotating or classifying toxicity. The Human Coders might have subjectively interpreted and labeled comments differently compared to the automated classification performed by the BanglaBERT model.

Finally, to gain valuable insights into the emotional landscape of the Bangladeshi Facebook community, sentiment analysis is employed. This analysis helps understand users' emotional reactions, such as Love, Like, Sad, etc., towards different comments, including Toxic, Neutral, Identity attack, etc. Figure 3.5 visually presents the distribution of reactions across the six predefined types of toxicity - Toxic, Insult, Profanity, Threat, Identity Attack, and Neutral, providing an overview of the sentiments expressed within the analyzed dataset by under Human Coder and BanglaBERT.

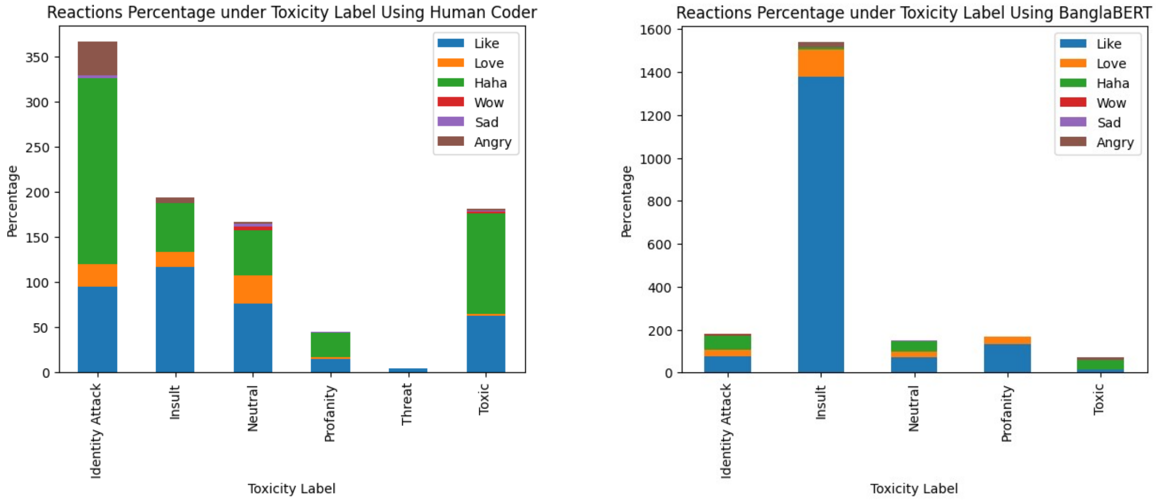


Figure 3.5: Sentiment analysis using Human Coder and BanglaBERT

According to the Human Coder approach, the most common reaction for different toxic comments is labeled as *Haha*, followed by *Like*. On the other hand, when utilizing the BanglaBERT model, the *Like* reaction emerges as the most frequent response. This sentiment analysis aids in gaining a deeper understanding of the emotional dynamics and user reactions within the Bangladeshi Facebook community.

3.5 Model Implementation

We employ the K-Prototype clustering algorithm on our final dataset to delve into the spatiotemporal analysis, user reaction traits, and social network dynamics. By utilizing this unsupervised learning technique, we can effectively handle the mixed nature of our dataset, which includes both numerical and categorical attributes. The K-Prototype clustering algorithm offers a comprehensive approach that facilitates the identification of meaningful clusters, allowing for a deeper understanding of the underlying structures and relationships among individuals within the dataset. This analysis aids in uncovering valuable insights related to the spatiotemporal patterns, user reactions, and social network dynamics in the Bangladeshi Facebook community.

Chapter 4

Results and Discussion

4.1 Temporal and Spatial Pattern of Online Toxicity

To address our first RQ, we conducted a comprehensive analysis of the temporal and spatial patterns observed within the identified clusters by using the K-Prototype. This analysis explores the relationship between the timing of posts and comments and their corresponding spatial distribution within the dataset. Figure 4.1 presents the temporal analysis of the clusters.

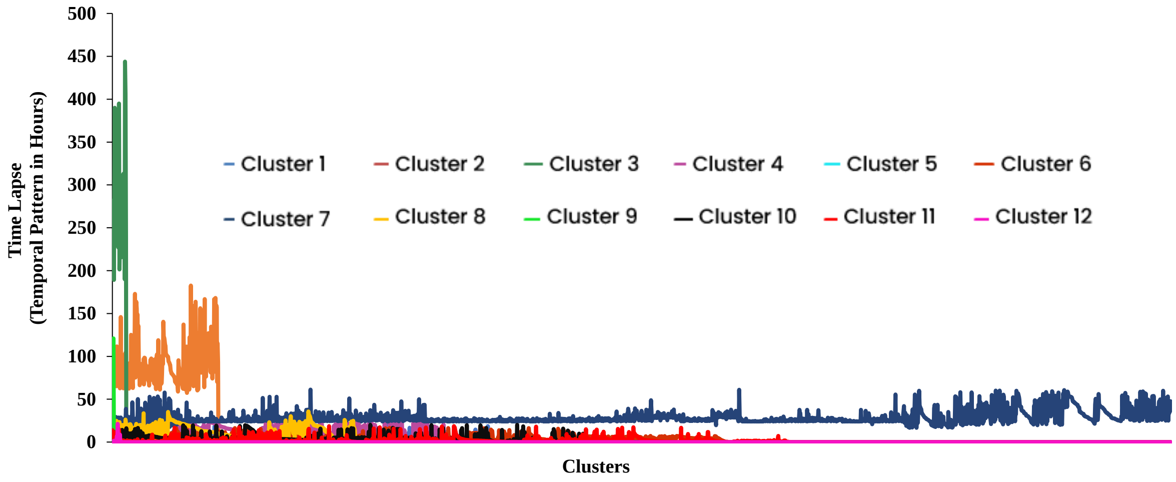


Figure 4.1: Temporal analysis of the clusters

Among the identified clusters, our attention is particularly drawn to *Cluster 3* due to its distinctive temporal characteristic. This specific cluster exhibited a significant time lapse, which refers to the time difference between when a post is created and when the corresponding comment is made. The graphical representations vividly illustrate this substantial time difference compared to the other clusters, enabling us to visually comprehend the temporal dynamics at play. Delving further into the temporal patterns exhibited by each cluster, we make an intriguing observation regarding the disclosure of location information by commenters on Facebook, depicted in Figure 4.2.

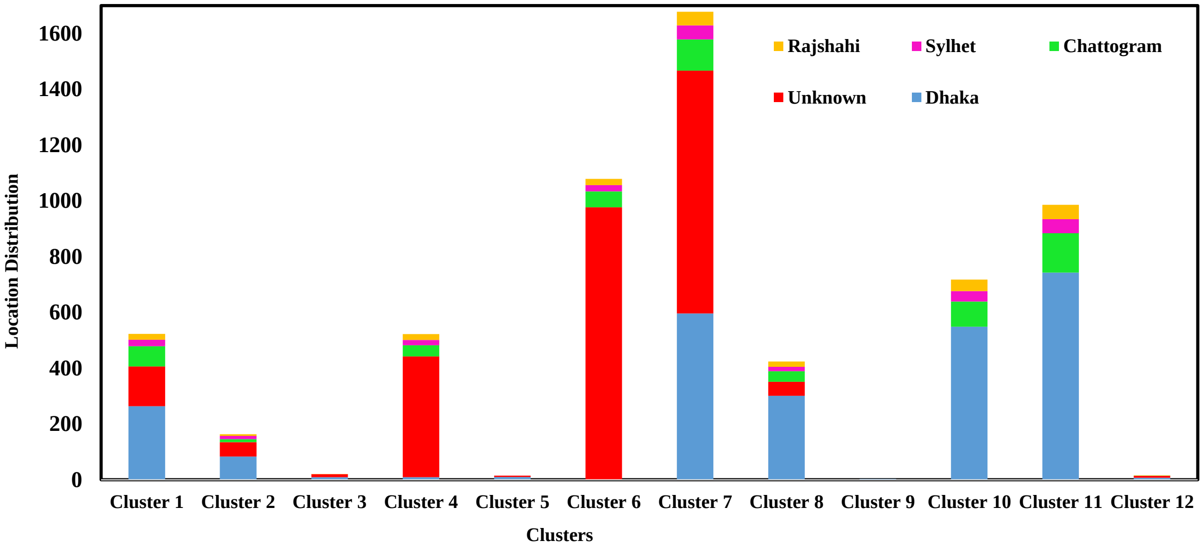


Figure 4.2: Spatial analysis of the clusters

Notably, a significant portion of commenters choose not to publicly disclose their geographical location. However, *Dhaka* emerged as the most frequently mentioned location among those who provided location information. Our analysis of the temporal and spatial patterns within the dataset uncovers intriguing insights regarding the timing of posts and comments and the disclosure of geographical information. These findings contribute to our understanding of online dynamics and provide a foundation for further exploration into the intricate relationships between time, space, and online toxicity in the Bangladeshi community.

4.2 Traits of Online Toxicity

To explore individuals' reactions to toxicity, we delved into our second RQ and analyzed clusters. We aim to understand how different types of toxicity elicited varying reactions from individuals. Figure 4.3 presents the toxicity analysis of the clusters.

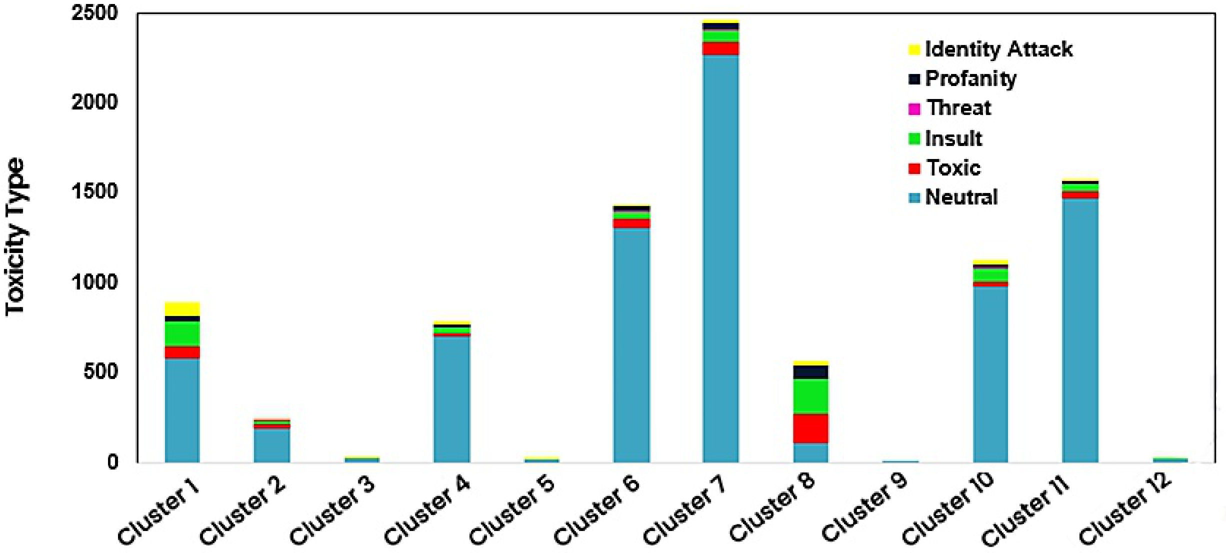


Figure 4.3: Traits of online toxicity of the clusters

Interestingly, we find that the most prevalent reaction type across most clusters is *Neutral*. This suggests that individuals tend to respond with a neutral stance when confronted with toxic content. This prevalence of neutrality can be attributed to the context-based nature of our analysis. Human Coders take into consideration the surrounding context of the posts when categorizing comments, and this likely influenced the neutral responses. These findings shed light on the complex dynamics of individuals' responses to toxicity. The understanding that context plays a crucial role in shaping reactions adds depth to our analysis and emphasizes the need for a nuanced understanding of online toxicity. By considering the context in which toxic content is presented, we can gain a more comprehensive understanding of how individuals engage with and respond to such content.

4.3 Social Dynamics of Online Toxicity

To delve into the social dynamics of online toxicity, we turn our attention to our final RQ. Figure 4.4 examines the distribution of clusters based on gender. However, it is worth noting that in some cases, the gender of commenters could not be determined from their profiles, and a few comments originated from random Facebook pages. To address this, we employ the letter P to represent unknown gender in one scenario and N in another.

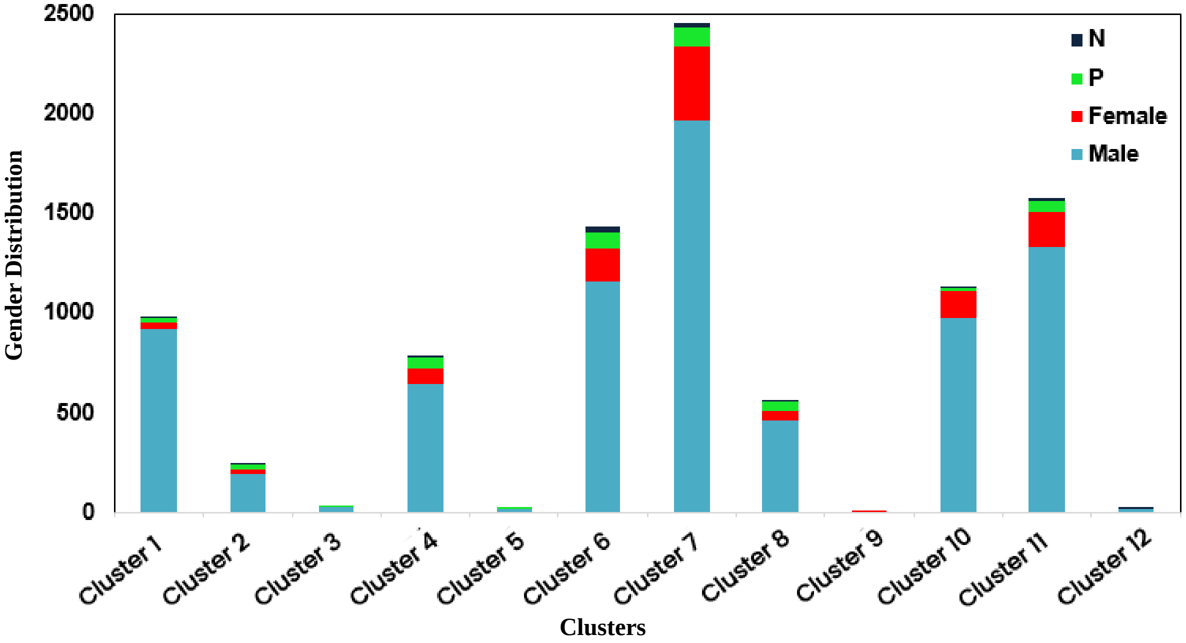


Figure 4.4: Gender distribution among the clusters

It reveals that the majority of commenters are identified as male. This finding sheds light on the gender composition of individuals actively participating in online discussions surrounding toxicity. It provides valuable insights into the social dynamics surrounding online toxicity and the gendered nature of engagement in these discussions. Additionally, we examine the total number of comments for each cluster, Figure 4.5. Notably, *Cluster 7* has the highest overall number of comments, followed by *Cluster 11* and *Cluster 6*.

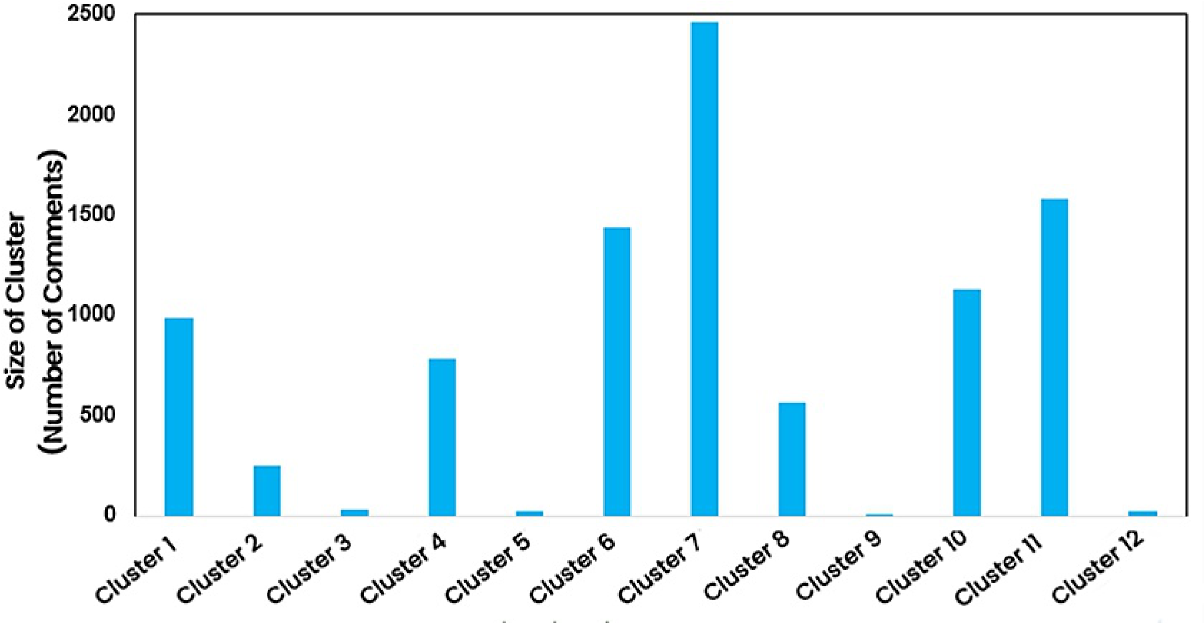


Figure 4.5: Comments distribution among the clusters

By exploring the gender distribution and the volume of comments across different clusters, we gain insights into the social dynamics of online toxicity. These findings contribute to our broader comprehension of the complex interplay between gender, participation, and the manifestation of toxicity within online communities. Understanding these dynamics is crucial for developing strategies and interventions to foster healthier and more inclusive online environments.

Chapter 5

Conclusion, Limitations, and Future Directions

This study explores the phenomenon of online toxicity within the Bangladeshi community using an unsupervised machine-learning approach. By analyzing a comprehensive dataset of toxic comments from popular Bangladeshi news channels and newspapers on Facebook, this study successfully uncovers patterns and dynamics of online toxicity, reactions of individuals to toxic activities, and social dynamics involved in toxic behavior. The findings of the study reveal valuable insights into the prevalence and locations of online toxicity in Bangladesh. The majority of commenters identified themselves as male, and Dhaka emerged as the most commonly mentioned location. The analysis also highlighted the influence of context in categorizing toxic comments, with human coders considering contextual information more extensively than the BanglaBERT model.

The proposed framework makes important contributions by creating a comprehensive database of toxic remarks made on Bangla Facebook, including spatiotemporal information. This database fills a gap in the existing literature and provides a foundation for targeted interventions and educational efforts. Additionally, the study explores the relationship between demographic characteristics and engagement in toxic commenting activity, which can inform prevention and intervention initiatives. The study's contributions extend to practical implications, as the knowledge gained can inform the development of targeted interventions and strategies aimed at mitigating the negative impact of online toxicity and fostering a healthier and more inclusive online environment within the Bangladeshi community.

Despite the valuable insights gained from this study, there are several limitations that should be acknowledged. First, the study focuses on toxic comments from popular Bangladeshi news channels and newspapers on Facebook, which may not fully represent online toxicity across all social media platforms or the entirety of the Bangladeshi community. The dataset's bias towards a specific platform and source of comments should be taken into account when generalizing the findings. Second, the dataset used for analysis is collected from Facebook, which has specific policies governing data use. Access to data may be limited, and automated data scraping is prohibited, requiring written permission for data collection. Additionally, Facebook moderators may delete content that violates community standards, which can affect the quantity and quality of the data. Furthermore, while efforts are made to create a comprehensive dataset, the size of the dataset may have limitations, and the findings may not be fully generalizable to the entire Bangladeshi community. Increasing the dataset size and analyzing social media behavior across different demographics could enhance the validity and generalizability of the research findings.

In terms of future research, there are several directions that can be pursued to further advance our understanding of online toxicity within the Bangladeshi community. One avenue is the development of predictive modeling approaches to study the characteristics of identified clusters and predict future behavior or trends related to online toxicity. This would provide insights into the long-term patterns and dynamics of toxic behavior, enabling proactive measures to mitigate its impact. Additionally, future studies could focus on expanding the dataset and conducting demographic analyses to enhance the validity and generalizability of the findings. This would involve collecting larger datasets from a broader range of social media platforms and demographics, providing a more comprehensive understanding of online toxicity patterns, and identifying specific risk factors among different population segments.

Another important area for future research is the development and evaluation of interventions and strategies to combat online toxicity. This could involve the design of educational programs, digital resilience training, and community-based initiatives tailored to the cultural and social context of Bangladesh. By implementing and assessing the effectiveness of such interventions, researchers can contribute to creating a healthier and more inclusive online environment. Additionally, comparative analyses of online toxicity patterns and dynamics across different cultural and linguistic contexts can provide valuable insights into the role of cultural and social factors in shaping online behavior. Such comparative studies would facilitate the development of culturally tailored interventions to address the unique challenges faced by the Bangladeshi community. By pursuing these future research directions, we can advance our understanding of online toxicity, develop effective interventions, and contribute to fostering a positive and respectful online environment within the Bangladeshi community.

References

S. Ahammed, M. Rahman, M. H. Niloy, and S. M. M. H. Chowdhury. Implementation of machine learning to detect hate speech in bangla language. In *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, pages 317–320, 2019. doi: 10.1109/SMART46866.2019.9117214.

A. A. Aporna, I. Azad, N. S. Amlan, M. H. K. Mehedi, M. J. A. Mahbub, and A. A. Rasel. Classifying offensive speech of bangla text and analysis using explainable ai. In M. Singh, V. Tyagi, P. K. Gupta, J. Flusser, and T. Ören, editors, *Advances in Computing and Data Sciences*, pages 133–144, Cham, 2022. Springer International Publishing.

M. A. Awal, M. S. Rahman, and J. Rabbi. Detecting abusive comments in discussion threads using naïve bayes. In *2018 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pages 163–167, 2018. doi: 10.1109/ICISET.2018.8745565.

I. Awan. Islamophobia and twitter: A typology of online hate against muslims on social media. *Policy & Internet*, 6(2):133–150, 2014.

F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, 38: 100311, 2020.

N. Banik and M. H. H. Rahman. Toxicity detection on bengali social media comments using supervised models. In *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*, pages 1–5, 2019. doi: 10.1109/ICIET48527.2019.9290710.

F. T. Boishakhi, P. C. Shill, and M. G. R. Alam. Multi-modal hate speech detection using machine learning. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4496–4499, 2021. doi: 10.1109/BigData52589.2021.9671955.

S. A. Castaño-Pulgarín, N. Suárez-Betancur, L. M. T. Vega, and H. M. H. López. Internet, social media and online hate speech. systematic review. *Aggression and Violent Behavior*, 58:101608, 2021.

P. Chakraborty and M. H. Seddiqui. Threat and abusive language detection on social media in bengali language. In *2019 1st International Conference* *on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–6, 2019. doi: 10.1109/ICASERT.2019.8934609.

N. Chetty and S. Alathur. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118, 2018.

A. K. Das, A. A. Asif, A. Paul, and M. N. Hossain. Bangla hate speech detection on social media using attention-based recurrent neural network. *Journal of Intelligent Systems*, 30(1):578–591, 2021. doi: doi:10.1515/jisys-2020-0060. URL https://doi.org/10.1515/jisys-2020-0060.

E. A. Emon, S. Rahman, J. Banarjee, A. K. Das, and T. Mittra. A deep learning approach to detect abusive bengali text. In *2019 7th International Conference on Smart Computing Communications (ICSCC)*, pages 1–5, 2019. doi: 10.1109/ICSCC.2019.8843606.

T. Ghosh, A. A. K. Chowdhury, M. H. A. Banna, M. J. A. Nahian, M. S. Kaiser, and M. Mahmud. A hybrid deep learning approach to detect bangla social media hate speech. In S. Hossain, M. S. Hossain, M. S. Kaiser, S. P. Majumder, and K. Ray, editors, *Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021*, pages 711–722, Singapore, 2022. Springer Nature Singapore.

M. I. Hossain Junaid, F. Hossain, and R. M. Rahman. Bangla hate speech detection in videos using machine learning. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pages 0347–0351, 2021. doi: 10.1109/UEMCON53757.2021.9666550.

A. M. Ishmam and S. Sharmin. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 555–560, 2019. doi: 10.1109/ICMLA.2019.00104.

M. Islam, M. S. Hossain, and N. Akhter. Hate speech detection using machine learning in bengali languages. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1349–1354, 2022. doi: 10.1109/ICICCS53718.2022.9788344.

M. R. Karim, S. K. Dey, T. Islam, S. Sarker, M. H. Menon, K. Hossain, M. A. Hossain, and S. Decker. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2021. doi: 10.1109/DSAA53316.2021.9564230.

M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, and B. R. Chakravarthi. Multimodal hate speech detection from bengali memes and texts. In A. K.

M, B. R. Chakravarthi, B. B, C. O'Riordan, H. Murthy, T. Durairaj, and T. Mandl, editors, *Speech and Language Technologies for Low-Resource Languages*, pages 293–308, Cham, 2023. Springer International Publishing.

L. Li, Z. Ma, and T. Cao. Data-driven investigations of using social media to aid evacuations amid western united states wildfire season. *Fire Safety Journal*, 126:103480, 2021. ISSN 0379-7112. doi: https://doi.org/10.1016/j.firesaf.2021.103480. URL https://www.sciencedirect.com/science/article/pii/S0379711221002228.

Mike Clark. How we combat scraping, April 15 2021. Retrieved April, 2021 from https://about.fb.com/news/2021/04/how-we-combat-scraping/.

N. S. Mullah and W. M. N. W. Zainon. Advances in machine learning algorithms for hate speech detection in social media: a review. *IEEE Access*, 9:88364–88376, 2021.

Napoleon Cat Stats. Social media in bangladesh - 2023 stats platform trends, April 6 2023. Retrieved April, 2023 from https://oosga.com/social-media/bgd/.

O. Oh, M. Agrawal, and H. R. Rao. Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises.

MIS Quarterly, pages 407–426, 2013.

M. R. Rezvi and M. R. Hossain. Exploring issues of online hate speech against minority religious groups in bangladesh. *Available at SSRN*, 2021.

N. Romim, M. Ahmed, H. Talukder, and M. Saiful Islam. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In M. S. Uddin and J. C. Bansal, editors, *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 457–468, Singapore, 2021. Springer Singapore.

N. Romim, M. Ahmed, M. S. Islam, A. Sen Sharma, H. Talukder, and M. R. Amin. BD-SHS: A benchmark dataset for learning to detect online Bangla hate speech in different social contexts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5153–5162, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.552.

M. Sarker, M. F. Hossain, F. R. Liza, S. N. Sakib, and A. 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, 2022. doi: 10.1109/ICAEEE54957.2022.9836407.

A. Saroj and S. Pal. Use of social media in crisis management: A survey. *International Journal of Disaster Risk Reduction*, 48:101584, 2020. ISSN 2212-4209. doi: https://doi.org/10.1016/j.ijdrr.2020.101584. URL https://www.sciencedirect.com/science/article/pii/S221242091931684X.

P. William, R. Gade, R. e. Chaudhari, A. B. Pawar, and M. A. Jawale. Machine learning based automatic hate speech recognition system. In *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pages 315–318, 2022. doi: 10.1109/ICSCDS53736.2022.9760959.

J. Zhu, F. Xiong, D. Piao, Y. Liu, and Y. Zhang. Statistically modeling the effectiveness of disaster information in social media. In *2011 IEEE Global Humanitarian Technology Conference*, pages 431–436. IEEE, 2011.