4WHContext: A Context Based Hate Speech Detection Framework From Social Media Posts

Md Jahangir Alam[1], Ismail Hossain[1], Sai Puppala[2], and Sajedul Talukder[3]

[1] University of Texas at El Paso TX, USA
{malam10,ihossain}@miners.utep.edu
[2] Southern Illinois University Carbondale IL, USA
saimaniteja.puppala@siu.edu
[3] University of Texas at El Paso TX, USA
stalukder@utep.edu

Abstract. Detecting hate speech online has become increasingly important due to the surge in harmful content on social media. This is particularly challenging for resource-constrained languages like Bengali. This paper presents a dataset specifically created for detecting contextual hate speech in Bengali, developed through extensive data collection, preprocessing, and both manual and automatic labeling. It comprises 15,000 annotated texts categorized into hate speech and non-hate speech, with a Cohen's kappa score of 0.88, reflecting strong agreement among annotators. We assessed the dataset using machine learning (ML), deep learning (DL), and BERT-based models. Among these, the BERT-based model XLM-R excelled, attaining an F1 score of 0.94 and an accuracy of 0.92 when context was considered, and an F1 score of 0.89 with an accuracy of 0.87 without context. These findings highlight that integrating context notably enhances the accuracy of hate speech detection, contributing to more effective methods for identifying and mitigating harmful online content.

Keywords: social network · hate speech detection · Bangla language · TF-IDF · contextual hate speech.

1 Introduction

Bangla, serving as the primary language for 98% of Bangladesh's population, holds formal recognition as the national language [17]. Its influence extends beyond national borders, with a significant Bangla-speaking diaspora found in regions such as the Middle East, Europe, and the USA [20]. The Digital Bangladesh initiative has further amplified Bangla's presence on digital platforms like Facebook, LinkedIn, and Twitter [12]. Facebook, particularly, stands as the predominant social platform in Bangladesh, boasting 33.71 million active Bengali users [13]. This manuscript focuses on the critical issue of hate speech detection within Bangla text on social media. Despite extensive research in languages like English, Russian, and Arabic, there remains a notable gap in hate speech detection methods that incorporate contextual information. Our study addresses this

gap by conducting experiments specifically tailored for Bangla datasets, with potential applicability to English datasets and beyond.

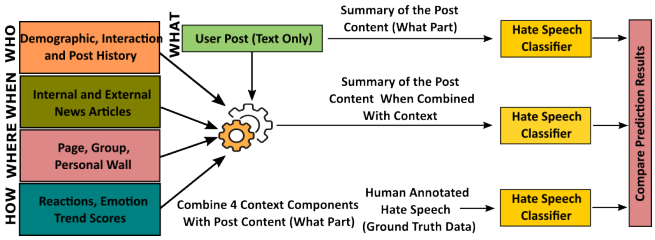

Fig. 1. Architecture of Contextual Hate Speech Detection

Hate speech detection in Facebook's conversational textual content has become a focal point for NLP researchers, given the widespread expression of opinions on Web 2.0 platforms. Developing an automatic system for Bengali presents challenges due to limited resources and standardized corpora. This paper outlines the creation of a hate speech dataset involving data collection, preprocessing, human labeling, automatic labeling, and label verification. The dataset comprises 15,000 texts categorized into hate speech and non-hate speech, annotated with a high Cohen's score of 0.88, indicating strong annotator agreement.

We conducted hate speech classification experiments using ML, DL, and BERT-based models. In testing, the BERT-based model XLM-R achieved the highest performance with an F1 score of 0.94 and accuracy of 0.92 when using context information. Without context, it achieved an F1 score of 0.89 and accuracy of 0.87. This highlights the effectiveness of context in improving hate speech detection accuracy. The main contributions of this paper encapsulate:

- **Hate Speech Dataset Development:** We collect user posts and develop a semi-automated data annotation process to annotate the dataset of size 15K.
- **Defining Context:** We define context with multiple components which are 'WHO', 'WHEN', 'WHERE', 'HOW'. Using these context elements we get close to human level understanding of the post which facilitates the hate speech detection.
- **Comparative Performance of Different Models:** We compare Experiment on hate speech detection with context and without context. We utilize different ML, DL and BERT-based models with TF-IDF feature approach.

2 Related Works

In this section we discuss the literature study for hate speech detection from Bangla text. In recent studies, various datasets and methods have been employed

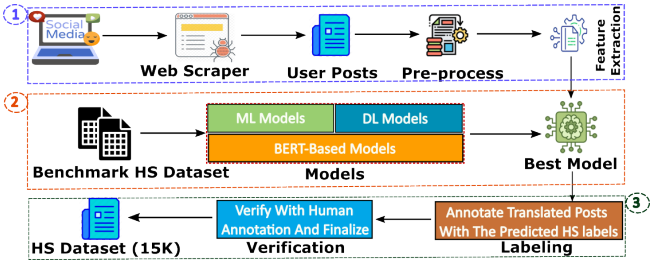

Fig. 2. Semi-automated dataset annotation process flow.

for hate speech detection in Bengali. Romim et al. [10] utilized 30,000 Facebook comments across nine categories, training multiple models including SVM and LSTM variants, with performance peaking at 87.5% accuracy for SVM.

Karim et al. [6] worked with 5,000 labeled examples from Facebook, YouTube, and newspapers, employing ML, DNN, and BERT variants, where an ensemble model achieved an 88% precision. Romim et al. [10] also gathered over 100,000 comments from YouTube and Facebook across seven categories but lacked detailed model information. Das et al. [18] focused on public pages of celebrities and other figures, training models like CNN with LSTM and GRU, achieving up to 77% accuracy.

Romim et al. [8] annotated 50,281 comments from social media, while Das et al. [1] examined Twitter data, using models such as m-BERT and XLM-Roberta, with XLM-Roberta achieving 79% accuracy. Ghosh et al. [2] leveraged a dataset of 35,000 hate statements, with models like XLM-RoBERTa and BanglaBERT attaining up to 90% accuracy. Jahan et al. [4] curated 15,000 comments from Facebook and YouTube, with BanglaHateBERT achieving 93.1% accuracy.

Kalita et al. [5] examined 2,180 Bengali texts, employing models such as CNN with FastText and Logistic Regression, with the highest accuracy reaching 81% when external datasets were used. Mathew et al. [7] enhances hate speech detection using textual context (Twitter, Gab), rationales by human annotators for interpretability, and annotations for targeted communities (race, religion, gender). BERT-MRP achieves 81.0% accuracy, F1 80.0%; Space-XLNet achieves 82.5% accuracy, F1 81.5%; Space-DistilBERT achieves 80.5% accuracy, F1 79.0%. These components improve accuracy in identifying harmful online content.

Gomez et al. [3] enhances hate speech detection using textual and visual contexts from tweets. Textual context includes tweet content processed with LSTM using GloVe embeddings and text extracted from images via Google Vision API's Text Detection module. Visual context involves images analyzed with CNN (Google Inception v3). Multimodal context integrates textual and visual data using models like FCM, SCM, and TKM. FCM achieves 75.4% accuracy, F1 score 74.2%; SCM achieves 77.8% accuracy, F1 score 76.3%; TKM achieves 78.5% accuracy, F1 score 77.1%. These models effectively combine visual and textual cues for improved hate speech detection. These studies emphasize the

variety of datasets and modeling methodologies in hate speech detection, showcasing the efficacy of BERT-based models in achieving superior accuracy. None of the studies explored the contextual dimensions of hate speech detection like ours.

3 Dataset

3.1 Data Collection

Demographic Dataset Collection We collect user demographic data from Facebook, including name, work experience, education, gender, and relationship status.

User Post Dataset Collection Our dataset comprises 15,000 posts sourced from diverse Facebook groups and pages, encompassing a range of demographics, interests, and locations. These posts include text, images, and videos, either standalone or in combined formats, as depicted in Figure 3. We employed a meticulous collection strategy guided by stringent criteria to maximize representativeness and generalizability.

We ensured efficient data gathering while adhering to ethical standards and respecting user privacy, in accordance with Facebook's terms of service and data regulations.

Interaction Dataset Collection We compile a dataset named "User Interaction History in Social Networks with Other Users' Posts." This dataset captures user activities such as commenting, sharing, and engaging in discussions on social networks. It provides insights into user behavior and social dynamics by exploring connections between users through their posts, usernames, and profile URLs. Our approach involves reverse engineering to understand information diffusion and community dynamics while strictly adhering to ethical standards and prioritizing user privacy, in compliance with platform policies and regulations.

Reverse Engineering Approach for Constructing User Interaction Dataset Using a reverse engineering approach, we extract user posts and metadata from engagements on social media platforms, constructing a cohesive user network based on profile URLs. This method reveals comprehensive user interaction patterns, highlighting individual behaviors and collective relationships within the online community. Analyzing this interconnected user base provides insights into social network structures, information diffusion, and community formation processes. Beyond collecting interaction data, this approach facilitates deeper understanding of user behavior, informing sociological studies and guiding content strategies on social media platforms.

3.2 Ethical Considerations

We adhered to ethical guidelines by obtaining permissions and consent from Facebook users and groups. Only anonymized data was stored, and sensitive information was carefully de-identified. Approval from the Institutional Review Board was secured to ensure compliance, with stringent measures in place to protect user privacy and uphold ethical standards throughout the study.

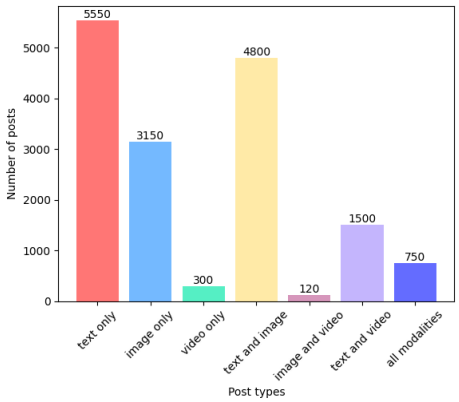

Fig. 3. Dataset Distribution: A Visual Representation of Text, Image, and Video Content Distribution

4 Methodology

In the following, we discuss the detailed methodology of our research. We first create the hate speech dataset by implementing a semi-automatic data annotation process. Then we discuss the methods of regular hate speech detection and contextual hate speech detection. Under contextual hate speech detection method we discuss the methods of the components of the context.

4.1 Semi-automated Hate Speech Dataset Development Process

Data Collection To experiment the contextual hate speech detection we develop our hate speech dataset of 15k collected from Facebook profiles, groups, pages etc. We collected posts in Bengali language. To develop dataset in Bengali is a critical challenge for any language processing task. One of the notable challenges is the scarcity of appropriate user posts dataset collected from social media platforms. We collect 15,000 data from social media platform and develop hate speech (HS) dataset following a semi-automatic annotation process. Figure 2 shows an overview of the development process of our HS dataset, which consists of five major phases: data collection, preprocessing, data annotation, BERT-based automatic label prediction and label verification respectively.

In our contextual hate speech detection project, we applied a semi-automatic data annotation process to a dataset of 15,000 posts, extending the work initiated in Ahmed et al. [19], with strict adherence to ethical and legal considerations. Targeting Facebook groups, profiles, and pages ranging from 1,000 to 5,000 members or followers, we collect data ensuring compliance with data usage policies and user privacy guidelines. Our approach encompassed approximately 500 unique users, ensuring diverse representation of Bangla language users and capturing the language's nuanced expressions comprehensively. Throughout our

data collection, we prioritized anonymity and respected user privacy, aligning with GDPR and local regulations in Bangladesh.

We extensively de-identified personal information using cryptographic hash functions to anonymize Facebook account names as needed. The dataset comprises 15,000 posts, categorized with 43% identified as hate speech and 57% as non-hate speech.

Preprocessing We addressed unique challenges specific to the Bangla language in social media, particularly within Facebook group discussions. These posts are often informal and include non-textual elements such as URLs, images, tags, and links. Our preprocessing focused on refining these texts by meticulously removing non-Bangla alphanumeric characters, punctuation, URLs, images, links, hashtags, and user tags. This process, consistent with our previous work [15, 16], prepared the text for hate speech detection analysis.

A crucial aspect of our preprocessing involved handling stop words, which are common words that add little semantic value and can introduce dataset noise. We tokenized the texts and systematically removed Bangla stop words using a comprehensive list available in a GitHub repository [14], referenced in the work of Tripto and Ali [11]. This approach ensures transparency and reproducibility in our research methodology.

Through these tailored preprocessing steps, our aim was to refine the dataset to accurately reflect the linguistic patterns essential for contextual hate speech detection from social media texts.

Manual Data Annotation The whole corpus was labeled manually, followed by majority voting to assign a suitable label. The labeling or annotation tasks were performed by a group of graduate students of Computer Science who are doing research on NLP. The majority voting mechanism was used to decide the final label of each post. The evaluation of the dataset, with a Cohen's score of 0.88, indicates strong agreement among annotators.

Label Prediction Using BERT-based Model As part of the semi-automatic annotation process we train several machine learning models and perform prediction for our dataset using the best performing model. We train ML, DL and BERT-based models with a benchmark dataset [4]. We find the BERT-based model mBERT performs the best.

Label Verification The majority voting by the annotators has decided the original label of data. We perform a label verification of the original labels with the predicted labels generated by model trained with benchmark HS dataset. If a mismatch found between labels for the same post we separate the post for further discussion and human labeling. We found around 41% match among original labels and the predicted labels. We consider the 59% data for relabeling by the human annotators. Thus we perform the label verification.

4.2 Regular Hate Speech Detection

Figure 5 shows the regular hate speech detection process. In this process, we pass the dataset through pre-processing phase, feature extraction phase and model training and classification phases. During regular HS detection dataset

contains only the text component of the user post. We do not consider the user information, user posting pattern, interaction pattern, contemporary news or events occurring at the time of the post, how user friends react with his/her post etc. We train several models for HS detection including ML, DL, BERT-based models.

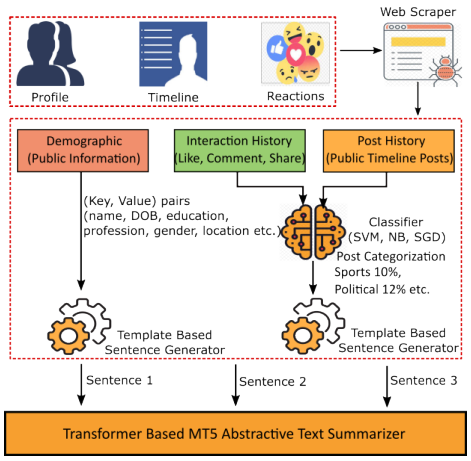

Fig. 4. Who component of the context

4.3 Contextual Hate Speech Detection

Figure 1 shows the overall process of contextual hate speech detection. In this process, there are 4 components of the context which are 'WHO', 'WHEN', 'WHERE', 'HOW'. During regular hate speech detection we consider only the text content of the user post. In contract, in contextual hate speech detection, we pass the summaries of the 4 components along with the text content of the user post for the classification. In the subsequent sub-sections we discuss the methods of the 4 components of the context.

4.4 "Who" Component Of The Context

In this section we describe how we perform the persona categorization, user post history categorization, user post interaction categorization. All these things comprise the who component of the context. Figure 4 depicts the flow of generating summary from the who component of the context.

User Persona Categorization Using ML Models We gather a comprehensive dataset of user demographics, including personal details like name, age,

gender, date of birth, education, profession, relationship status, and interests. This dataset is annotated with 12 distinct persona categories such as political, businessperson, entrepreneur, educator, religious, artist, athlete, technologist, scientist, healthcare professional, traveler, and musician. Using machine learning models, we classify users into these persona categories. These insights enhance our understanding of user characteristics and traits, thereby improving contextual hate speech detection.

User Post Categorization Using ML Models We collect user posts from their Facebook walls and use machine learning algorithms to categorize them into 19 distinct categories. This categorization reveals the predominant topics and themes in a user's posts, such as political discourse or sports-related content. These insights into user interests and preferences enable targeted content delivery and engagement strategies. Additionally, we calculate the percentage of posts in each category, providing a breakdown like 10% political posts, 12% sports posts, 15% religious posts, and so on.

User Interaction Post Categorization Using ML Models In the Facebook ecosystem, users engage actively with posts from friends, groups, and pages. We curate and annotate a list of these interaction posts, categorizing them into 19 distinct categories using machine learning algorithms. This categorization reveals patterns in user engagement, indicating preferences for topics, interests, or ideologies. Understanding these dynamics informs tailored content recommendations and targeted engagement strategies. Similarly, we calculate percentages for each category, similar to the approach used for categorizing user posts.

4.5 "When" Component Of The Context

Our methodology for temporal analysis of social media content revolves around leveraging state-of-the-art Natural Language Processing (NLP) techniques, with a particular emphasis on the "when" aspect of user-generated text. The approach entails several key steps, including search query preparation through tokenization and keyword extraction, internal search, external search and aggregating the search results. We describe these steps in below sections.

Tokenization and Preprocessing: The user's post undergoes tokenization, where it is broken down into individual words or tokens. Subsequently, stop words (common words like "the", "and", "is", etc.), punctuation marks, and non-alphanumeric characters are stripped away, leaving behind only meaningful content.

Keyword Extraction: After preprocessing, significant keywords are extracted from the post. These keywords serve as essential indicators of the central themes or topics within the user's message. Techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) or NLP (Natural Language Processing) algorithms are often employed to identify and rank the importance of these keywords.

Fetching Contemporary News from Facebook: Leveraging the extracted keywords, a search is conducted within the Facebook platform to retrieve contemporary news articles or posts that are closely related to the identified topics. For this purpose we utilize the Facebook search box to search across the Facebook. Also we search posts from user's friend list, Facebook pages, Facebook groups which are public.

Fetching Contemporary News from Google Search Engine: Similarly, the identified keywords are utilized to initiate a search on the Google search engine. This search retrieves news articles, blog posts, or other web content from across the internet that is pertinent to the topics identified in the user's post. For this purpose we develop a python script to load the google search page and perform search by keywords.

Aggregation and Analysis: The retrieved news articles and posts from both Facebook and Google are aggregated and analyzed to provide a comprehensive snapshot of contemporary events and topics that align with the themes of the user's post.

By employing this systematic approach, users are provided with access to real-time, relevant news events that coincide with the themes and topics they are discussing in their posts. This enhances their online experience by keeping them informed about current affairs and facilitating engagement with timely and pertinent content.

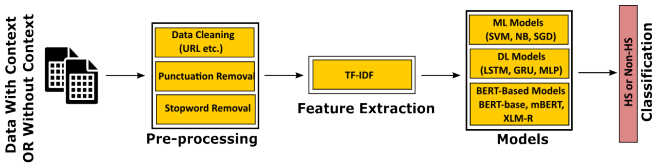

Fig. 5. Hate Speech (HS) Classification Process.

4.6 "Where" Component of the Context

We construct a comprehensive user post dataset by employing a diverse data collection strategy, gathering posts from various sources within the Facebook ecosystem. This includes aggregating posts from Facebook pages, user profiles, and groups, among others. By systematically traversing these channels, we capture a broad spectrum of user-generated content, ensuring the inclusion of diverse perspectives, topics, and engagement levels. This rich repository serves as a valuable resource for our research and analysis, enabling a holistic understanding of user behavior and interaction dynamics on Facebook. Figure 6 shows the architecture of collecting information of where part of the context.

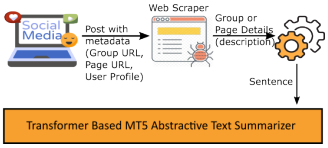

Fig. 6. Where component of the context.

4.7 "How" Component of the Context

Audience engagement on social networks involves reactions, comments, and shares, reflecting diverse opinions. In our study [16], we introduced a method to analyze sentiment trends and developed a template to illustrate these trends (Figure 7). By evaluating interactions such as likes, comments, and shares, we assess their collective impact and audience sentiment, offering insights into online discourse dynamics.

We integrate these insights with contextual elements (who, when, where) and pass the combined data to a summarizer. The summarized content, along with the original post, is then processed by our contextual hate speech (HS) classifier, as shown in Figure 1.

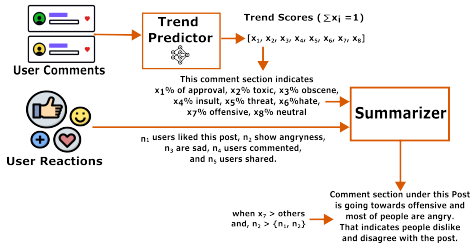

Fig. 7. How Component of the Context

5 Experiment

In this section we discuss the experimental details of the context components, experimental details of regular and contextual hate speech detection.

5.1 Who Part Experiment

In this section we describe experiment related to who part of the context object. We experiment on user persona categorization, user post categorization, and user interaction post categorization. We categorize posts and calculate the percentage of post categories.

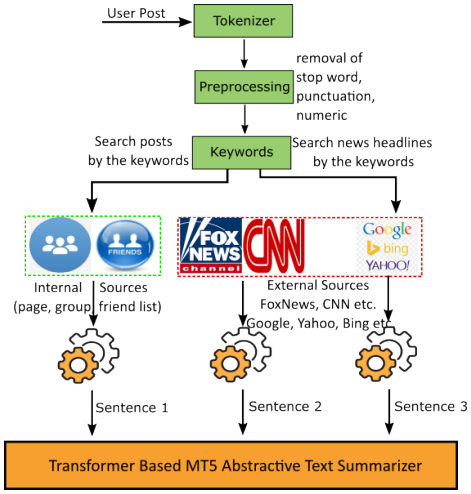

Fig. 8. When component of the context.

User Persona Categorization We collect user demographic dataset which consists of user personal information (name, date of birth, living place, gender etc.), education information, professional information, user interest (page likes) list. We analyze the dataset and create a feature set of n number. We collect 500 user demographic data. we train traditional machine learning models (SVM, NB, SGD etc.) to categorize users based on these features. We label demographic dataset using 12 categories.

User Post Categorization We collect 15000 user posts from 500 Facebook users. We manually annotate categories of these posts from a set of 19 categories. We train traditional machine learning models SVM, NB, SGD with their default parameters. We choose the best performing model comparing the accuracy values. We predict and compute the percentage of the categories for the given user's posts collected from his Facebook wall. Thus we get an idea of that user which reveals that user's distribution of topics of his content he posted in Facebook.

User Interaction Post Categorization Similar to the user post categorization we train machine learning models to categorize the posts in which the given user interacted until the given time when we perform the contextual hate speech detection.

5.2 When Part Experiment

In this section we describe how we perform the experiment to collect contemporary events news, articles from Facebook ecosystem which we call internal sources and from external sources utilizing the search engine.

Contemporary Events News And Articles From Internal Sources We outline a streamlined process for analyzing user-generated content on social media platforms in four steps. First, content extraction retrieves data followed by tokenization, breaking text into manageable units. The preprocessing phase removes stop words, numeric characters, and punctuation to enhance clarity. Leveraging Social context-based GPT and NLP techniques, significant keywords are extracted to reveal underlying themes. Lastly, these keywords guide the search through users' Facebook networks, pinpointing posts that align with research criteria for focused analysis.

Contemporary Events News And Articles From External Sources Initial steps like tokenization, preprocessing, and keyword extraction remain consistent for extracting content from external sources. However, the final step involves using these keywords to search across diverse platforms such as news websites and search engines like Google, Bing, and Yahoo. This approach ensures efficient discovery and analysis of articles relevant to research objectives, enhancing the precision and relevance of findings. The workflow is illustrated in Figure 8 for clarity.

Summarize The Contemporary Events News And Articles Using GPT4 We employ an advanced method to distill insights from diverse sources including news articles, reports, and Facebook posts. Using GPT-4 and prompt engineering via the completion API, we generate individual summaries focused on extracting core details from each article. These summaries are then combined into a cohesive overview using GPT-4, providing a comprehensive perspective on global events. This synthesized text enriches our 'context object,' enhancing our analysis with nuanced background information and ongoing discussions. This approach ensures the relevance and contextual awareness of our findings within a broader narrative framework.

5.3 Where Part Experiment

We identify where the user posted the post. We collect post metadata during collecting user post. A user can post in his/her personal Facebook wall, groups or pages. We Collect and process the where information as a component of the context which provides some context information like user interacts where, with what type of group or page. During the posts collection process we collect metadata like is it collected from user profile, group or page.

5.4 How Part Experiment

We fine-tuned RoBERTa using our referenced dataset [16], achieving an F1-score of 0.75 and an AUC of 0.92. The model's output was used to train XGBoost, which outperformed others with an F1-score of 0.79 in predicting trend scores. RoBERTa analyzes each comment in the comment section to generate emotion

scores, aggregated at the reply tree level using a bottom-up approach. These aggregated emotions are fed into XGBoost to derive final probability scores for trends. These scores are crucial for generating contextual insights, aiding in the detection of hate speech.

6 Results

6.1 Performance Comparison of ML Models in Post Categorization

We train machine learning models for categorizing user posts. We implement SVM, NB, SGD for this purpose. Table 1 shows the performance of the models in post categorization. We choose the best performing model which is SGD with the highest accuracy of 0.82.

Model	Accuracy	F1
SVM	0.80	0.85
NB	0.78	0.80
SGD	**0.82**	**0.88**

Table 1. Performance Comparison of ML Models in Post Categorization

Model	Accuracy	F1
SVM	0.76	0.78
NB	**0.80**	**0.82**
SGD	0.78	0.81

Table 2. Performance Comparison of ML Models in Persona Categorization

6.2 Performance Comparison of ML Models in Persona Categorization

We train machine learning models for categorizing user persona (political, athletic, business, religious etc.). We implement SVM, NB, SGD for this purpose. Table 2 shows the performance of the models in post categorization. We choose the best performing model which is NB with the highest accuracy of 0.80.

7 Testing in the Wild: Comparison Between Regular Hate Speech Detection and Contextual Hate Speech Detection

We established ground truth data through annotation of 100 user posts by multiple human judges to ensure diverse perspectives and mitigate biases. Discrepancies were resolved through rigorous discussions among judges, enhancing the

Hyperparameter	**M1**	**M2**	**M3**
Learning rate	2e-5	3e-5	5e-5
Epochs	10	10	10
Batch size	16	16	16
Dropout	0.2	0.2	0.2
Max seq length	128	128	128

Table 3. List of hyperparameters for BERT-based models. Here M1 = BERT-base, M2 = mBERT-uncased, M3 = XLM-RoBERTa-uncased.

reliability of our annotated dataset. This dataset served as a benchmark for evaluating hate speech detection using various machine learning (ML), deep learning (DL), and BERT-based models. Table 3 details the hyperparameters used for training BERT-based models on our dataset, while Table 4 presents performance metrics across different models. Incorporating contextual information generally improved performance for BERT-based models, with XLM-R achieving the highest accuracy (0.92) and F1 score (0.94). In contrast, SVM and MLP showed slightly higher accuracy and F1 scores without context compared to with context. These results underscore the significant advantage of leveraging contextual information in enhancing hate speech detection accuracy with BERT-based models.

Model	**With Context**		**Without Context**	
	Accuracy	**F1**	**Accuracy**	**F1**
ML Models				
SVM	0.82	0.84	0.85	0.87
NB	0.83	0.85	0.78	0.79
SGD	0.86	0.88	0.81	0.83
DL Models				
LSTM	0.88	0.90	0.83	0.84
GRU	0.87	0.89	0.82	0.83
MLP	0.84	0.86	0.85	0.87
BERT-based Models				
BERT-base	0.90	0.92	0.85	0.86
mBERT	0.91	0.93	0.86	0.88
XLM-R	**0.92**	**0.94**	**0.87**	**0.89**

Table 4. Performance Metrics for Binary Hate Speech Detection with and without Context

8 Conclusion & Future Work

Our research pioneers a novel approach to hate speech detection by integrating four distinct categories of contextual information, distinguishing it from previ-

ous studies. We gathered data from diverse sources including Facebook groups, pages, and individual profiles, utilizing a semi-automated annotation method to train a BERT-based model for automatic hate speech prediction utilizing a benchmark dataset [4]. Statistical analysis of our dataset was conducted, and experiments evaluated different contextual components, detailed in Table 1 and Table 2. We also assessed BERT-based models' effectiveness in detecting hate speech on 100 user posts, highlighting the impact of contextual information. Future work involves scaling experiments to larger datasets and exploring other BERT-based and language models to enhance contextual hate speech detection capabilities.

9 Acknowledgment

This research was supported by NSF grant CNS-2153482.

References

1. Das, M., Banerjee, S., Saha, P., Mukherjee, A.: Hate Speech and Offensive Language Detection in Bengali. arXiv preprint arXiv:2210.03479 (2022)
2. Ghosh, K., Senapati, A.: Hate Speech Detection: A Comparison of Mono and Multilingual Transformer Models with Cross-Language Evaluation. In: Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation, pp. 853–865 (2022)
3. Gomez, R., Gibert, J., Gomez, L., Karatzas, D.: Exploring Hate Speech Detection in Multimodal Publications. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1470–1478 (2020)
4. Jahan, M. S., Haque, M., Arhab, N., Oussalah, M.: BanglaHateBERT: BERT for Abusive Language Detection in Bengali. In: Proceedings of the Second International Workshop on Resources and Techniques for User Information in Abusive Language Analysis, pp. 8–15 (2022)
5. Kalita, G., Halder, E., Taparia, C., Vetagiri, A., Pakray, P.: Examining Hate Speech Detection Across Multiple Indo-Aryan Languages in Tasks 1 4. In: FIRE (Working Notes), pp. 474–485 (2023)
6. Karim, M. R., Dey, S. K., Islam, T., Sarker, S., Menon, M. H., Hossain, K., Hossain, M. A., Decker, S.: Deephateexplainer: Explainable Hate Speech Detection in Under-Resourced Bengali Language. In: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10 (2021)
7. Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., Mukherjee, A.: Hatexplain: A Benchmark Dataset for Explainable Hate Speech Detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 17, pp. 14867–14875 (2021)
8. Romim, N., Ahmed, M., Islam, M. S., Sharma, A. S., Talukder, H., Amin, M. R.: BD-SHS: A Benchmark Dataset for Learning to Detect Online Bangla Hate Speech in Different Social Contexts. arXiv preprint arXiv:2206.00372 (2022)
9. Romim, N., Ahmed, M., Islam, M. S., Sharma, A. S., Talukder, H., Amin, M. R.: HS-BAN: A Benchmark Dataset of Social Media Comments for Hate Speech Detection in Bangla. arXiv preprint arXiv:2112.01902 (2021)

10. Romim, N., Ahmed, M., Talukder, H., Islam, M. S.: Hate Speech Detection in the Bengali Language: A Dataset and Its Baseline Evaluation. In: Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020, pp. 457–468 (2021)
11. Tripto, N. I., Ali, M. E.: Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments. In: International Conference on Bangla Speech and Language Processing (ICBSLP), pp. Sept 21-22, Sylhet, Bangladesh (2018)
12. Simon Kemp.: DIGITAL 2021: BANGLADESH. (2021). https://datareportal.com/reports/digital-2021-bangladesh. Accessed 2 August 2024
13. StatCounter Global Stats.: Social Media Stats in Bangladesh. (2024). https://gs.statcounter.com/social-media-stats/all/bangladesh. Accessed 2 August 2024
14. stopwords-iso.: Stopwords Bengali. (2024). https://github.com/stopwords-iso/stopwords-bn. Accessed 2 August 2024
15. Hossain, I., Puppala, S., Alam, Md J., Talukder, S.: Monitoring Dynamics of Emotional Sentiment in Social Network Commentaries. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, pp. 51–55 (2023)
16. Hossain, I., Puppala, S., Alam, Md J., Talukder, S.: A Visual Approach to Tracking Emotional Sentiment Dynamics in Social Network Commentaries (2024)
17. Eglitis-media.: worlddata.info. (2024). https://www.worlddata.info/languages/bengali.php. Accessed 2 August 2024
18. Das, A. K., Al Asif, A., Paul, A., Hossain, M. N.: Bangla Hate Speech Detection on Social Media Using Attention-Based Recurrent Neural Network. Journal of Intelligent Systems, vol. 30, no. 1, pp. 578–591 (2021)
19. Ahmed, S., Alam, Md J., Talukder, S., Hossain, I.: Towards Addressing Identity Deception in Social Media using Bangla Text-Based Gender Identification. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, pp. 72–76 (2023)
20. Chung Hwan Kwak.: New World Encyclopedia. (2020). https://www.newworldencyclopedia.org/entry/Bengali_language. Accessed 2 August 2024