Bengali Misogyny Identification with Deep Learning and LIME

Shafakat Sowroar Arnob*, M. A. Ahad Shikder, Tashfiq Alam Ovey, Samiun jahan awishy and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{shafakat.sowroar.arnob*, m.a.ahad.shikder, tashfiq.alam.ovey, samiun.jahan.awishy}@g.bracu.ac.bd, annajiat@gmail.com

***Abstract*—The increase of misogyny across social media platforms highlights the urgent need to create efficient tools for recognizing and responding to gender-based online abuse. This study explores the complex problem of identifying instances of sexism in the Bengali language, a field that has a limited amount of research conducted due to a lack of financial resources and academic interest. We study the performance of BERT-based architectures, in recognizing misogynistic language by using the capabilities of deep learning models. Our research hypothesis is that enhancing the mBERT model with linguistic and cultural variety by employing multilingual training such as merging Bengali, Hindi, and English data for training will improve the ability to detect misogyny in Bengali, potentially transcending language barriers. We offer two extensive experiments that assess the performance of the models and give insight into the strengths and limits of those models. In addition, we employ LIME to uncover the decision-making processes of the models, enhancing their interpretability. Our results contribute to the development of improved methods for identifying online sexism, offering insights for the creation of safer digital environments. This study lays the groundwork for future research on language-specific nuances and cross-lingual trends in the field of gender-based abuse detection.**

***Index Terms*—Misogyny, Deep Learning, BERT, LIME**

I. Introduction

The disturbing continuation of patriarchal ideas in today's society is shown by the rise of misogyny, which is manifested in sexism, hostility, and degrading comments directed toward women and is especially popular on social media platforms. The widespread nature of this behavior makes it all the more critical that its social and mental health consequences be addressed. The internet has contributed to the spread of sexism in a society that is becoming more reliant on the digital sphere. Alarming figures show that over three-quarters (73%) of women who use the internet are victims of various forms of cybercrime, with just 23% choosing to report these crimes. More than three-quarters (73.71%) of victims are young adults (18 to 30), and a startling 30 percent have never sought help before [1]. Victimization may take many forms, including but not limited to cyberbullying, trolling, and the sharing of offensively sexist material. The use of misogynistic language, hate speech, and negative phrases are all instruments used to diminish the victim's sense of self, psychological health, and social status.

It is undeniably important to identify and mitigate such hostile and misogynistic behavior inside social networks, but doing so manually is hard and ineffective. Given the nuances and complexities of the problem, developing a universal automated method to identify misogynistic speech is a formidable challenge. The widespread availability of such harmful information online has motivated academics to focus on creating advanced speech and aggression detection systems.

Within this framework, we shift our attention to the difficulty of detecting misogyny in Bengali, a language that has received very little attention in the field of natural language processing. This study lays the groundwork for further research into the complex linguistic and cultural factors that contribute to sexism and misogyny. Our research focuses on a deep learning model with a BERT-based architecture, which has shown extraordinary success in text categorization challenges.

Our research endeavors are split into two distinct phases. The first step is an evaluation of deep learning models for misogyny detection in Bengali, including the monolingual BERT, BanglaBERT, and the multilingual mBERT. Afterward, we explore the potential of using multilingual datasets to improve the mBERT model's performance on Bengali misogyny identification. To do this, we train the model using Bengali, English, and Hindi data. This work is motivated by the hypothesis we have that incorporating additional linguistic and cultural data will improve the mBERT model's ability to detect Bengali misogyny, demonstrating the potential for bridging linguistic boundaries to deepen our understanding of the problem.

In addition to our modeling efforts, we use LIME (Local Interpretable Model-agnostic Explanations) to decipher the intuition behind the forecasts made by our deep learning models, BERT, BanglaBERT, and mBERT. This interpretability method provides insight into the categorization decisions made by these models and the underlying characteristics.

The next sections detail the dataset and preprocessing methods, the model architectures we developed, the experiments we ran, and our final thoughts on the outcomes. At its core, our study aims to illuminate the complex problem of misogyny detection in the Bengali language, using the capabilities of

advanced deep learning and explainability approaches to arrive at a comprehensive understanding of this crucial subject.

II. Related Works

Here, we take a look back at the research done on misogyny detection, deep learning models, and interpretability methods. Notable works using BERT and mBERT models for similar purposes are also highlighted.

A. Misogyny Identification :

Researchers in the field of social psychology have spent a lot of time trying to figure out how to detect sexist rhetoric and what it means for women. Recent efforts have centred on naming hate speech, which includes sexist language. Hateful comments in online user feedback can be identified with the help of neural networks built by Djuric et al. (2015), which represent text in a semantically coherent vector space [8]. To classify tweets as hate speech, offensive language, or neutral, Davidson et al. (2017) used a multi-class classifier trained on a hate speech lexicon [6]. Using Support Vector Machines (SVM) and critical race theory criteria, Waseem and Hovy (2016) created a dataset of sexist tweets [19]. While informative, these research have a narrow emphasis on sexist hostility [10].

B. Hate Speech:

Current study dives further into the analysis of hate speech, although typically only in monolingual settings. For the purpose of studying bigotry in all its forms—whether it be abuse, racism, sexism, or religious extremism—a new multilingual dataset has been established. This burgeoning area of study looks at how hate speech has affected different industries and social movements across time. Existing models have difficulty generalising to new data, which demonstrates the difficulty of the challenge. [12]. Much of the current research on the analysis of hate speech focuses on monolingual and monolingual categorization tasks. In this study, we propose a new dataset for the analysis of hate speech in many languages, including English, Hindi, Arabic, French, German, and Spanish. The dataset covers abuse, racism, sexism, religious hatred and extremism [20]. This inspection concentrates on works on hate speech, especially in the Web of Science-indexed law and communication studies. It examines published work in both English and Spanish and obscures the most common academic fields in which these studies have been authored, as well as their trends over time, by race, and by document type. This analysis is developed to identify discussion, study areas that are of greatest appeal, and develop hypotheses [15].

C. Aggression Detection :

The ability to identify hostile material has also attracted considerable interest. In order to spot antisocial features in Italian tweets, researchers have used a multi-agent classification technique [2]. Using lexical and semantic features with logistic regression and machine learning algorithms [18], [16], other studies have concentrated on specific languages such as Hindi and English.

D. Monolingual and Multilingual Approaches:

Research has carefully studied the effect of pre-training and task-specific training on models across languages, both from a monolingual and multilingual perspective. When it comes to spotting misogyny, monolingual models, and especially BERT-based ones, have proven to be exceptionally effective. HASOC-2019, TRAC, HatEval, and GermEval-2018 have all contributed to the development of the academic landscape surrounding the identification of abusive language and hate speech. While there have been numerous cross-lingual studies, there are still unanswered questions that could benefit from additional research [14]. A parallel endeavor undertaken by S.S. Saruar Jahan and Peom Dutta entailed the automatic development of a lexicon to identify instances of misogyny across multiple languages. This approach involved two discrete phases: "Identifying Misogyny" (Task A) and "Categorizing Misogynistic Behavior and Targets" (Task B). Employing an SVM classifier with RBF, parameters C = 5 and gamma = 0.10 for English, and gamma = 0.01 for Italian, this methodology yielded marginally superior results compared to human-curated lexicons [11]. Furthermore, recent years have witnessed a proliferation of academic events and collaborative undertakings centered around English, Spanish, Italian, German, Mexican-Spanish, and Hindi languages. This surge reflects the burgeoning interest within the NLP community towards hate speech and related matters, exemplified by the insights shared [3]. In the monolingual study, the researchers meticulously examined the impact of pre-training and task-specific training on models using exclusively English, Italian, and Spanish monolingual data. Remarkably, single-language BERT models emerged as frontrunners, achieving unparalleled performance in detecting misogyny within tweets across all three languages. An intriguing revelation from their error analysis was that, despite divergent training methodologies, both monolingual and multilingual models exhibited analogous error patterns [13].

E. BERT and mBERT Utilization :

Studies have explored the feasibility of utilizing BERT and mBERT models for a range of text analysis applications. Researchers conducted a multilingual study to evaluate these models' ability to transfer learning to the task of identifying misogynistic language in texts written in other languages. While BERT models trained on a single language performed admirably, mBERT and other multilingual setups showed promise. Error patterns were found to be consistent between monolingual and multilingual models despite significant differences in training methods [7].

III. Dataset and Preprocessing

This research makes use of a dataset that is multilingual in nature, including data in Bengali, English, and Hindi. The dataset used for our studies comes from Bhattacharya et al. (2020) [5]. It went through a series of preprocessing procedures to guarantee that it was suitable for our needs before it was used.

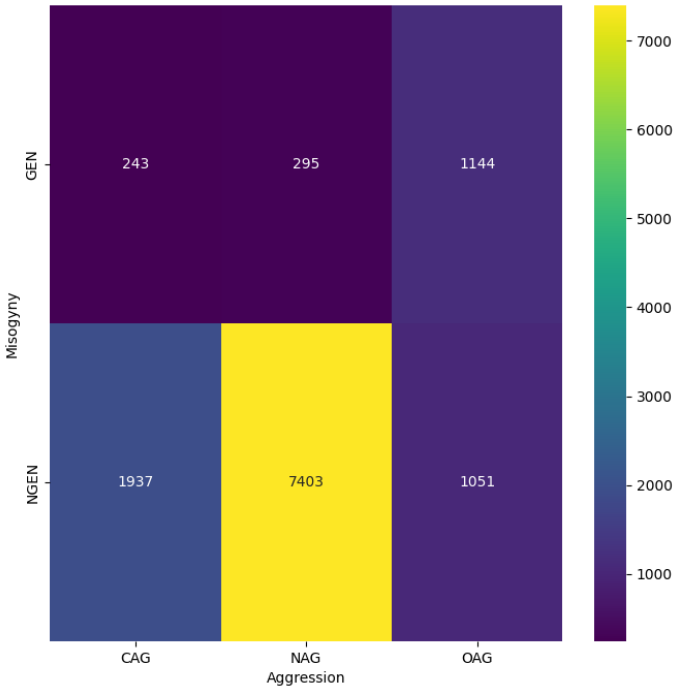

Fig. 1. Heatmap of Total Dataset with Aggression Labels and Misogyny Labels.

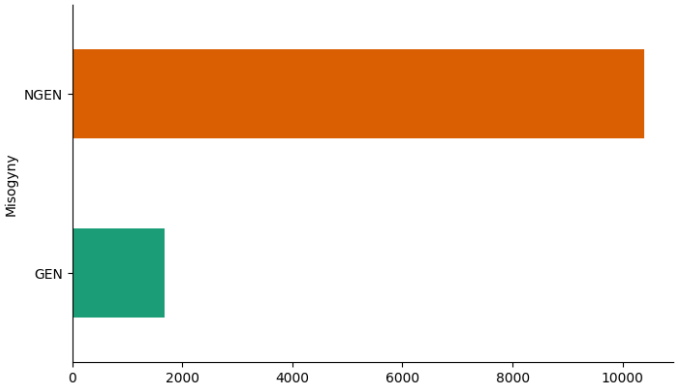

Fig. 2. Bar chart of Total Misogyny Labels in the Dataset.

The TRAC-2 dataset was designed to capture the nuances of specific language use, as well as different degrees of misogyny. The data were annotated into two misogyny classes: Gendered (GEN) and Non-Gendered (NGEN). We labeled the data for our pre-processing needs. Labeling helped realize these divisions, gendered data were given the value 1 (GEN), while non-gendered expressions were given the value 0 (NGEN). In this dataset, the data are also annotated in these 3 aggressive classes: Not Aggressive (NAG), Covertly Aggressive (CAG), and Overtly Aggressive (OAG). Some Examples:

In Fig. 1 and Fig. 2, we can see that the GEN Data are much less compared to the NGEN DATA. If the NGEN and GEN data were somewhat equal then the trained model would have performed better. However, Bengali is a very low-resourced language, and the dataset for misogyny in this language is very rare.

TABLE I
DATA SET EXAMPLES

1.	*"Mollar bachara chara kau ai kaj korta parana"*	(NGEN)
2.	"Re Vai bon er somporko ta khrap koro na.."	(GEN)

From the above examples, we can see that some Bengali data are written with English letters. So, there is a mixture of Bengali and English text in the Bengali data part. We preprocessed the data and a Unicode Normalization procedure was used to ensure optimal understanding and standardization for our model training.

There were many other measures taken in text preprocessing besides language normalization. Methods for cleaning text were used to improve the data quality by getting rid of irrelevant information including letters, symbols, and formatting. To further prepare the text as input for our deep learning models, Tokenization, the process of segmenting text into discrete units, was carried out.

To ensure that the categorical labels for gendered language, non-gendered language, and aggressiveness levels were translated into numeric values, an encoding procedure called label encoding was implemented. Because of this adjustment, the data could be used in the models' training and testing procedures with minimal effort.

We have prepared a solid groundwork for our deep learning studies to uncover misogyny within the multilingual data by carefully organizing and preparing the TRAC-2 dataset, which included language normalization, text cleaning, tokenization, and label encoding.

IV. METHODOLOGY

This section discusses our research technique. We explain the monolingual BERT, BanglaBERT, and multilingual mBERT models' architecture, training procedure, parameter settings, evaluation setup, and LIME model interpretation [7] [4].

A. Model Architecture

Bidirectional Encoder Representations from Transformers (BERT), a language modeling pioneer, inspired our models' architecture. BERT's usage of the Transformer's attention mechanism enables bidirectional training, surpassing previous methods that relied on left-to-right or combination training. The bidirectional method gives language models a deep sense of context and coherence. Bidirectional training is possible using Masked Language Modelling (MLM). Transformer, a text attention mechanism, captures complex contextual links between words or sub-words in BERT. The Transformer's default setup includes an encoder for text inputs and a decoder for task prediction. BERT's language modeling capabilities rely exclusively on the encoder [9]. This architecture allows

our models to understand the Bengali language and recognize misogynist statements.

B. Training Process, Parameter & Evaluation Metrics

Our Bengali misogyny-detecting algorithms were rigorously trained. The training procedure used back-propagation to alter model weights throughout successive iterations. To improve model performance, we used weight decay, warm-ups, logging strategy, etc. Using iterative testing and best practices, learning rate, batch size, and optimizer were chosen.

A thorough set of assessment measures assessed our models' effectiveness. Calculated accuracy, precision, recall, F1-score, and loss. These metrics assess the models' gendered and non-gendered expression classification and performance.

C. Multilingual Training Experiment

The Bengali data from the TRAC-2 dataset were used in our initial experiment to fine-tune the monolingual BERT, BanglaBERT, and mBERT models. This was done completely using Bengali data. Following that, in the second experiment we conducted, we broadened the scope of our investigation to include a multilingual training paradigm. In this stage of the process, the mBERT model received further training with data including all three languages: Bengali, English, and Hindi. Our hypothesis suggested that including language and cultural variety in the mBERT model should improve its sensitivity to the identification of Bengali sexism. our theory served as the inspiration for our multilingual approach. Through the use of this experiment, we were able to investigate if the mBERT model could make use of information from a variety of languages in order to enhance its performance in the detection of misogyny in Bengali.

D. LIME for Model Interpretation

LIME, which stands for Local Interpretable Model-agnostic Explanations, is what we went to in our effort to figure out how our deep learning models get to their conclusions. Through the use of LIME, we are able to understand the thought process that goes into model predictions. This interpretability technique takes complicated model behavior and breaks it down into insights that are easier to understand by providing locally faithful approximations of model behavior for particular situations. By using LIME in our BERT, BanglaBERT, and mBERT models, we want to provide insight into the elements that drive their misogyny predictions, improving our knowledge of how models behave.

V. Experiments and Results

A. Experiment 1: BERT, BanglaBERT and mBERT

Experiment 1 was carried out by us in order to determine how effective our models are in recognizing instances of sexism within the Bengali language. This endeavor required extensive preparation measures to assure the quality of the data, such as text cleaning, the elimination of null cells, and the normalization of Unicode characters. The dataset was painstakingly labeled to make the classification of phrases into gendered (GEN) and non-gendered (NGEN) categories easier. After performing tokenization to segment the texts, the results were then turned into encoded tensor slices to provide the foundation for the training of the model.

TABLE II
Comparison of Evaluation Metrics

Model	Loss	Accuracy	F1 Score	Precision	Recall
BERT	*0.222441*	*0.914794*	*0.853394*	*0.885316*	*0.829449*
mBERT	*0.281855*	*0.902771*	*0.826363*	*0.876121*	*0.794358*
BanglaBERT	*0.358021*	*0.870883*	*0.715384*	*0.923560*	*0.669163*

Following that, we worked on refining the monolingual BERT, BanglaBERT, and multilingual mBERT models using the Bengali data that was obtained from TRAC-2. Our assessment was based on important parameters that included accuracy, F1-score, precision, and loss. The performance of the model was further improved by the implementation of several tactics, such as weight decay, warm-up phases, logging steps, etc.

The assessment metrics of BERT, mBERT, and BanglaBERT, which are shown in Table 1, revealed insights into the various performances of the three systems. The findings highlight the efficacy of the monolingual BERT in recognizing Bengali misogyny since it achieved the maximum accuracy and F1 score possible. mBERT comes in a close second, displaying competitive accuracy and an F1 score while simultaneously supporting several languages. Although BanglaBERT has a somewhat poorer performance, it still has the highest precision and respectable numbers for accuracy and recall. Notably, all of the models exhibit a trade-off between accuracy and recall, which is an indication of the delicate balance that must be maintained between reducing the number of false positives and capturing the actual positives.

The remarkable performance of BERT is suggestive of its capacity to recognize nuanced details of misogyny in Bengali. Despite a little decrease in performance, mBERT's adaptability shows through as it handles the multilingual component in a decent manner. While BanglaBERT's accuracy suggests that it is capable of accurately categorizing gendered utterances, its recall indicates that there is a need for development in terms of recognizing more instances of misogyny. In spite of optimization tactics, model performance is affected by the quality of the data as well as the amount of the dataset. In conclusion, Experiment 1 highlighted the various strengths of BERT, mBERT, and BanglaBERT in the detection of Bengali misogyny, highlighting paths for further research and the potential for improvement.

B. Experiment 2: Multilingual Training

Experiment 2 examines the effects of training the multilingual mBERT model using Bengali, English, and Hindi datasets. Data augmentation to provide language and cultural variety may improve the model's Bengali misogyny detection.

Our second experiment included Bengali, English, and Hindi datasets. A complete pretreatment pipeline included text

TABLE III
EVALUATION METRICS OF MBERT AFTER TRAINING IT ON MERGED DATA

Model	Loss	Accuracy	F1 Score	Precision	Recall
BERT(Bengali)	*0.222441*	*0.914794*	*0.853394*	*0.885316*	*0.829449*
mBERT(Bengali)	*0.281855*	*0.902771*	*0.826363*	*0.876121*	*0.794358*
mBERT(Bengali+Hindi+English)	*0.275801*	*0.907513*	*0.838960*	*0.890131*	*0.805122*

cleaning, null cell removal, and gendered and non-gendered expression labeling. The tokenized and encoded tensor slices were used for model training. The multilingual mBERT model was fine-tuned using three languages' data to improve its linguistic adaptability. Weight decay, warm-up stages, and logging scheme improved model performance.

Table 2 provides a comparative examination of the performance of the model under a variety of distinct training settings. We observe that mBERT trained on merged datasets exhibits improved performance in terms of accuracy, F1-score, precision, and recall compared to the mBERT only trained in Bengali. This result provides credibility to our hypothesis that including data from a wider variety of languages increases the model's capacity to recognize misogynistic language in Bengali. Now, the mBERT model's ability to leverage linguistic diversity becomes evident, underpinning the potency of multilingual training in enhancing cross-lingual proficiency.

It is important to highlight that while the findings are encouraging, the improvement is not very significant. This might be defined as the difficulties provided by the linguistic variances between Bengali, English, and Hindi, as well as the possible noise created by various language data. Additionally, this could be linked to the fact that there is a wide variety of language data. In addition, the performance improvement was probably helped by model optimization tactics including weight decay and warm-up phases.

VI. LIME EXPLANATIONS

LIME, which stands for Local Interpretable Model-agnostic Explanations, is a technique that we used so that we could acquire a more in-depth understanding of the decision-making process that our models through [17]. This method assists us in comprehending the elements that drive the models' predictions and sheds light on the behaviors that lie under the surface of such models.

An example, "ei meye ta magi hoye geche oke chude de," was chosen for analysis, and LIME explanations were used to categorize it using BERT, BanglaBERT, and mBERT models.

A. BERT LIME Explanation

The probabilities that are given to words in the BERT model represent how important those words are to the final classification decision. The word "magi" has the greatest likelihood of 0.570 in the expression "ei meye ta magi hoye geche oke chude de," followed by the word "chude" with a value of 0.210. This suggests that both terms contribute significantly to the model's prediction of "GEN" for this specific occurrence. When the probability value is larger, a word is more relevant in its capacity to influence the model's decision.

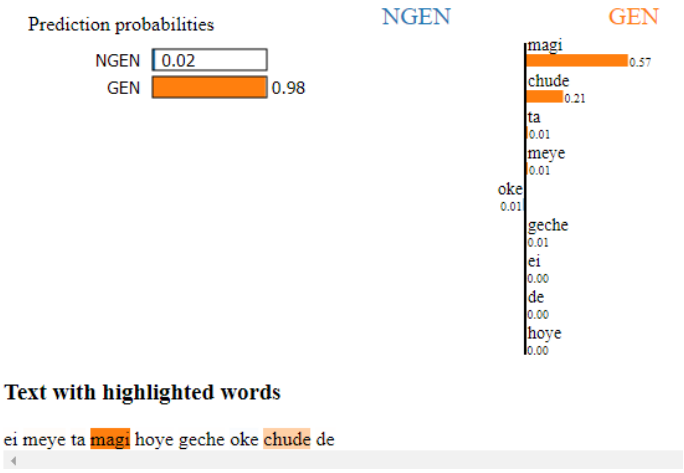

Fig. 3. BERT: LIME Explanation

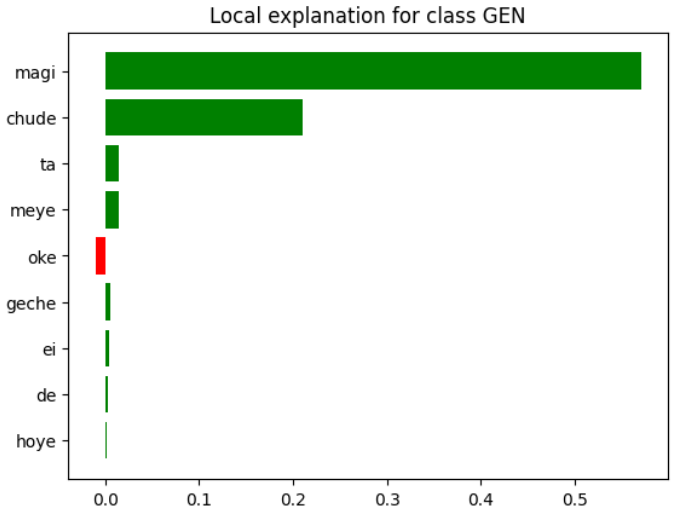

Fig. 4. BERT: Probability Plot of words in the Instance

B. mBERT LIME Explanation

In the example of mBERT, which also labeled the occurrence as "GEN," the "magi" explanation is highlighted by the LIME explanation with an even greater likelihood of 0.889, reiterating its significant influence on the prediction. On the other hand, and here is where things get interesting, the word "chude" has been given a far lower likelihood of 0.010, which suggests that mBERT thinks "chude" to be less definitive than "magi" in this context.

C. BanglaBERT Explanation

The word "magi" is given a probability of 0.751 according to the LIME explanation when applied to BanglaBERT, which also categorizes the case as "GEN." This demonstrates how significant it is in terms of deciding the forecast. This explanation, in contrast to the others, does not place a

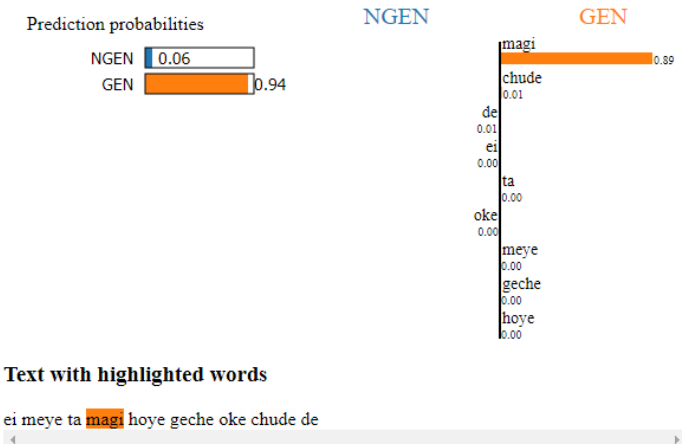

Fig. 5. mBERT: LIME Explanation

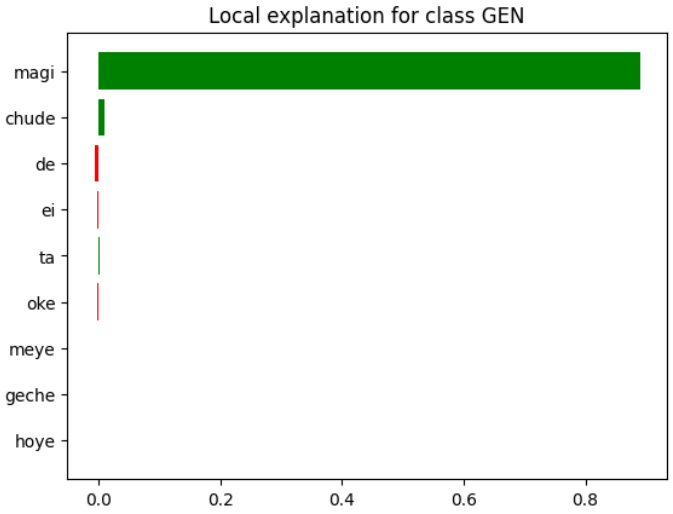

Fig. 6. mBERT: Probability Plot of words in the Instance

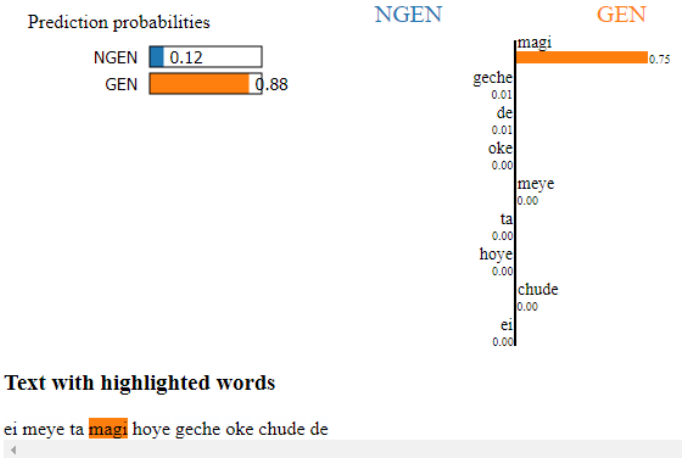

Fig. 7. BanglaBERT: LIME Explanation

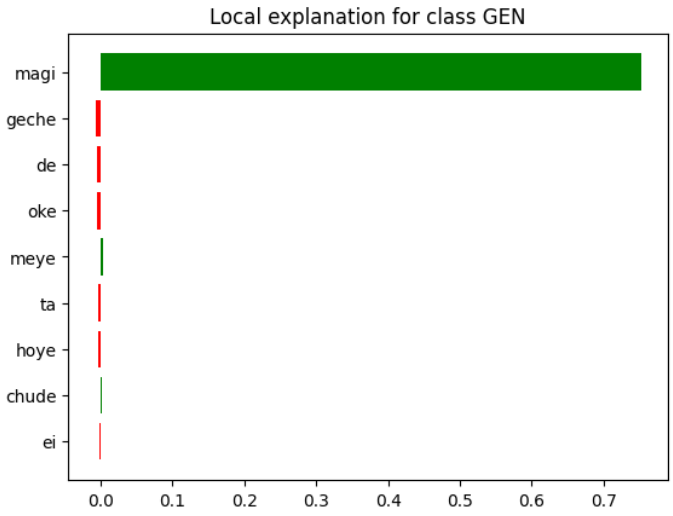

Fig. 8. BanglaBERT: Probability Plot of words in the Instance

considerable emphasis on the word ”chude,” as seen by the comparatively low chance of 0.002 that it will be used.

These probability values highlight the different degrees of effect that certain words have on the models' predictions, which can be found in the previous sentence. The greater the likelihood that is associated with a term, the more significant its function in steering the model towards a certain category will be. It is important to note that although the word ”magi” regularly emerges as a vital feature, the emphasis on ”chude” varies across the models. This is something that should be taken into consideration. This variance might be the result of the models' different understandings of the language context and patterns.

Overall, these probability values give a detailed picture of how words contribute to the judgments made by the models, which enables us to study the complexities of their classification procedures in recognizing misogyny in the Bengali language.

VII. Discussion

Through the use of deep learning models, the purpose of this research was to take on the difficult task of identifying instances of misogyny in the Bengali language. The most important data that we have obtained shed light on the possibilities of BERT-based designs, the influence of multilingual training, and the interpretability benefits that LIME offers.

In the context of identifying sexism, our investigation of the BERT, BanglaBERT, and multilingual mBERT models revealed some very interesting results. The performance of the models was encouraging, with BERT obtaining an accuracy of 0.914, mBERT achieving 0.903, and BanglaBERT achieving 0.871. These findings highlight the efficacy of deep learning models in capturing the complexity of sexist language in

Bengali and demonstrate their potential to contribute to the identification of gender-based online harassment.

The results of our studies provide credence to the idea that the performance of the models might be improved with the use of multilingual training. By training mBERT on a dataset that included Bengali, English, and Hindi, we were able to increase its performance in terms of accuracy (0.908) and F1-score (0.839) for the detection of sexism in Bengali. This outcome is in accordance with our predictions since the addition of language and cultural variety seems to expand the model's knowledge and discriminatory ability, presenting a more complete view of the topic being studied. In support of the hypothesis:

- **Shared Semantics:** mBERT is designed to understand multiple languages. While Bengali, English, and Hindi are linguistically distinct, there might be shared semantic concepts or syntactic structures that the model could leverage to enhance its understanding of misogyny.
- **Cross-Lingual Patterns:** By training in multiple languages, the model might learn cross-lingual patterns and gain a broader perspective on the features that contribute to identifying misogyny.
- **Indic Language Group:** Bengali and Hindi belong to the same Indic language family, so there could be linguistic similarities that assist the model in understanding the context and nuances in Bengali text.
- **Data Abundance:** Adding more data generally helps improve model performance, as long as the data is relevant and diverse enough to capture different aspects of the problem.

Nevertheless, despite the fact that our data provide credence to the idea, there are a few important caveats to take into account. The differences in semantics and patterns that may be shared across Bengali, English, and Hindi may also be obscured by the complexities and confusions that might be introduced by the languages' linguistic variances. The problem still lies in finding a happy medium between learning from a variety of languages without adding any preconceived notions or leading to misunderstanding. In addition, the success of multilingual instruction is highly dependent on the quality of the additional material as well as its applicability to the situation. The possibility of the potential advantages of greater data volume being nullified or diminished due to the existence of noisy or irrelevant data is not out of the question.

In addition, we were able to offer interpretability to the model's predictions via the use of LIME, which facilitated a better comprehension of the decision-making processes that they engaged in. We acquired insights into the characteristics that lead the algorithms' detection of sexism by analyzing the contribution of individual words to the predictions. Not only may these explanations assist in establishing faith in the conclusions that the models produce, but they can also help uncover possible areas in which the models might be improved.

In conclusion, the findings of this study highlight the potential of deep learning models for the detection of sexism in the Bengali language. The efficacy of BERT-based designs, the advantages of multilingual training, and the interpretability made possible by LIME all contribute, together, to a more in-depth comprehension of this very important matter. Our results provide the groundwork for additional developments in automated detection systems that might contribute to safer digital environments for all users, which is an urgent issue given the ongoing prevalence of online gender-based abuse.

VIII. Conclusion and future work

Within the scope of this research, we investigated the detection of misogyny in the Bengali language in order to address the critical problem of gender-based online harassment. Our study provided a number of key advances, extending our knowledge of the efficacy of deep learning models, the influence of training in several languages, and the part that interpretability plays in increasing the trustworthiness of models.

The most important results from our study show that BERT-based models have a lot of potential, as seen by the excellent performance of BERT, BanglaBERT, and the multilingual version of mBERT when it came to detecting misogynistic language in Bengali. We demonstrated that these models are capable of accurately capturing the subtleties of sexist speech, which contributes to the development of more secure settings online.

In addition, the results of our studies with multilingual training provided support for our hypothesis, which stated that include elements of language and cultural diversity might potentially improve model performance. The better findings achieved by training mBERT on a combined dataset of Bengali, English, and Hindi demonstrate the potential advantages of using cross-lingual patterns to expand one's knowledge of sexism. [Cross-lingual patterns] are patterns that appear in more than one language.

Despite these contributions, it is essential to recognize that there are certain constraints. The breadth of our investigation was limited as a result of the dearth of labelled datasets in Bengali, particularly those that were focused on sexism. In addition, despite the fact that our models have shown promising performance, it is possible that they still display biases or have shortcomings when it comes to managing certain language subtleties.

There are a number of directions that should be looked at for future study. Extending the scope of our research to include additional languages may provide helpful new perspectives on the extent to which our results are generalizable. It may be possible to achieve optimal model performance by investigating a variety of fine-tuning procedures, such as changing learning rates or training methodologies. Additionally, testing with additional explainability approaches outside LIME, such as attention visualization techniques or feature significance analysis, might give a variety of viewpoints on the decision-making process for the model.

In conclusion, the findings of our study provide a contribution to the continuing conversation about the detection of sexism by demonstrating the potential of deep learning models, multilingual training, and interpretability tools. Even if our work offers a solid basis, the ever-changing nature of online abuse necessitates ongoing study and innovation in order to guarantee the development of digital environments that are both secure and welcoming for all users.

REFERENCES

[1] Md Sayeed Al-Zaman. Gendered communication and women's vulnerability in digital media of bangladesh. 2021.

[2] Giuseppe Attanasio and Eliana Pastor. Politeam@ ami: Improving sentence embedding similarity with misogyny lexicons for automatic misogyny identification in italian tweets. In *EVALITA*, 2020.

[3] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63, 2019.

[4] Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States, July 2022. Association for Computational Linguistics.

[5] Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France, May 2020. European Language Resources Association (ELRA).

[6] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[8] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30, 2015.

[9] Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online, November 2020. Association for Computational Linguistics.

[10] Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16, 2017.

[11] Farina Mir. *The social space of language: Punjabi popular narrative in colonial India, c. 1850–1900*. Columbia University, 2002.

[12] Arka Mitra and Priyanshu Sankhala. Multilingual hate speech and offensive content detection using modified cross-entropy loss. *arXiv preprint arXiv:2202.02635*, 2022.

[13] Arianna Muti and Alberto Barrón-Cedeño. A checkpoint on multilingual misogyny identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 454–460, 2022.

[14] Endang Wahyu Pamungkas and Viviana Patti. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, pages 363–370, 2019.

[15] María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. Hate speech: A systematized review. *Sage Open*, 10(4):2158244020973022, 2020.

[16] Abir Rahali, Moulay A Akhloufi, Anne-Marie Therien-Daniel, and Eloi Brassard-Gourdeau. Automatic misogyny detection in social media platforms using attention-based bidirectional-lstm. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2706–2711. IEEE, 2021.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[18] Niloofar Safi Samghabadi, Parth Patwa, Srinivas Pykl, Prerana Mukherjee, Amitava Das, and Thamar Solorio. Aggression and misogyny detection using bert: A multi-task approach. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 126–131, 2020.

[19] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[20] Ankit Yadav, Shubham Chandel, Sushant Chatufale, and Anil Bandhakavi. Lahm: Large annotated dataset for multi-domain and multilingual hate speech identification. *arXiv preprint arXiv:2304.00913*, 2023.